Longitudinal clinical assessment of undergraduate dental students: Building an argument for its validity

Jamie Daniel Robert Dickie

BDS, MFDS RCPS (Glasg), MEd

Submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

School of Medicine, Dentistry, and Nursing

College of Medical, Veterinary, and Life Sciences

June 2021

Acknowledgements

There are numerous people I wish to thank for their help and support at various points over the past three-and-a-half years. Firstly, I would like to acknowledge the assistance of Dr Kurt Naudi, Dr Andrea Sherriff and Dr Michael McEwan, for their invaluable mentoring in their role as my supervisors throughout this study.

To those who volunteered to participate in the focus groups I am particularly grateful.

The early contribution extended by Professor Vince Bissell in relation to his ideas for dental education research and the initiation of this study is gratefully acknowledged. Gratitude is also due to Professor Luke Dawson for his discussions on longitudinal assessment and the LIFTUPP© system, and to Dr Aileen Bell for her input on the assessment processes used at Glasgow Dental School.

In relation to obtainment of the assessment data sets, my appreciation is expressed to Dan Smith, Marion Howat, Claire Rodgers, Jessica Brewster, Susan Johnston, Frank Bonner, Jimmy Boyle, Tracey Gill, and Patrick Maitland-Cullen. For his expertise and hard work in linking all the data sets together, Dr Alex McMahon's assistance has been invaluable.

I want to thank my educational supervisors, Dr Laura Cross and Dr Douglas Robertson, for their guidance, good counsel, and reassurance during my time away from the Restorative Dentistry specialist training programme to complete this study.

Professor Jeremy Bagg, Dr Donald Thompson and Dr David Felix kindly offered me the opportunity to pursue this PhD study as an adjunct to my specialist training post and offered their continued support (and patience) throughout. I must also express my gratitude for the funding of this study, which was provided by the Dorothy Geddes Scholarship and NHS Education for Scotland.

My appreciation goes to Mr David Watson for his attention to detail in proofreading, as well as his previous contributions to the development of my academic writing.

Thanks to all my friends for their good humour and encouragement over the years, and for keeping me grounded.

Finally, my parents John, Denise, sister Chloe and the rest of my family sustain me with love and support throughout my studies, helping me to realise my ambitions.

List of Contents

List of Tables

List of Figures

List of Appendices

List of Abbreviations

ADEA - American Dental Education Association

ADEE – Association for Dental Education in Europe

ADEPT(s) - Dental evaluation of performance test(s)

AIC- Akaike information criterion

AMEE - Association for Medical Education in Europe

ANOVA - Analysis of variance

App – Application

AT(s) - Assessment Task(s)

AvePP - Average posterior probability

BDS – Bachelor of Dental Surgery

BIC(s) - Bayesian information criterion(s)

BRM - Borderline regression method

CBD - Case-based Discussion

CDP - Clinical Development Panel

COVID-19 – Coronavirus disease 2019

CVI – Content validity index

CVR – Content validity ratio

DOPS - Direct observation of procedural skills

DPIA - Data Protection Impact Assessment

DVT - Dental vocational training

EMQ(s) – Extended matching question(s)

EPA(s) - Entrustable professional activities

EU – European Union

GBTM - Group-based trajectory model/modelling

GCC - General Chiropractic Council

GCM - Growth curve model/modelling

GDC – General Dental Council

GDPR – General Data Protection Regulations

GLM - Generalised linear modelling

GLMM - Generalised linear mixed-effects model/modelling

GMC – General Medical Council

GMM - Growth mixture model/modelling

GOC - General Optical Council

GOsC- General Osteopathic Council

GPC - General Pharmaceutical Council

HCP(s) - Healthcare professional(s)

HCPC - Health and Care Professions Council

HEA - Higher Education Academy

ICO - Information Commissioner's Office (UK)

ID – Identification

IUA - Interpretation/use argument

KR-20 - Kuder–Richardson

LCGM - Latent class growth modelling

LEP(s) – Longitudinal evaluation(s) of performance

LIFTUPP© - Longitudinal Integrative Foundation Training Undergraduate to Postgraduate Pathway

LO(s) - Learning outcome(s)

Ltd – Limited company

MCQ(s) – Multiple-choice question(s)

MEQ(s) - Modified essay question(s)

Mini-CEX(s) - Mini-clinical evaluation exercise(s)

MMIs - Multiple mini Interviews

MSA(s) – Multiple-short answer(s)

MSF - Multi-source feedback

MVLS - Medical, Veterinary and Life Sciences

NES – NHS Education for Scotland

NHS – National Health Service (UK)

NMC - Nursing and Midwifery Council

OSCE(s) – Objective structured clinical examination(s)

PfP – Preparing for Practice (GDC)

PhD – Doctor of Philosophy

PI(s) – Performance indicator(s)

PTSD - Post-traumatic syndrome disorder

QAA - Quality Assurance Agency for Higher Education

Q1 – Lower (first) quartile

Q3 – Upper (third) quartile

r - Pearson correlation coefficient

R^2 – R-squared

ROC - Receiver operating characteristic

SAQ(s) – Short answer question(s)

SEM - Standard error of measurement

SBA – Single best answer

SCT - Script concordance test

SD – Standard deviation

SJT(s) - Situational judgement test(s)

SOHRC – Scottish Oral Health Research Collaboration

SRL - Self-regulated learning

TJE – Triple jump exercise

TLA - Teaching/Learning activities

UCAT - University Clinical Aptitude Test

VDP – Vocational dental practitioner

VT – Vocational training

VDT – Vocational dental training

WBA(s) - Workplace-based assessment(s)

USA – The United States of America

UK – The United Kingdom of Great Britain and Northern Ireland

W - Kendall's coefficient of concordance

WGEE(s) - Weighted generalised estimating equation(s)

Author's Declaration

Parts of the research work included in this thesis have been submitted to the European Journal of Education for publication. The details of the submitted article are as follows:

Title – Longitudinal clinical assessment of undergraduate dental students: Building evidence for validity

Authors – Dickie, J., McEwan, M., Sherriff, A., Bell, A., Naudi, K.

In addition, some of the study results have been presented at the following national and international and conferences:

National

Scottish Oral Health Research Collaboration, 1st October 2019, Dundee UK
Oral presentation: Longitudinal assessment of dental students

NHS Education for Scotland Annual Virtual Conference, 27th – 28th May 2021, online
Oral presentation: The relationship between undergraduate longitudinal clinical performance and examination outcomes

International

Association for Medical Education in Europe (AMEE): The Virtual Conference, 7th – 9th September 2020, online
Oral presentation: Longitudinal assessment of dental students: Building an argument for its validity

I declare that, except where explicit reference is made to the contribution of others, this thesis has resulted from my work and has not been submitted for any other degree or to any other institution.

Jamie Dickie

Summary

Background

Assessment of healthcare professionals plays a pivotal role in safeguarding patients by ensuring practitioners have been appropriately trained before being permitted onto professional registers. This prevents the public from being treated by those who are not fit to practise healthcare subjects, including dentistry.

In the UK, dental schools must provide the General Dental Council (GDC) with evidence that students have attained the necessary educational outcomes and are suitable to join the professional register. The GDC delegates responsibility of choosing appropriate assessment methods to obtain such evidence to the dental schools themselves. As part of their undergraduate assessment repertoire, some UK dental schools have adopted longitudinal assessment methods to measure development and consistency of competent performance in clinical environments. Although these longitudinal methods create a rich database of multiple points of evaluation over the duration of the Bachelor of Dental Surgery (BDS) curriculum, there is currently little evidence to support their use for assessing development of clinical competence. Therefore, there is a need to conduct thorough analyses of longitudinal clinical data using robust statistical methods and create evidence to support their validity for this purpose.

Aims

This thesis aims to investigate the content and criterion validity and reliability of longitudinal clinical assessment, which will contribute towards a validity argument on its use in assessing the development of clinical competence among undergraduate dental students. It will also explore how the evidence for validity could be used to enhance assessment within dental education.

Research design

A mixed methods approach, with quantitative and qualitative approaches, was adopted to address the study aims. For the quantitative component, statistical

descriptions, and group-based trajectory models (GBTMs) tracking individual undergraduate's clinical performance over time were produced from longitudinal clinical assessment (LIFTUPP©) data for three dental student cohort's (2017-19; n=234). Content validity was investigated using LIFTUPP© performance indicator 4 as the threshold for competence. Distinct trajectories were created using a performance indicator 5 as the threshold, which were then used to investigate the concurrent and predictive subtypes of criterion validity.

Concurrent validity was investigated by linking and cross-tabulating LIFTUPP© trajectory group memberships with BDS examination performance (mean scores and a "top 20%" performance in each BDS year). Predictive validity was investigated by linking and cross-tabulating undergraduate LIFTUPP© trajectory group memberships with postgraduate clinical performance trajectory group memberships generated from Longitudinal Evaluations of Performance (LEPs). Reliability was calculated using Cronbach's alpha.

For the qualitative component of the study, a series of online focus groups with key stakeholders within dental education were conducted. Participants were presented with the results of the quantitative analyses and their opinions on how these data could be used to enhance assessment within dental education were canvassed. Transcripts of the focus group discussions were analysed using thematic analysis to identify themes (i.e., patterns) of interest within the data.

Results

LIFTUPP© GBTMs with a threshold performance indicator of 4 resulted in all students following a single trajectory in all three cohorts and showed progressive development of clinical competence over three BDS clinical years, satisfying criteria for content validity. GBTMs with a threshold performance indicator of 5 provided at least two distinct trajectories of student clinical performance. According to the Bayesian information criterion (BIC), models with two distinct trajectories fitted the data best and a "better" performing trajectory was identifiable in each cohort.

In the two most recent cohorts, students who were more likely to belong to the "better performing" LIFTUPP© trajectory scored higher (on average) in the

undergraduate examinations for each BDS year. This association was not observed for cohort 1. Students allocated to "better performing" LIFTUPP© trajectories were more likely to also be assigned to "better performing" LEP trajectories in all three cohorts. Reliability for the undergraduate examinations was high in all three cohorts ($\geq$0.88) and did not change substantially when longitudinal clinical assessment data were included.

Comments from focus group participants appeared to provide further support for content validity. However, quantitative results were met with a degree of mistrust that seemed to stem primarily from previous experiences of operational issues associated with the LIFTUPP© assessment process and the absence of contextual data within the quantitative analyses.

Conclusions

The upward trend of LIFTUPP© trajectory patterns suggested there is evidence that longitudinal clinical performance data have content validity for the assessment of clinical competence. Associations between better LIFTUPP© performance and better undergraduate examination outcomes and better postgraduate clinical performance in the two most recent cohorts were indicative of criterion validity. The lack of association in cohort 1 may have been due to poorer calibration among assessors following the initial adoption of LIFTUPP© into the BDS curriculum.

Evidence for LIFTUPP© data reliability was inconclusive. This uncertainty may have resulted from using probabilities of student trajectory group membership as the metric for longitudinal clinical assessment in the calculation of Cronbach's alpha. Therefore, further investigations on LIFTUPP© data reliability are required.

Data processing procedures and suggestions from focus group participants revealed there is a need to improve current assessment practices and data collection to allow other investigations on validity to be pursued and to further increase confidence in the results produced by this study. Some data collection issues encountered in relation to LIFTUPP© and undergraduate examinations have since been resolved, meaning studies involving subsequent student cohorts

should seek to incorporate LIFTUPP© communication, management and leadership, and professionalism data as well data from clinical case presentation examinations and one-off clinical competence tests.

Overall, the study provides an early contribution towards a validity argument on the use of longitudinal clinical assessment in assessing development of competence in undergraduate dentists and provides a starting point from which consequent studies can be based. The study should now be expanded into different settings, e.g., other dental schools and disciplines (such as medicine, nursing, and veterinary medicine), to confirm and build upon these initial findings.

Chapter 1 - Introduction and literature review

Assessment is an essential component of dental education. It determines whether dental trainees have been trained to the required standards before being permitted onto professional registers, thus preventing the public from being treated by individuals who are not fit to practise dentistry.

For providers of undergraduate dental education, there is a constant challenge to provide regulatory bodies with evidence that their graduates have attained the necessary educational outcomes and are suitable for initial entry onto the professional registers. Many assessment methods have been developed and adopted in dental education and institutions need to ensure they choose methods which are valid.

The following chapter explores assessment within dental education as a narrative literature review and identifies a need for further research on the validity of longitudinal clinical assessment.

1.1 Background

1.1.1 Regulation of healthcare professions

In the UK, multiple regulatory bodies (Table 1.1) set professional standards for the training and conduct of healthcare professionals (HCPs). These regulatory bodies maintain registers of individuals who have met these standards and, therefore, possess the training, skills and experience required to treat members of the public safely and competently (UK Health and Safety Executive, Accessed 2021). Ultimately this protects the public from harm by ensuring medical treatment is only provided by individuals who are deemed fit to practice.

Each regulatory body operates independently of one another and publishes their own educational standards documents. These documents provide educational institutions with a list of learning outcomes (LOs) that must be achieved by those wishing to be admitted onto the respective registers, i.e., they provide a

framework against which educational institutions can assess if their graduates have met the required standards.

Table 1.1 - Regularity bodies for UK healthcare professions (UK Health and Safety Executive, Accessed 2021).

UK healthcare regulatory body	Profession(s) regulated
General Medical Council (GMC)	- Doctors/Medics/Physicians
Nursing and Midwifery Council (NMC)	- Nurses - Midwives
General Dental Council (GDC)	- Dentists - Dental nurses - Dental technicians - Clinical dental technicians - Dental hygienists - Dental therapists - Orthodontic therapists
General Optical Council (GOC)	- Optometrists - Dispensing opticians - Student opticians - Optical businesses
General Chiropractic Council (GCC)	- Chiropractors
General Osteopathic Council (GOsC)	- Osteopaths
General Pharmaceutical Council (GPC)	- Pharmacists - Pharmacy technicians - Pharmacy premises
Health and Care Professions Council (HCPC)	- Arts therapists - Biomedical scientists - Chiropodists/podiatrists - Clinical scientist - Dieticians - Hearing aid dispensers - Occupational therapists - Operating department practitioners - Orthoptists - Paramedics - Physiotherapists - Practitioner psychologists - Prosthetists/orthotists - Radiographers - Speech and language therapists

The specific content of each regulatory body's educational standards document(s) can vary considerably due to differences in priorities, knowledge, skills, and behaviours between each healthcare discipline. Some of these documents simply provide a list of the outcomes that individuals must achieve by the end of their professional training (e.g., Guidance for Osteopathic Pre-Registration Education – General Osteopathic Council (GOsC, 2015); Core Competencies - General Optical Council (GOC, 2016); Education Standards - General Chiropractic Council (GCC, 2017)). Others provide both a list of outcomes and summative descriptions of the standard of graduate expected. For example, the General Medical Council (GMC) state in their "Outcomes for Graduates" (GMC, 2018) that the "overarching outcome for graduates" is that, in accordance with their "Good Medical Practice" guidelines (GMC, 2014):

> "Newly qualified doctors must make the care of patients their first concern, applying their knowledge and skills in a competent, ethical and professional manner and taking responsibility for their own actions in complex and uncertain situations".

This thesis will now focus on the content of the General Dental Council's (GDC's) educational standards document, entitled "Preparing for Practice" (GDC, 2015a), which is of direct relevance to this study.

1.1.2 The General Dental Council's Educational Standards

Like the GMC, the GDC provide a list of the educational LOs as well as a summative description of the required standard for new dental graduates from Bachelor of Dental Surgery (BDS) courses. Their Preparing for Practice (PfP) document (GDC, 2015a) divides 150 LOs across four key domains (clinical, communication, management and leadership, and professionalism) and twelve subdomains (Table 1.2) of competent clinical practice and states that, prior to admittance onto their professional register, those who wish to practise dentistry must be trained to the point of "safe beginner".

Table 1.2 - The General Dental Council's key domains and subdomains for learning outcomes (GDC, 2015).

Domain	Sub-domains
Clinical	Individual patient care - Foundations of practice - Comprehensive patient assessment - Diagnosis - Treatment planning - Patient management - Patient and public safety - Treatment of acute oral conditions - Health promotion and disease prevention - Management and treatment of periodontal disease - Hard and soft tissue disease - Management of the developing dentition - Restoration and replacement of teeth
	Population-based health and care
Communication	Patients, their representatives, and the public
	Team and the wider healthcare environment
	Generic communication skills
Management and leadership	Managing self
	Managing and working with others
	Managing the clinical and working environment
Professionalism	Patients and the public
	Ethical and legal
	Teamwork
	Development of self and others

The GDC describe a "safe beginner" as:

> "A rounded professional who, in addition to being a competent clinician and /or technician, will have the range of professional skills required to begin working as part of a dental team and be well prepared for independent practice. They will be able to assess their

> own capabilities and limitations, act within these boundaries and will know when to request support and advice".

Within this definition, the GDC acknowledge that, despite UK undergraduate dental curricula taking four to five-years of full-time study to complete, they do not expect dental students to be "experts" at the point of graduation. However, they do expect that new dental graduates must have demonstrated attainment of the LOs within each of the four key domains of competent clinical practice to be regarded as "safe beginners" (GDC, 2015a) (Figure 1.1). Following graduation from the BDS course, those who wish to practice independently within the UK NHS must complete a year of postgraduate dental vocational training (DVT). A summary of the UK dental training pathway (for general dental practice) is provided in Figure 1.2.

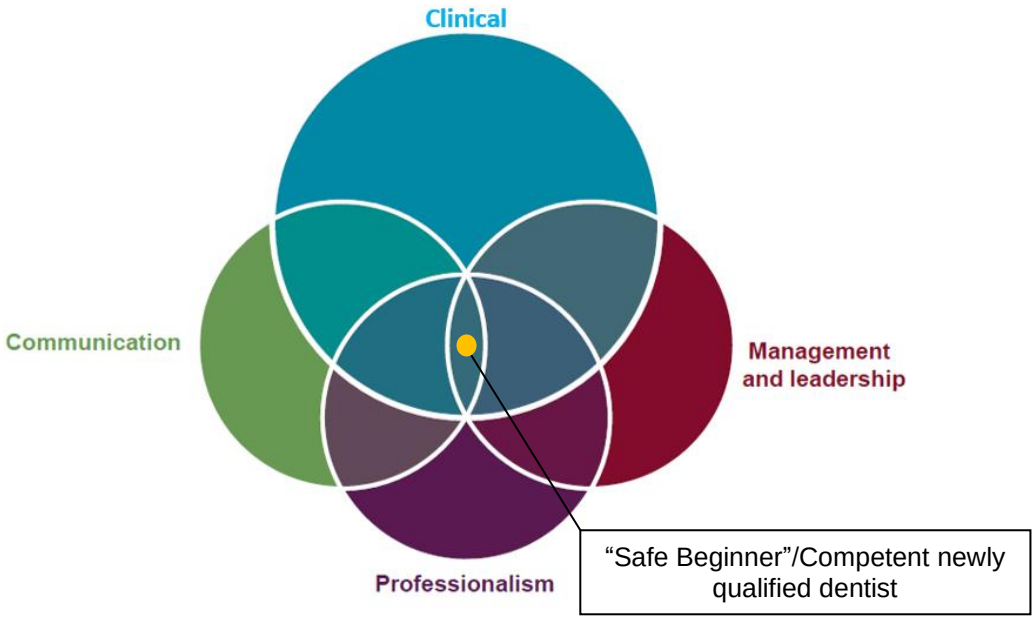


Figure 1.1 - The General Dental Council's four key domains of competent clinical practice. The "safe beginner" must have attained the learning outcomes within each of these domains - Modified from General Dental Council's Preparing for Practice (2015).

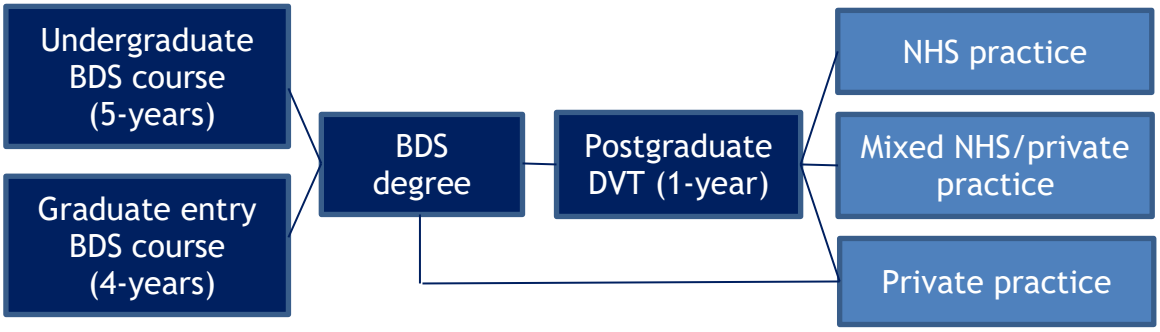


Figure 1.2 - UK dental training pathway (for general practice). BDS = Bachelor of Dental Surgery. DVT = Dental vocational training.

At the time of writing this thesis, there were indications of a forthcoming shift in the terminology used within dental education in the UK. An impending update of the GDC's PfP document (GDC, 2015a) is likely to replace "competence" with "capability". How such a change in terminology would impact understanding and assessment of attainment of the LOs remains to be seen.

This thesis uses the term "competence" as a measure of student performance in accordance with the current version of the GDC's PfP document (GDC, 2015a) and much of the existing literature within dental education. However, studies on dental student assessment conducted after the updated GDC guidance is published will need to take account of any changes to the terminology.

Regardless of terminology, training individuals to the point of "safe beginner" requires appropriate teaching, learning, and assessment that are satisfactory to the GDC. However, the GDC do not stipulate which teaching and assessment methods should be used by educational institutions to prove their undergraduate dental students have achieved all the LOs. Instead, the responsibility of choosing suitable methods lies with the educational institutions themselves.

1.2 Curriculum design

The delegation of responsibility of choosing suitable assessment methods allows dental schools to design their curricula in a manner that suits their individual circumstances providing they are consistent with the GDC's LOs. The GDC monitor and inspect UK dental school curricula and require each school to produce evidence that their students are being assessed appropriately (GDC, 2015a). These inspections reassure the GDC that graduates from UK dental

schools are safe to practise and, therefore, can join their register. As a result, dental schools must ensure that they select good and appropriate assessment methods, continually review them, and seek to adopt best practice in accordance with the available evidence.

Dental schools may also consider how their chosen assessment methods link to the teaching/learning activities (TLAs) they adopt – another responsibility delegated to them by the GDC. How the TLAs relate to the LOs may also be considered. Institutions which opt to establish strong links between the LOs, TLAs and assessment methods demonstrate the use of "constructive alignment" in their curriculum design.

Constructive alignment was originally described by Biggs (1996) who proposed that, in well-designed curricula, there must be continuity between the LOs, TLAs and assessment tasks (ATs) (Figure 1.3) - which are the three key components of a curriculum. The first of these components (i.e., the LOs) ensures students (and assessors) are aware of the educational goals which must be met over the duration of the course. The second (the TLAs) ensures there is a conscious effort to provide students with appropriate teaching methods which encourage them to learn the knowledge and/or skills associated with the LOs. And finally, the third component (the ATs) ensures appropriate evidence is collected to demonstrate that students have attained the LOs.

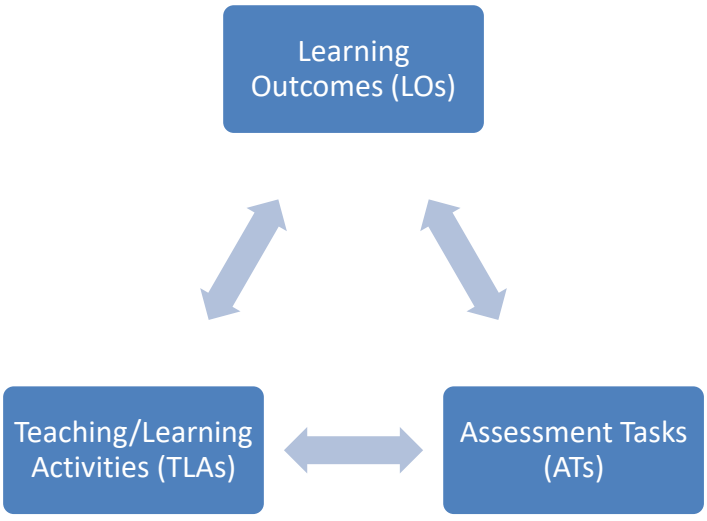


Figure 1.3 - Constructive alignment curriculum design (Biggs, 1996).

A simplified example of constructive alignment within dental education is provided in Table 1.3.

Table 1.3 - A simplified example of constructive alignment within a dental education curriculum. NOTE: Since tooth extraction is a practical clinical skill, practical clinical teaching/learning activities and assessment methods have been selected.

Learning outcome (LO)	Teaching/Learning Activities (TLAs)	Assessment Tasks (ATs)
Extraction of teeth	- Clinical demonstration(s) - Opportunities to perform tooth extractions (simulated and real patients)	- Competence test (see section 1.4.10) - Objective Structured Clinical Examination (OSCE) (see section 1.4.9)

The Association for Dental Education in Europe (ADEE) previously recommended all dental curricula be constructively aligned (ADEE, 2010). However, from the currently available literature, it is not possible to determine how different dental schools plan and design their curricula, and, therefore, information on which dental schools have followed the ADEE's recommendations was not readily available at the time of this study.

1.3 Selecting assessment methods for clinical assessment

1.3.1 Determining the purpose(s) of assessment

Before educational institutions can choose which assessment methods to use, they need to determine the intention(s) of the assessment. As detailed in section 1.1.2, the overarching goal of assessment for UK dental schools is to demonstrate to the GDC that their graduates have attained the necessary LOs to be certified as "safe beginners" who are ready to begin practising dentistry independently and competently. The LOs are categorised across four domains (clinical, communication, management and leadership, and professionalism) and within each domain there are a broad range of attributes and skills competent dental practitioners are expected to attain (GDC, 2015a). As a result, a panel of different assessment methods is required to assess different skills related to clinical practice (van der Vleuten, 1996; van der Vleuten et al., 2010; van der Vleuten et al., 2012; van der Vleuten, 2016) as no single method would be suitable or appropriate for assessing all the GDC's LOs. This thesis focuses on the

assessment of competence in performing practical clinical skills (i.e., attainment of LOs within the GDC's "clinical" domain (GDC, 2015a).

Miller (1990) previously developed a model to summarise various aspects of clinical competence in medical education. This model, known as Miller's four-tiered triangle of clinical competence or, simply, Miller's pyramid/triangle (Figure 1.4), is frequently cited within literature on medical education subjects (including dental education) (Williams et al., 2015). As well as illustrating different elements of clinical competence, Miller's model can help evaluate the progressive development of attributes and skills required for clinical competence.

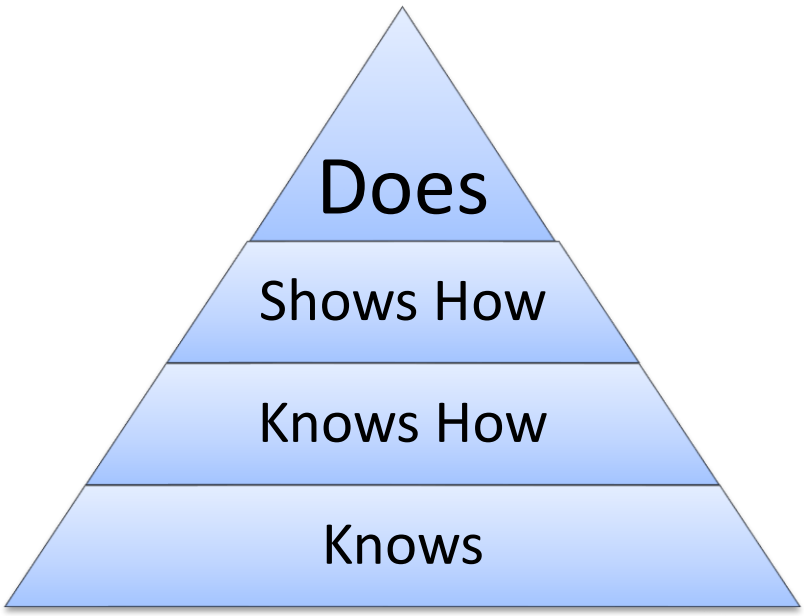


Figure 1.4 - Miller's Triangle of Clinical Competence (Miller, 1990).

Miller's triangle also overlaps Bloom's taxonomy for learning, teaching and assessing (Bloom et al., 1956; Bloom, 1984). This taxonomy was originally developed to illustrate and classify different cognitive processes (Figure 1.5) so that LOs (and therefore teaching and assessment methods) could be aligned with them. However, Bloom (1984) recognised not all LOs relate to cognitive processes and that some could be considered as "psychomotor", which is particularly relevant for teaching, development and assessment of practical clinic skills. This led to a modification of Bloom's taxonomy by Dave (1970) which took the learning and assessment of psychomotor skills into consideration (Figure 1.6).

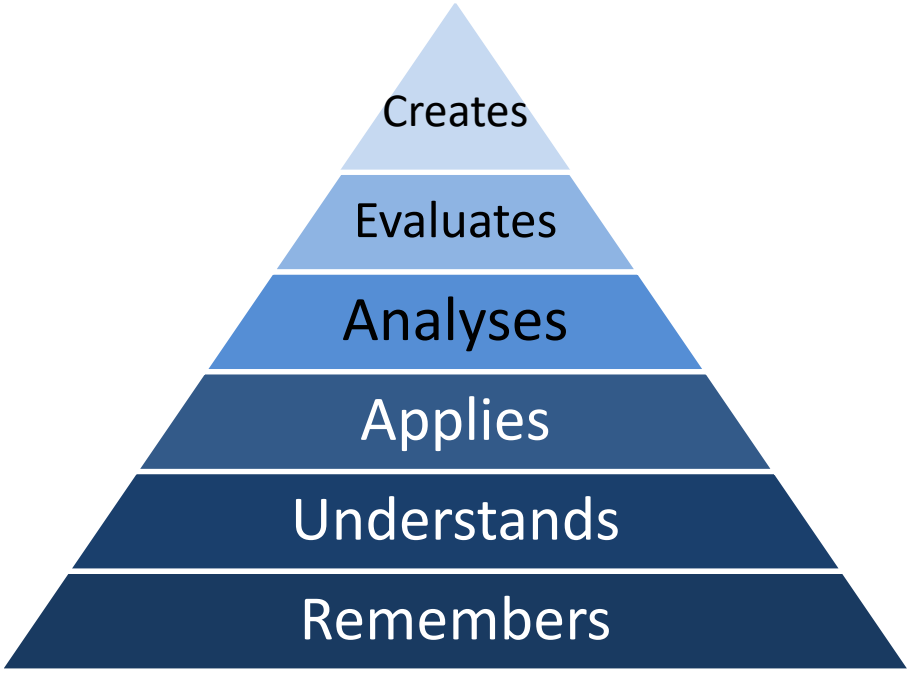


Figure 1.5 - Bloom's taxonomy for learning, teaching, and assessing (Bloom et al., 1956). Revised by Anderson et al. (2001).

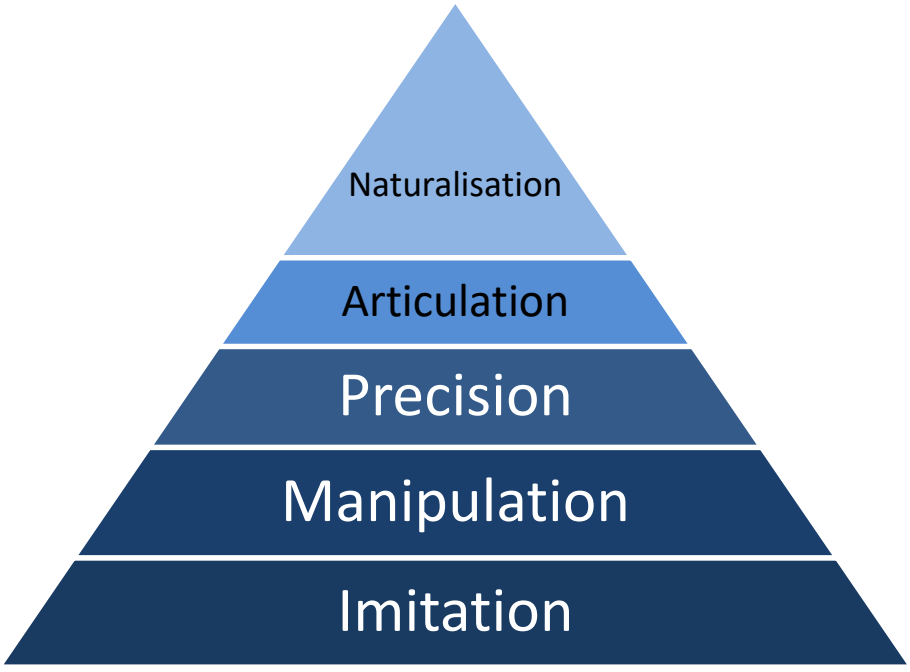


Figure 1.6 – Dave's (1970) revision of Bloom's taxonomy to consider learning, teaching, and assessment of psychomotor skills.

However, within dental education literature, Dave's modification of Bloom's taxonomy is rarely cited when discussing assessment methods. In contrast, Miller's triangle is regularly referred to – possibly because it appears easier to align assessment methods used within dental education with each of the four tiers of Miller's triangle.

At the first level of Miller's model ("Knows"), students can factually recall and comprehend information out with the context of patient care. At the second level ("Knows How"), students can problem solve and make decisions on patient care through applying their knowledge in the context of written assignments/assessments and/or simulated clinical scenarios. Students should also be able to explain, in their own words, the cause(s) and progression of basic disease processes. At the "Shows How" level, students should be able to apply practical patient care skills in real-life situations or simulated clinical settings where patient interaction is incorporated, thus making them as close to real-life medical working environments as possible. Students performing at this level can demonstrate they are able to work as a HCP in a controlled, well-supervised environment. The fourth and final level ("Does") determines whether students have demonstrated the fundamental competencies necessary for unsupervised practice and can consistently reproduce these skills to the standard(s) expected over time. Ideally, assessors in dental education want students to progressively develop towards the "Does" level so they can be confident students have become "entrustable".

This degree of confidence in a student's clinical abilities is important not only for education, but also for patient care, as faculty must be sure students are able to perform key clinical activities with reasonable chances of success. The need to assess whether students have developed to this point has resulted in "Entrustable Professional Activities" (EPAs) becoming increasing prevalent within medical education subjects (Pittenger et al., 2016; Chesbro, Jensen and Boissonnault, 2018; Duijn et al., 2019; Lau et al., 2020; Tonni et al., 2020).

EPAs can be described as units of practice (or tasks) that students can be expected to perform independently once they have demonstrated they can perform them competently (Ten Cate and Taylor, 2020). Their concept suggests the level of supervision has an inverse correlation with student competence (Ten Cate, 2013). Therefore, when competence is achieved, no or very little supervision is required.

EPAs must focus on tasks routinely faced by clinicians in daily practice. Each patient encounter requires performance of multiple skills, and ideally this should be reflected in EPAs, i.e., instead of students being assessed against lists of

individual competencies, skills which are typically performed together are assessed in tandem. Therefore, due to their nature, workplace-based assessment (WBA) methods (e.g., Direct Observation of Procedural Skills (DOPS) (see section 1.4.10) and longitudinal assessment (see section 1.4.11)), can serve as means for assessing EPAs (Mulder et al., 2010). These assessment formats are typically aligned with the "Does" level of Miller's triangle (see section 1.3.1).

However, although there are authors who advocate the use of EPAs for medical education subject curricula, there is currently little evidence on their practical application (Ten Cate and Taylor, 2020). This may explain why some authors within dental educational literature still recommend Miller's triangle as a basis for planning assessment (Patel et al., 2018). However, there are also no readily available data to suggest which, if any, dental schools follow this specific approach. Regardless, Miller's model does highlight how the intention(s) of an assessment must be clarified to help identify which assessment methods are suitable for a particular purpose. For example, if the purpose was to assess student knowledge and understanding (i.e., the "Knows" and "Knows How" levels of Miller's model) then written tests could be used for this purpose. If the purpose was to assess student ability in performing a clinical procedure on a patient, then a more practical assessment format would be required (see example provided in Table 1.3. in previous section (1.1.2)). Further discussion on assessment methods and how they can be aligned with the four tiers of Miller's model is provided in section 1.3.1.

Once the assessment purpose(s) have been specified, dental schools can then identify which assessment methods may be suitable and evaluate their utility to ensure they are using the best available methods.

1.3.2 Factors influencing choices

Within the literature, there is heterogeneity on which factors are taken into consideration when assessment methods are being evaluated and selected. Examples include discriminatory power (Kline, 2000), utility, acceptability, educational impact (van der Vleuten, 1996; van der Vleuten and Schuwirth, 2005), defensibility (Hecker and Violato, 2009), costs (Brown et al., 2015) and ease of implementation (Jolly and Dalton, 2019). The broad range of factors

reported may be due to variation in individual circumstances between subject areas and educational institutions. However, within educational subjects, a commonly used means of determining whether an assessment method is "good" (i.e., fit for purpose) is psychometric testing/psychometrics.

Psychometrics is an area of study concerned with measuring the characteristics of reliability and validity. Typically, assessments which are considered to be "good" possess high reliability and high validity (Kline, 2000).

Section 1.4 of this chapter reviews the literature on assessment methods which are commonly used within dental education. It will primarily focus on their reliability and validity since these psychometric properties are consistently considered across multiple dental educational publications which investigate, review, and compare dental education assessment methods. However, other assessment properties will be acknowledged for publications which discuss the strengths and weaknesses of an assessment method based on other factors. The remainder of this section (1.3) describes reliability and validity in further detail and how these psychometric properties may be measured.

1.3.3 Psychometric properties and their use in evaluating assessment methods

1.3.3.1 Reliability of assessments

Reliability refers to how reproducible the results of an assessment are (Schuwirth and van der Vleuten, 2014). Highly reliable assessments are likely to produce the same or similar results each time they are used. This is traditionally investigated by performing a "test-retest analysis" - i.e., having the same candidate(s) repeat the exact same assessment. The two sets of scores gathered through "test-retest analysis" should display good correlation if the assessment is reliable (Kline, 2000).

However, in many cases it may not be possible to run a repeat test. As an alternative approach, psychometrists might opt to retrospectively split a test into two halves and treat one half as the initial "test" and the other as the "re-test". Many prominent reliability measurements which use mathematical models, such as Kuder–Richardson (KR-20) (Kuder and Richardson, 1937) and

Cronbach's alpha (Cronbach, 1951), are based on this approach. Both Kuder-Richardson and Cronbach's alpha produce reliability coefficients, however there is no consensus on what their value should be for an assessment to be considered reliable. Some authors suggest they should be $\geq$0.7 (Gravetter and Wallnau, 2000), whereas others propose they should be higher (0.8-1.0) (Keynan, Friedman and Benbassat, 1987; Reznick et al., 1997; Bould, Crabtee and Naik, 2009), especially if the assessments are for "high-stake" purposes, such as certification. Alternatively, some studies have measured reliability through correlation coefficients, such as Pearson's (r) (Beanland et al., 1999; Polit and Hungler, 1999; Gravetter and Wallnau, 2000; Al-Osail et al., 2015) and Spearman's rank (Al-Osail et al., 2015). In general, correlation coefficients $\geq$0.7, $\geq$0.8 and $\geq$0.9 are accepted as indications of acceptable, good and excellent reliability, respectively (Karras, 1997), however their statistical significance (p-values) may be influenced by a variety of factors (such as sample size) and therefore should not be taken at face value and solely relied upon. Instead, there is a need for triangulation with other data sources to confirm reliability.

Other difficulties in determining test reliability may relate to "real changes", the timing of testing and test length/number of assessment items. A "real change" could be when candidates have demonstrated progress by improving their knowledge/skills between the test and the retest - therefore, it is important to know how much time has lapsed between the two. The time at which a test takes place could be linked to additional factors which affect performance such as candidate mood/state of mind/health status and the conditions under which the assessment is taken (Kline, 2000). Tests with more items (i.e., longer tests) have been shown to be more reliable (Nunnally, 1978); however, if the test becomes too long then candidate boredom and fatigue could become a factor which adversely affects its reliability (Kline, 2000).

Therefore, adequate investigation of a test's reliability requires sufficiently large sampling which considers as many potential sources of error (e.g., assessment items, test conditions etc.) as possible (Schuwirth and van der Vleuten, 2014). The desired outcome of these investigations is for a test to display high reliability. However, taken in isolation, high reliability may not

necessarily indicate that a test is good since it is possible for a test to be reliable but not valid (Beanland et al., 1999; Polit and Hungler, 1999; Kline, 2000).

1.3.3.2 Validity of assessments

Validity refers to how well an assessment measures what it proposes to measure (Kline, 2000). Much like a ruler acting as a valid method for measuring length (not weight or speed), assessments can only be valid for a defined purpose and therefore, when discussing validity, it is important to clarify the intention of an assessment from the outset. Although defining the intended purpose may seem simple, proving an assessment measures what it intends to measure can be challenging since there is no single numerical index that can be used to test validity (Kline, 2000). Instead, establishing assessment validity requires evidence to be collected from a variety of sources and perspectives (Schuwirth and van der Vleuten, 2014).

The investigative approaches used signify what type(s) (and subtypes(s)) of validity can be attributed to an assessment method and, therefore, the strength of evidence available for its validity. Various publications (Kline, 2000; Hecker and Violato, 2009; Kane, 2013) have described the four main types of validity as:

1. Face validity.
2. Content validity.
3. Criterion validity (includes concurrent and predictive subtypes).
4. Construct validity.

Face validity

Face validity denotes if an assessment appears to measure what it intends to (Kline, 2000; Hecker and Violato, 2009). It is established through superficial subjective opinion and, as a result, is not considered to be a strong form of validity (Kline, 2000). However, it can make assessments appear reasonable to those undertaking them (Kline, 2000) and, therefore, may determine if an

assessment will be taken seriously (Hecker and Violato, 2009). Face validity could also be important in deciding whether further investigation of the validity of an assessment method is merited.

Content validity

This form of validity is sometimes referred to as "direct validity" (Schuwirth and van der Vleuten, 2019). It describes how well an assessment represents what it aims to measure. For example, an assessment on root canal treatment should not just contain items on tooth/root canal anatomy, but should also include other necessary and relevant items such as instruments and equipment, canal preparation techniques, dental materials etc. To ensure an appropriate sample of items is selected, assessments are usually drawn up against a blueprint. The blueprint acts as a template against which assessors can select relevant items in relation to the subject or category being tested (Hopkins, 1998; Hecker and Violato, 2009; Roudsari, 2017; Schuwirth and van der Vleuten, 2019).

It is also worth noting that, along with subject matter, the cognitive process being assessed should also be clarified when determining content validity (Hecker and Violato, 2009). Cognitive processes usually refer to the levels of Bloom's taxonomy (Bloom et al., 1956) (see Figure 1.4 in section 1.3.1), but are also applicable to the first two levels Miller's triangle ("Knows" and "Knows How") (Miller, 1990) (see Figure 1.3 in section 1.3.1).

Once the purpose of the assessment has been determined, appropriate items (set against a blueprint) have been selected and the cognitive (or practical) processes being examined are defined, assessments can then be evaluated by subject experts to establish content validity (Schuwirth and van der Vleuten, 2019). If there is agreement between a panel of experts that the assessment adequately assesses what it aims to assess, then the assessment is said to have content validity.

Criterion validity

Criterion validity is the extent to which the results of an assessment relate to other outcome measures – ideally those which are considered to be "gold

standard" or have previously been validated. It is divided into two subtypes: concurrent and predictive. The former (concurrent) refers to how an assessment correlates with other forms of assessment which are used for the same purpose (Kline, 2000; Hecker and Violato, 2009) and are already considered valid (Fink, 2010) or "gold standard" (Stokes, 2011; Prince, 2012; Bellamy, 2015). For example, an assessment employed to test the anatomical knowledge of first year dental students should correlate with another (valid) assessment that has also been used to test the same criteria in the same group of students. The latter (predictive) refers to how current student performance forecasts future performance (Kline, 2000; Hecker and Violato, 2009). An example would be using secondary school examination results to predict performance in higher education.

Determining both concurrent and predictive validity requires acquisition of robust data (Hecker and Violato, 2009). Such data could be acquired through statistical analysis, e.g., correlations between the assessment method under investigation and established valid assessment methods (Kline, 2000; Hecker and Violato, 2009; DePoy and Gitlin, 2016). However, investigations could be challenging if there is a general lack of "gold standard" assessment methods to compare against or the perceived "gold standard" has insufficient evidence to support its own validity (Bellamy, 2015).

Construct Validity

Before construct validity (occasionally referred to as "indirect validity" (Schuwirth and van der Vleuten, 2019)) can be understood, it is necessary to define what is meant by a "construct". A construct is a psychological quality or concept that cannot be observed directly but is suspected to exist (MacCorquodale and Meehl, 1948; Rowntree, 1987; Hecker and Violato, 2009; Schuwirth and van der Vleuten, 2019). A typical example of a construct would be intelligence (Schuwirth and van der Vleuten, 2019), and examples within educational and other health sciences would be communication and professionalism (Hecker and Violato, 2009). In terms of this study, competence would be considered as a construct (as would capability – should it become the preferred terminology used by the GDC and UK dental schools).

Construct validity concerns whether an assessment correctly reveals the construct(s) being measured (Stuart, 2007; Hecker and Violato, 2009). Schuwirth and van der Vleuten (2019) provide an example on assessment of problem-solving skills within medicine, stipulating that those with good problem-solving skills will perform better than those with poorer problem-solving skills in an assessment with good construct validity, i.e., the individuals being assessed would "behave" as (hypothetically) anticipated.

It has been suggested there is no single best way to investigate construct validity and, in most cases, evidence from a variety of sources and perspectives is required. For example, those investigating construct validity may build evidence through content analysis, correlation studies, factor analysis, analysis of variance (ANOVA) studies, intervention studies, factor analysis, multi-trait/multi-method studies, etc. (Brown, 2000). The more evidence gathered for the various other types of validity (face, content, and criterion), the greater the support for construct validity.

The various types of validity for assessment methods have resulted in a range of viewpoints within the literature. For example, some authors believe that construct validity is the most important type and other forms (such as face and content validity) should be discounted on the basis that they are not supported by sufficient evidence (Downing and Haladyna, 2004). Other authors (such as Kane (2006)) adopt a more universal approach and advocate that although there are different types of validity, they collectively contribute evidence for investigating the validity of an assessment for a defined purpose (Schuwirth and van der Vleuten, 2019). This approach to establishing validity is known as building a "validity argument" and is explored further in the following section (1.3.3.3) and is further related to the work of this study in section 1.5.

It is also worth noting that for an assessment to be valid, it must be reliable (Beanland et al., 1999; Polit and Hungler, 1999). However, although reliability is necessary, it does not constitute a sufficient component of validity (Feld and Brennan, 1989; Downing, 2003), i.e., a valid assessment does not need to have high reliability, but it does need to be generally reliable. If an assessment had high validity but no/little reliability, then it would assess what it intends to but would very inconsistent and therefore its outcomes could not be trusted.

1.3.3.3 Building evidence for validity

The concept of providing a body of evidence that evaluates if appropriate interpretations are being made from assessment scores has previously been described as a “validity argument” or an “argument-based approach to validation” (Cronbach, 1988; Kane, 1992; Shepard, 1993; American Educational Research Association, American Psychological Association and National Council on Measurement in Education, 2004).

Kane (2006) explored this concept in depth and suggested that (as previously discussed in section 1.3.1) the proposed interpretations and uses (i.e., the purpose) of an assessment must be made clear if we are to begin scrutinising the conclusions and decisions generated from it. In a subsequent publication, Kane (2013) stipulated any inferences and assumptions that influence the proposed interpretations and uses need to be developed as an “interpretation/use argument (IUA)”. Once an IUA has been created, it provides “a framework for validation that defines the claims that need to be checked”, i.e., we evaluate how well the evidence supports the IUA inferences and assumptions (Cronbach, 1971; Messick, 1989; Kane, 2006; Moss, 2007) (Figure 1.7).

Figure 1.7 - Argument-based validation process.

The amount of evidence required depends on how ambitious the assessment purpose is. Highly ambitious interpretations (e.g., those involved in high stakes decisions) require more evidence and justification compared to less ambitious ones, however the evidence gathered must relate to the purpose of the assessment. For example, if there was no intention for an assessment to be used as a means of predicting future performance, then there would be no point in obtaining evidence to support that it can be used in this manner. It is also worth noting assessment interpretations and uses (and therefore the IUA) can evolve over time as new evidence becomes apparent (Kane, 2013).

1.4 Clinical assessment methods within dental education

1.4.1 Levels of clinical competence assessment

A variety of assessment methods have been developed and implemented within dental education. The range of assessment methods used by dental schools in the USA was previously investigated by Albino et al. (2008) using the results of an online survey conducted by the American Dental Education Association (ADEA). The survey received responses from 931 members of faculty across fifty-three (out of fifty-six) USA dental schools and identified seventeen different assessment methods had been used.

Albino et al. (2008) subsequently aligned these seventeen methods against the four levels of Miller's triangle of clinical competence ("Knows", "Knows How", "Shows How" and "Does") to highlight how the spectrum of learning should be considered when designing or selecting appropriate assessment methods. Williams et al.'s "A Guide to Assessment in Dental Education" (2015) also categorises a variety of assessment methods according to the four levels of Miller's triangle. The process(es) and/or rationale behind these alignments are not described by either publication, therefore it is possible that some methods may have been miscategorised if the authors were unfamiliar with how each dental school had used the assessment methods. Despite this lack of clarity, aligning assessment methods against the four tiers of Miller's triangle suggests which methods may be suitable for tracking the development of clinical competence in students at various stages of their training.

Between Albino et al. (2008) and Williams et al.'s (2015) publications, a total of twenty-four assessment types were listed (Table 1.4).

Table 1.4 - Assessment methods in dental education. Compiled from Albino et al. (2008) and Williams et al. (2015).

Assessment methods in dental education	
- Multiple-choice questions (MCQs) - Modified essay questions (MEQs) - Extended matching questions (EMQs) - Short answer questions (SAQs) - 'Spotter' tests - [Traditional] essays - Oral examinations (Viva /Viva Voce) - Triple jump exercises (TJE) - Objective structured clinical examination (OSCE) - Clinical or laboratory simulated practical tests - Clinical competency examinations/Direct observation of procedural skills (DOPS)	- Case-based discussions (CBD) - Multi-source feedback (MSF) - Script concordance test (SCT) - Mini-Clinical evaluation exercises (mini-CEXs) - Longitudinal [clinical] assessment - Dental evaluation of performance tests (ADEPTs) - Unit requirements and daily evaluations - Chart-stimulated evaluation - Portfolios - Critical appraisal - Student reports - Computer-based simulations - Student self-assessment

Williams et al. (2015) proposed that some assessment methods could be aligned with both the "Knows" and "Know How" levels of Miller's triangle depending on how the assessment questions are formatted. For example, if MCQs are formatted in a manner which assesses simple factual recall, then it will be more aligned with "Knows". By comparison, MCQs which are formatted to assess application of knowledge and evaluation on information result in an assessment will be more aligned with "Knows How".

Despite Williams et al.'s suggestions for aligning some assessment methods with the two "lower" tiers of Miller's, there was a general consensus between both publications on assessment methods which align with each of the two "higher" tiers ("Shows How" and "Does"). For example, both associated OSCEs and clinical or laboratory simulated practical tests with "Shows How", and clinical competency tests/DOPS, portfolios, and longitudinal assessment with "Does".

At this stage, it is worth noting that any form of assessment repeated over time could be described as "longitudinal". However, in both Albino et al. (2008) and Williams et al. (2015) (and within this thesis), "longitudinal assessment" refers to practical clinical assessments carried out during routine patient care on a

regular basis to assess students' clinical skills. This also explains why Albino et al. (2008) and Williams et al. (2015) aligned longitudinal [clinical] assessment with the "Does" level of Miller's triangle.

Figure 1.8 combines the alignments of the twenty-four assessment methods listed between Albino et al. (2008) and Williams et al. (2015) against the four-tiers of Miller's triangle of clinical competence. Assessment methods which (according to Williams et al. (2015)) can potentially belong to either the "Knows" and "Knows How" levels are highlighted.

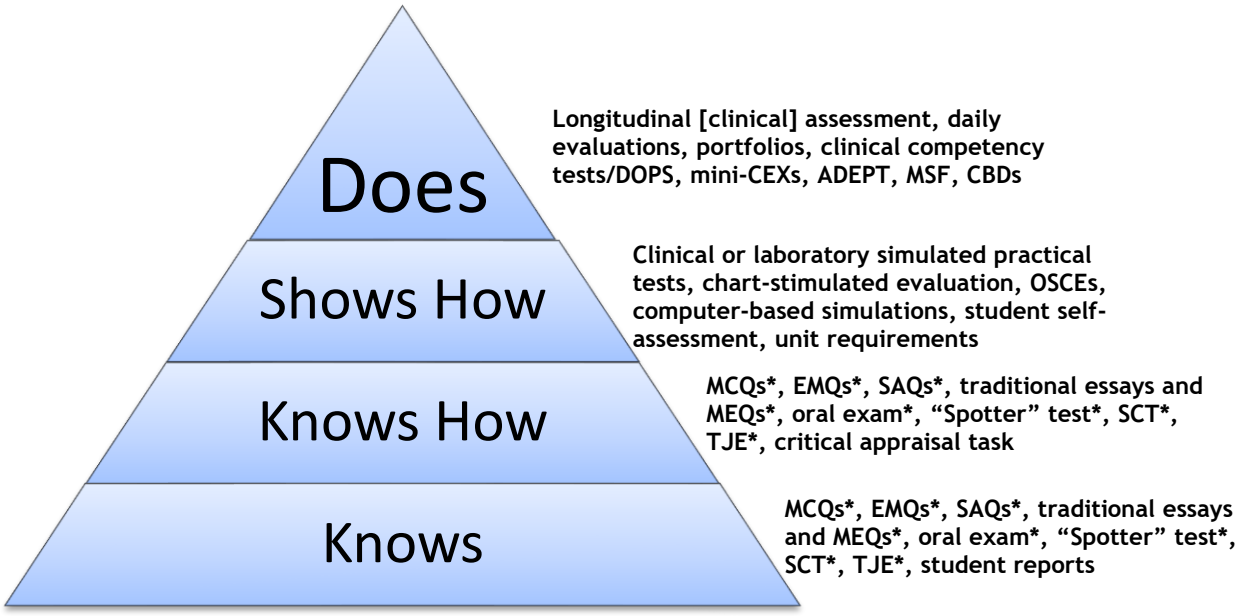


Figure 1.8 - Distribution of twenty-four assessment methods within dental education against Miller's triangle of clinical competence (combined from Albino et al. (2008) and Williams et al. (2015)). Methods marked with an * could be aligned with both the "Knows" and "Knows How" tiers.

Both publications highlight the range of assessment methods which have been adopted within dental education. However, these two publications do not cover all methods which may (or have been) used. For example, Roudsari (2017) presents reflective exercises, true-false question examinations, role play, situational judgement tests (SJTs) and case studies as other possible forms of assessment used by dental schools.

A review produced by van der Vleuten and Verhoeven (2013) also provides a modified version of Miller's triangle (summarised in Table 1.5). However, instead of aligning specific assessment methods with each level of Miller's, van der Vleuten and Verhoeven provide an overview of the nature and format of

assessments at each level regardless of which specific method is used. For example, their outline of "Knows How" (see Table 1.5) assessments resonates with each of the assessment types aligned with this level by Albino et al. (2008) and Williams et al. (2015) (i.e., case-based MCQs, essays, oral exams, critical appraisal tasks, and TJEs).

Table 1.5 - Summative descriptions of assessment methods associated with each level of Miller's triangle of clinical competence according to van der Vleuten and Verhoeven (2013).

Level of Miller's triangle of clinical competence	Description of assessment methods associated with each level
"Does"	Assessment of habitual clinical performance in the authentic working context by professionals (including other healthcare professionals, patients, and the self).
"Shows How"	Assessment of clinical performance in standardised simulated performance situations by trained professionals (including simulated patients).
"Knows How"	Written, computer-based or oral assessment that test for factual knowledge that assess application of knowledge (usually scenario-based).
"Knows"	Written, computer-based or oral assessments that test for factual knowledge.

Within the same publication, van der Vleuten and Verhoeven (2013) stress that - due to the limitations and weaknesses associated with each method - there is no single form of assessment which adequately measures all the attributes expected of HCPs, or assesses all levels of Miller's triangle of clinical competence. As a result, good assessment programmes should use multiple assessment methods from all levels of Miller's triangle.

The need for multiple forms of assessment within dental education has been acknowledged by the ADEE (2010) and is imperative for "programmatic assessment" (van der Vleuten et al., 2012) - an approach which is becoming increasingly prevalent in medical education subjects (Dannefer et al., 2005; Fishleder, Henson and Hull, 2007; Schuwirth, van der Vleuten and Durning, 2017).

1.4.2 Programmatic assessment

Programmatic assessment seeks to optimise both “assessment for learning” (e.g., formative assessment) and “assessment of learning” (e.g., summative assessment) for the development of competence. The former is enhanced by using individual assessments to drive student learning and permit feedback, and high-stake decisions associated with the latter are based on information generated through aggregation of all the assessments (Driessen et al., 2012; van der Vleuten et al., 2012; Schuwirth, van der Vleuten and Durning, 2017; Norcini and Zaidi, 2019; Schuwirth and van der Vleuten, 2019). However, although the use of multiple forms of assessment has been encouraged, using too many different formats could lead to problems with “assessment literacy”.

Assessment literacy refers to students (and assessors) knowing about the rules, processes, benefits, and limitations of an assessment method. If students become familiar with how an assessment functions and what it is measuring, they can develop a greater understanding on how to evaluate their own performances. In turn, this can lead to students having a greater understanding of their own learning and how they might take control of it (Price et al., 2012).

Issues with assessment literacy can arise when many different modes of assessment are used, and students don’t get a chance to become familiar with any one type. By contrast, subjecting students to a few different methods, in a formative setting, allows them to become familiar (i.e., “literate”) with these assessments. Once students become “assessment literate”, they can progress their learning and development further since they are able to make the most of formative feedback by having a clearer understanding of what the assessment is testing and how feedback on their performance relates to achievement of the LOs being tested (Price et al., 2012). Assessors should also be assessment literate (Webb, 2002) so they can identify high quality assessment methods, implement them appropriately, and demonstrate “good assessment practices” (Price et al., 2012).

The Higher Education Academy (HEA) have advocated assessment literacy as “essential” to everyone involved in assessment (HEA, 2012). However, although the concept of assessment literacy makes sense from a theoretical perspective,

there is little robust evidence to suggest it directly improves student learning. Smith et al. (2013) serves as one of the few studies which has investigated its impact and concluded that increasing student assessment literacy improved their learning. Regardless, if assessment literacy is accepted as a "foundation" of assessment *for* learning (as described by Price et al. (2012)), then a stronger argument supporting its impact on student learning can be made since assessment *for* learning is an initiative which is now widely accepted within educational literature (Schuwirth and van der Vleuten, 2004a; McLachlan, 2006; Wormald et al., 2009; Deeley and Bovill, 2017). However, there is currently no readily available evidence which demonstrates how assessment literacy influences assessors when selecting assessment methods for curricula.

As highlighted by Figure 1.8 (section 1.4.1), there are a wide range of assessment methods from which assessors within dental education could choose. Roudsari (2017) recently conducted a survey of summative assessment methods used by UK dental schools. This survey invited fourteen UK dental schools to participate and requested details on which assessment methods were used within each year of their BDS curricula. Responses were received from nine institutions and the results revealed that, between them, the most common assessment methods (i.e., those used by ≥10% of BDS years across the nine responding dental schools) included MCQs, EMQs, SAQs, "Spotter" tests, [traditional] essays and MEQs, reflective write-ups, project presentations, poster presentations, oral exams (unseen or seen), OSCEs, skills tests, DOPS/competency tests, portfolios, logbooks, and longitudinal clinical assessment (Table 1.6). However, it is worth noting that the popularity of the methods presented within this survey may be misleading if some UK dental schools were using the same assessment methods for each BDS year whereas others only adopted certain methods for a select number of BDS years (or not at all).

Table 1.6 - Results of a survey on summative assessment methods used by UK dental schools (adapted from Roudsari, 2017 and permitted for presentation within this thesis by Roudsari). Methods used by the University of Glasgow Dental School as part of their end year of BDS examinations are marked with an *.

Assessment method	Percentage of BDS years using assessment method across the nine responding UK dental schools
MCQ/SBA*	93%
Essay/Modified essay/Assignment*	88%
DOPS/Competency assessment*	68%
SAQ/MSA*	65%
OSCE*	58%
Longitudinal clinical assessment*	53%
Reflective write up	38%
Project presentation	38%
Skills test	30%
Portfolio	25%
Unseen oral exam	23%
Logbook	23%
Poster presentation*	23%
EMQ	20%
Spotter*	20%
Seen oral exam	18%
Oral exam*	15%
Long case/Case study	5%
Role play	5%
SJT	2.5%
True-False	2.5%
Discussion forum	0.0%

The subsequent sections of this chapter (1.4.3-1.4.11) will discuss the summative assessment methods which are used by the University of Glasgow Dental School to assess knowledge and clinical skills - and are directly relevant to this thesis - in more detail. An overview of the summative assessment methods used by the University of Glasgow Dental School for each year of the curriculum is provided in Table 1.7.

Table 1.7 - Assessment methods used by the University of Glasgow Dental School per BDS year. MCQ = Multiple-choice question. SAQ = Short answer question. MSA = Multiple-short Answer. OSCE = Objective structured clinical examination.

BDS year	Assessment methods
1	- Summative assignment (traditional essay) - MCQ - SAQ/MSA - OSCE*
2	- MCQ - SAQ/MSA - OSCE
3	- Anatomy examination (Spotter) - MCQ - SAQ/MSA - OSCE - Longitudinal clinical assessment
4	- SAQ/MSA - Clinical case presentation (oral and poster presentation examination) - Longitudinal clinical assessment
5	- OSCE - Longitudinal clinical assessment

*BDS1 sit a structured clinical examination which is a mix between a spotter test and an OSCE. For this thesis, it will simply be referred to as the BDS1 OSCE.

1.4.3 Multiple Choice Questions (MCQs)

MCQs are one of the most prominent assessment methods within medical education subjects (Grainger et al., 2018; Javaeed, 2018; Abdus et al., 2020) – including dentistry (Albino et al., 2008; Williams et al., 2015; Roudsari, 2017). However, the term "MCQ" may refer to a variety of assessment designs. Some MCQ assessments provide "lead-in" questions, statements (or stems) followed by a list of possible answers options from which students choose one answer. This format is known as "single best answer" (SBA) (Williams et al., 2015; Jolly and Dalton, 2019) and, according to Roudsari (2017), is the format used by the majority of UK dental schools when referring to MCQs.

Other formats include True/False style questions (Williams et al., 2015; Jolly and Dalton, 2019), sentence completion, asserted reasoning, negative marking (True-False-Abstain), elimination scoring, and confidence scoring (Williams et al., 2015). However, Case and Swanson (2001) previously recommended that the use of these formats should be avoided due to the problems associated with them. Examples of such problems included an increased chance of students guessing the correct answers, difficulties in distinguishing between correct and incorrect answers, and ambiguities that cannot be easily clarified (Case and Swanson, 2001). As a result, the use of these formats appears to be diminishing (Williams et al., 2015; Jolly and Dalton, 2019) - unlike SBAs, which remain a popular choice (Albino et al., 2008; Williams et al., 2015; Roudsari, 2017).

Within medical education subjects, MCQs are typically used to assess knowledge/factual recall (Considine, Botti and Thomas, 2005; Vanderbilt, Feldman and Wood, 2013; Williams et al., 2015; Gerhard-Szep et al., 2016; Patel et al., 2018), hence their association with the "Knows" and "Knows How" levels of Miller's triangle of professional competence (Albino et al., 2008; Williams et al., 2015) (see section 1.3.1). Precisely which level they associate with depends on how the assessment questions have been constructed. MCQ items which encourage students to recall facts will assess at the "Knows" level, whereas items which are presented as clinical scenarios can assess "Knows How" (Case and Swanson, 2001; Schuwirth et al., 2001; van der Vleuten et al., 2010; Williams et al., 2015; European Board of Medical Assessors, 2017; Scully, 2017; Patel et al., 2018). Although devising items which test higher orders of thinking

(i.e., “Knows How”) can be challenging (Bridge et al., 2003; Abdulghani et al., 2015; AlFaris et al., 2015; Scully, 2017), the process may be aided by using guidelines on MCQ writing - of which there are many examples (Case and Swanson, 2001; Medical Council of Canada, 2010; European Board of Medical Assessors, 2017; Abdus et al., 2020; Joint Commission on National Dental Examinations, 2020) – and training (Abdulghani et al., 2015; Abdulghani et al., 2017; Dellinges and Curtis, 2017; Tenzin, Dorji and Tenzin, 2017).

Advantages of MCQs highlighted within the literature include:

- Cost-effectiveness (Medical Council of Canada, 2010; Williams et al., 2015).

- Objective scoring (Kemp, Morrison and Ross, 1994; Newstead and Dennis, 1994; Kniveton, 1996; Considine, Botti and Thomas, 2005; Collins, 2006; Escudier et al., 2011; Tarrant and Ware, 2012; Brame, 2013; Sam et al., 2016) and the reduction of inter-examiner marking variability (Coughlin and Featherstone, 2017).

- Efficiency – i.e., MCQs permit a broad range of knowledge across multiple subject areas to be assessed over a short period of time (Schuwirth and van der Vleuten, 2003; McCoubrie, 2004; Considine, Botti and Thomas, 2005; Collins, 2006; Escudier et al., 2011; Tarrant and Ware, 2012; Williams et al., 2015; Sam et al., 2016; Javaeed, 2018), which, therefore, facilitates coverage of a blueprint of LOs (Williams et al., 2015; Coughlin and Featherstone, 2017).

- Computer/machine/digital marking – which makes scoring simple and quick (Morrison and Free, 2001; Epstein, 2007; Williams et al., 2015; Coughlin and Featherstone, 2017; Jolly and Dalton, 2019) and facilitates collation of results and feedback to students.

Disadvantages include:

- The need for time consuming question writing and standard setting/quality assurance processes (Collins, 2006; Williams et al., 2015).

- Difficulties in writing good quality questions (Collins, 2006; Tarrant, Ware and Mohammed, 2009; Williams et al., 2015) - especially if five conceivable answers (i.e., one correct answer and four “distractor” answers) are to be listed per question (Tarrant, Ware and Mohammed, 2009).

- Cueing (i.e., the correct answers can be worked out by eliminating those that are obviously incorrect, and/or the correct answers can be recognised without candidates knowing the fact in question) and guesswork (Case and Swanson, 2001; Downing, 2002; Collins, 2006; Memon, Joughin and Memon, 2010; Williams et al., 2015; Sam et al., 2016; Jolly and Dalton, 2019).

- Encouragement of superficial factual learning and regurgitation rather than deep approaches to learning (van der Vleuten, 1996; Scouller, 1998; Cobb et al., 2013; Williams et al., 2015).

- Limited scope for student feedback to aid learning and development (Williams et al., 2015).

It should be noted that these advantages and disadvantages are listed or described anecdotally within the literature and are seldom accompanied by robust evidence to support claims made about MCQs. This presentation - or lack of – throughout the literature may be due to the strengths and weaknesses of MCQs being extremely dependent on how the questions are designed. However, similar assumptions could be made for all types of assessment (see subsequent sections (1.4.3–1.4.11)).

In terms of psychometric properties, MCQs are predominantly regarded as a reliable form of assessment (Norcini et al., 1985; Newstead and Dennis, 1994; Kniveton, 1996; McCoubrie, 2004; Considine, Botti and Thomas, 2005; Medical Council of Canada, 2010; Panczyk and Gotlib, 2015; Williams et al., 2015; Abdulghani et al., 2017; Javaeed, 2018; AlKhatib et al., 2020). This largely due to their efficiency since they can assess a large sample of topics in a short time compared to other formats of written assessment (McCoubrie, 2004). However, MCQ reliability (like for any assessment type) is ultimately dependent on the

how well the items have been designed. Reliability coefficients - such as KR-20 (Kuder and Richardson, 1937) and Cronbach's alpha (Cronbach, 1951) - can be affected by question format (AlKhatib et al., 2020), question difficulty, number of questions (Downing and Haladyna, 2004), number of available answers (AlKhatib et al., 2020), "function" of distractor answers (Ali, Carr and Ruit, 2016), and standard deviation of the results (Karras, 1997).

The KR-20 and Cronbach's alpha coefficients are often used to measure MCQ reliability since the internal consistency of this type of assessment is of interest. (Downing and Haladyna, 2004). Internal consistency is based on the average correlation between the items within the test (Nunnally and Bernstein, 1994) and determines the degree to which items assess similar areas of knowledge (Beanland et al., 1999; Polit and Hungler, 1999). Of all the reliability coefficients, Cronbach's alpha is the most frequently used (Downing and Haladyna, 2004; De Champlain, 2010). For an MCQ assessment to be reliable, its Cronbach's alpha coefficient must be $\geq$0.7 (Beanland et al., 1999; Polit and Hungler, 1999; Gravetter and Wallnau, 2000; Williams et al., 2015) - although some publications have advocated that it should be $\geq$0.8 or even $\geq$0.9 for high-stakes assessments (Downing and Haladyna, 2004; De Champlain, 2010) (see section 1.3.3.1).

In terms of MCQ validity, there is (currently) less assurance available within the literature. Many publications propose MCQs are valid, acknowledge that investigating their validity is important, and/or give recommendations on how MCQs can be designed to be valid. However, few publications present or refer to robust evidence which support their claims and/or recommendations. Examples of such publications include works by Bridge et al. (2003), Collins (2006), Medical Council of Canada (2010), Coughlin and Featherstone (2017) and Capan Melser et al. (2020).

The lack of good quality evidence for the validity of MCQs has been acknowledged for over 30-years (Violato, 1991; Masters et al., 2001; Surry, Torre and Durning, 2017) but there have been suggestions on how this could be improved upon (Haladyna, 1999; Considine, Botti and Thomas, 2005; Surry, Torre and Durning, 2017). Whether such recommendations will result in better quality studies on the validity of MCQs remains to be seen.

Recently, there have been some publications which depict a more measured and evidence-based approach to their findings on the validity of MCQs. For example, a study by Surry, Torre and Durning (2017) declares the purpose of the MCQs (determining clinical reasoning) from the outset and subsequently presents arguments and evidence on how the assessment can serve this purpose. This presentation is not reminiscent of many previous publications, but it is unclear if this study followed recommendations on how the quality of evidence on the validity of MCQs might be improved or if these improvements were made on initiative.

Finally, many studies (both historic and recent) seldom refer to the type(s) of validity (i.e., face, content, criterion, and construct (see section 1.3.3.1)) which are attributable to MCQs. Although some publications (Considine, Botti and Thomas, 2005) are an exception to this observation, the current lack of clarification once again makes it difficult to determine how valid MCQs are as an assessment method within medical educational subjects (including dentistry).

In summary, there appears to be more evidence to support the reliability of MCQs than their validity. It may be that the validity of MCQs is possibly being sacrificed for reliability (as suggested by Sam et al. (2016)) but calls for better quality studies on the validity of MCQs and recommendations on how this may be achieved may provide greater clarity in due course.

1.4.4 Extended Matching Questions (EMQs)

EMQs (also known as EMIs – Extended Matching Items) are another form of assessment which require candidates to select answers from a list. However, unlike MCQs, candidates are provided with the title or theme of the topic in question followed by a list of potentially acceptable options which are either numbered or lettered. This is followed by "lead in" statements which link the list of options to the questions asked. The questions are usually in the form of a clinical scenario. Candidates answer each question by selecting the best option from the list and - depending on how the questions and marking scheme have been written - may be required to select one of the options more than once to answer other questions under the same theme. Some answers on the list may not

need selected at all (Case and Swanson, 1993; Case and Swanson, 2001; Beullens et al., 2002; Williams et al., 2015; Jolly and Dalton, 2019).

Like MCQs, EMQs are aligned with both the "Knows" and "Knows How" levels of Miller's Pyramid of Professional Competence (Williams et al., 2015) (see section 1.4.1). Exactly which of these levels they will associate with depends on how the questions are written. However, it appears EMQs are more associated with testing clinical application of knowledge (e.g., diagnostic abilities and clinical judgement) (Case and Swanson, 1993; Veloski et al., 1999; Beullens et al., 2002; Beullens, Struyf and Van Damme, 2005; Beullens, Struyf and Van Damme, 2006; van Bruggen et al., 2012) rather than simple factual recall which, therefore, suggests they are typically more aligned with "Knows How".

EMQs also share some of the same advantages as MCQs. They are an objective form of assessment (Skakun, Maguire and Cook, 1994; van der Vleuten and Newble, 1994; Fowell and Bligh, 1998), can test a wide number of subjects in a short time (Beullens et al., 2002; Duthie et al., 2006) and can be computer marked (Kreiter, Ferguson and Gruppen, 1999; Schuwirth and van der Vleuten, 2003; Duthie et al., 2006; Baird, 2010). However, they are less prone to cueing and guesswork compared to MCQs (Case and Swanson, 1993; Skakun, Maguire and Cook, 1994; van der Vleuten and Newble, 1994; Fowell and Bligh, 1998; Veloski et al., 1999; Duthie et al., 2006; Baird, 2010; Williams et al., 2015) and offer a good degree of discrimination when testing higher levels of ability (Case and Swanson, 1993; Fenderson et al., 1997; Williams et al., 2015). They are also considered to be one of the fairest forms of assessment (McCoubrie, 2004).

Although it has been advocated that writing EMQ items is quicker and easier compared to other forms of written assessment (Case and Swanson, 1993; Fenderson et al., 1997; Schuwirth and van der Vleuten, 2003), others have contradicted this proposal, suggesting it can still be time consuming (Williams et al., 2015) and challenging for assessors to develop good quality EMQ items, especially for certain topics - e.g., surgical management (Beullens et al., 2002) and psychiatry (Samuels, 2006). Another potential disadvantage of EMQs is that certain topics or themes could be under-represented since it can be difficult to ask questions on certain themes and topics using the EMQ format (Schuwirth and van der Vleuten, 2003). There is also a risk of reducing the breadth of topics or

themes covered when items are linked together and, like MCQs, the scope for good quality feedback for students can be limited (Williams et al., 2015).

In terms of psychometric properties, EMQs have been suggested to be both reliable and valid. EMQs have been shown to return high reliability coefficients (Case and Swanson, 1993; Gruppen et al., 1994; Fenderson et al., 1997; Veloski et al., 1999; Beullens et al., 2002; Coderre et al., 2004) and there are well-designed studies providing evidence to support their content validity (Beullens et al., 2002; Coderre et al., 2004; Beullens, Struyf and Van Damme, 2005). There is also evidence to support their criterion validity in the assessment of the clinical application of knowledge (Gruppen et al., 1994; Fenderson et al., 1997; Wass, McGibbon and van der Vleuten, 2001; Beullens et al., 2002). Criterion validity was established in these studies through comparisons with other assessment types which were designed to test the clinical application of knowledge. However, it is worth considering this evidence could be potentially misleading if the validity of the methods EMQ are compared against was questionable to begin with. It was not clear within this group of studies whether the validity of these assessment methods had previously been thoroughly investigated or established.

Ultimately, the advantages, disadvantages, and psychometric properties of EMQ assessments will be influenced by how well they are designed. Like for MCQs, there are multiple publications available (e.g., Case and Swanson (1993) and Jolly and Dalton (2019)) to guide assessors on constructing good quality EMQs (which are reliable and valid) by demonstrating how title/topic headings, lead in statements, scenarios/stems and list of options/answers should be written and formatted.

Research on EMQ formats has resulted in a reduction of the recommended number of options from 15-20 to eight as there was evidence to support that streamlining the list of potential answers to eight did not significantly impact the psychometric properties of EMQ assessment (Swanson et al., 2005; Swanson, Holtzman and Allbee, 2008). Reducing the number of options also means there is potential for the time candidates spend on each question to be reduced (Swanson et al., 2005; Swanson, Holtzman and Allbee, 2008), which gives scope for a greater number and breadth of questions to be asked over the duration of the assessment. Increasing the number and breath of questions can result in

greater spread of scoring among candidates and, therefore, improve reliability and validity (Case and Swanson, 1993). EMQ assessments consisting of at least 100 questions have previously been shown to produce favourable psychometrics (Beullens et al., 2002).

1.4.5 Short Answer Questions (SAQs)/Multiple-short Answers (MSAs)

SAQs (also known as constructed responses, MSAs and Short Structured Answers (SSAs)) are another form of written assessment which are aligned with both the "Knows" and "Knows How" levels of Miller's triangle of professional competence (Williams et al., 2015) (see section 1.4.1). Unlike MCQs and EMQs, candidates are required to formulate a brief response to the questions asked instead of selecting answers from a list of options. Their responses may take the form of single words, a list, several sentences, short paragraphs, or short essays depending on how the questions are constructed (Rodriguez, 2003; Kramer et al., 2009; Williams et al., 2015; Jolly and Dalton, 2019; Royal College of Physicians and Surgeons of Canada, 2019). Although it should be noted that questions requiring short essay answers are regarded as a different type of assessment – known as modified essay questions (MEQs) – within medical education subjects (Wallerstedt, Erickson and Wallerstedt, 2012).

SAQs can be used to formulate a series of questions focused on a topic or theme which - in medical education subjects - are usually based on clinical scenarios (Kramer et al., 2009; Williams et al., 2015; Jolly and Dalton, 2019; Royal College of Physicians and Surgeons of Canada, 2019). Like MCQs and EMQs, they measure knowledge and the application of knowledge (Edwards and Arthur, 2007; Kramer et al., 2009; Jolly and Dalton, 2019) and, therefore, are aligned with the "Knows" and "Knows How" levels of Miller's triangle of clinical competence (Williams et al., 2015) (see section 1.4.1). SAQs are typically chosen over MCQs and EMQs when assessors wish to determine if candidates can generate spontaneous answers (Schuwirth and van der Vleuten, 2004b), i.e., if candidates can "recall" rather than "recognise" information (Royal College of Physicians and Surgeons of Canada, 2019). This format reduces guessing and cueing (Kramer et al., 2009; Williams et al., 2015), which is the main advantage of SAQs over MCQs and EMQs.

Other advantages of SAQs which have been suggested are that they:

- promote long-term retention of information (compared to MCQs) (McDaniel, Roediger and McDermott, 2007; Larsen, Butler and Roediger, 2008; Wood, 2009; McConnell, St-Onge and Young, 2015);

- facilitate provision of feedback (Williams et al., 2015; Sam et al., 2019);

- are easier to write (compared to essays) (Damjanov et al., 1995; Fenderson et al., 1997; Williams et al., 2015);

- are easier to mark (compared to essays) (Shumway and Harden, 2003; Williams et al., 2015; Jolly and Dalton, 2019);

- can be scored relatively objectively (Edwards and Arthur, 2007).

However, the ease of marking and objectivity of scoring is dependent on the provision of clear marking schedules outlining the correct responses to assessors. Objective scoring could be enhanced by using computer marking, which can be easily adopted for SAQs that require one-word answers but not those which require more extensive candidate responses (i.e., short sentences, paragraphs, and essays). Technologies which permit computer marking of longer responses have been developed and trialled (Leacock and Chodorow, 2003; Jordan and Mitchell, 2009; Sam et al., 2018; Sam et al., 2019) but have not yet transitioned into widespread use.

At present, UK dental schools do not appear to have adopted computer marking for SAQs (Roudsari, 2017). This suggests that “hands on” marking methods are still used, which can be more intensive in terms of time and administration (compared to MCQs) (Rademakers, Ten Cate and Bar, 2005; Edwards and Arthur, 2007; Williams et al., 2015). Another disadvantage is that SAQ scoring may be prone to subjectivity and influenced by assessors penalising candidates for poor handwriting, spelling, and grammar (Kramer et al., 2009), despite the use of marking schedules.

Previously, there was little robust evidence on the reliability and validity of SAQs. Some publications have suggested they are more reliable than essays (Grant, 1957; Schuwirth and van der Vleuten, 2004b) - largely because they avoid issues surrounding the scoring of longer student responses (e.g., more subjective marking) and can test a larger sample of course content within a given timeframe (Jolly and Dalton, 2019). Others have proposed that they are less reliable than EMQs (Baird, 2010). However, the evidence on which these claims (made in relation to both essays and EMQs) appears to be unclear within these publications. Despite this, SAQs have previously been shown to produce high reliability coefficients (Rademakers, Ten Cate and Bar, 2005) and have exhibited a degree of criterion validity (Edwards and Arthur, 2007) in some appropriately designed studies.

In recent years, new evidence has begun to emerge from the literature for both the reliability and validity of SAQs. An initial pilot study by Sam et al. (2016), which consisted of 266 student participants, concluded that, since students were less likely answer correctly in a SAQ format compared to MCQs, there was a possibility that SAQs were a more valid form of assessment of student knowledge. This pilot was followed by two larger studies which adopted statistical methods to investigate both the reliability and validity of SAQs. The first of these studies compared the reliability coefficients (Cronbach's alpha) of both SAQs and MCQs and found that the former returned a higher reliability coefficient over 60 assessment items than the latter. These findings were generated from 299 medical student participants (Sam et al., 2018).

The second was a large, multi-centre cross-sectional study involving 1417 medical students across 20 medical schools. In this study, the reliability of SAQs was once again compared with MCQs (using Cronbach's alpha) and findings on their validity were based on calculations which determined the rate of the cueing effect in both SAQs and MCQs (i.e., the less cueing there was, the more valid the assessment). The study concluded that, compared to MCQs, SAQs appeared to be a more reliable and valid method for assessing student knowledge since they produced higher Cronbach's alpha coefficients and lesser rates of cueing. However, the study acknowledged that further investigations were required – particularly for validity (Sam et al., 2019).

A more recent study also concluded that SAQs appeared more reliable and valid than MCQs (Puthiaparampil and Rahman, 2020). Although this study had less medical student participants than those conducted by Sam et al. (2018) (2019), it used different methodological approaches to investigate the psychometrics of SAQs. T-tests, Pearson correlation coefficients (r) and Chi-square tests were used to investigate reliability and the opinions of key stakeholders (students and faculty) were used to investigate validity. Overall, the study appeared to provide better evidence for reliability than for validity since the former was established through robust statistical measurements and the latter was simply based on opinions generated from a short questionnaire. It was also unclear which type(s) of validity (face/content/both) was (were) being investigated.

Both the recent studies by Sam et al. (2018) (2019) and Puthiaparampil and Rahman (2020) signify a shift towards publication of evidence for the reliability and validity of SAQs supported through statistical measurement and/or obtained via appropriate study methods.

Like for MCQs and EMQs, it is well understood that the psychometric properties of SAQs will be influenced by how well they are designed. Reliability has been said to improve when assessors are provided with clearly structured marking schedules (since they can help reduce subjective scoring) and by having at least two assessors score the candidate's answers independently of one another (Williams et al., 2015) - but there remains a lack of confirmatory research.

Multiple guides on the construction and scoring of SAQs items are available within the literature (examples include publications by Jolly and Dalton (2019) and the Royal College of Physicians and Surgeons of Canada (2019)). Like for MCQs and EMQs, these guides aim to help assessors maximise the advantages and psychometric properties of the SAQ format.

NOTE: Although "SAQ" appears to be the more commonly used term within the literature, from this point onwards this thesis will use "MSA", as this is the term used with the University of Glasgow to describe this form of assessment (University of Glasgow, Accessed 2021).

1.4.6 Spotter tests

Depending on the institution, “Spotter” tests may be known by different terms to describe their format - examples of which include “Spot”, “Timed Stations” (Williams et al., 2015), “Bell Ringer”, “Steeplechase” (Inuwa et al., 2011; Williams et al., 2015) and objectively structured practical examination (Tirpude et al., 2019). They have generally been described as a series of stations containing a specimen, labelled dissection, or radiograph. Candidates move between the stations and answer the question(s) within them. For some questions, only one-word answers are required, whereas others necessitate more comprehensive responses (Williams et al., 2015).

Spotter tests are aligned with both the “Knows” and “Knows How” levels of Miller’s triangle of clinical competence (Williams et al., 2015) (see section 1.4.1). They have traditionally been used within anatomy (Inuwa et al., 2011; Smith and McManus, 2015) to assess if students can identify anatomical structures and - in some cases - their function. Assessment of anatomical knowledge is a component of medical and dental curricula and, therefore, Spotter tests have also been utilised for this purpose within medical (Chirculescu, Chirculescu and Morris, 2007; Tirpude et al., 2019) and dental education (Williams et al., 2015). Pathology and radiology knowledge are also known to have been assessed in dental education via Spotter tests (Williams et al., 2015).

Spotter tests have been integrated with other assessment methods - particularly the OSCE (Yaqinuddin et al., 2013; Smith and McManus, 2015). This is because the formats of both Spotter tests and OSCEs are very similar (see section 1.4.6) and it is possible that, as a result, the term “OSCE” has superseded the term “Spotter” test (and the various other terms that have been used to describe their format). It could also explain the lack of literature available on Spotter tests.

The lack of literature makes it difficult to compile a list of advantages and disadvantages and describe the evidence available on the psychometric properties of Spotter tests in detail. However, it could be that, with respect to

this assessment format, these aspects are now more commonly reported in publications which concern OSCEs.

Based on what little literature is currently available, Spotter tests have returned good reliability scores – although this finding stems from a single, small-scale study (Tirpude et al., 2019). The same study also proposed that Spotter tests had "fair validity", but it was not clear how this conclusion was reached, or which type of validity had been determined. Another publication has also suggested that Spotter tests are valid; however, this was based on "the author's experience" (Zafar et al., 2013) rather than robust evidence.

1.4.7 Traditional essays/assignments

Traditional essay assessment formats require candidates to write long, comprehensive answers to the question(s) asked. They are "open" forms of assessment as they provide students with little or no structural guidance on how the question(s) should be answered. Candidates may be provided with the question(s) on a theme or topic in advance and are required to compose and submit their answers by a deadline or within a set timeframe under examination conditions (a "seen" essay). Alternatively, candidates may not be presented with the question(s) until they are under examination conditions (an "unseen" essay) (Jolly and Dalton, 2019).

Essay formats (both seen and unseen) are designed to assess depth and/or application of knowledge and, therefore, are aligned with the "Knows" and "Knows How" levels of Miller's triangle of clinical competence (Albino et al., 2008; Williams et al., 2015) (see section 1.4.1). Good essay questions require candidates to process information, think critically and/or apply their knowledge (Day et al., 1990; Shumway and Harden, 2003; Schuwirth and van der Vleuten, 2004b). Essay questions that test these skills are more associated with the "Knows How" level. However, Hift (2014) and Jolly and Dalton (2019) have argued that essays often just assess factual recall, which would align them more with the "Knows" level of Miller's triangle.

Some of the proposed advantages of essays are that they:

- can be easily set as an assessment method (Palmer and Rideout, 1995).
- potentially drive deep learning (compared to MCQs) (Scouller, 1998).
- determine how capable candidates are at constructing clear, detailed responses which are grammatically correct and organised in manner that addresses the question posed (Schuwirth and van der Vleuten, 2003; Jolly and Dalton, 2019).
- provide insight into how well candidates can apply knowledge to new situations (Schuwirth and van der Vleuten, 2003).
- facilitate written feedback [to aid student learning] – although the process for this can be time consuming (Williams et al., 2015).

Despite these advantages, the traditional essay is recognised as an assessment method which is very prone to cheating and plagiarism (Bilic-Zulle et al., 2005; Williams et al., 2015; Lynch et al., 2017; Javaeed et al., 2019; Jolly and Dalton, 2019). Submissions need to be carefully checked to ensure the work presented is the candidate's own; however, there is software available to assist faculty in detecting plagiarism during marking (e.g., "Turnitin" (Heckler, Rice and Hobson, 2013)). The marking process itself can be difficult (Palmer and Rideout, 1995), resource intensive (Wainer and Thissen, 1993; Williams et al., 2015), susceptible to assessor bias (Williams et al., 2015; Jolly and Dalton, 2019) and subjective (Hift, 2014). Variability in scoring between assessors has been well recognised in relation to essays (Bloxham et al., 2016) and their scoring could be negatively affected by poor grammar, sentence and paragraph structure (Linn, Klein and Hart, 1972), and handwriting (Markham, 1976). Another disadvantage of essays is that they only sample a narrow area of candidate knowledge in depth (Hift, 2014; Williams et al., 2015; Jolly and Dalton, 2019) during a lengthy time period (Williams et al., 2015).

Due to these disadvantages, traditional essays are associated with low reliability (Palmer and Rideout, 1995; Schuwirth and van der Vleuten, 2004b; Williams et al., 2015). Statistical evidence to support this consensus is lacking but this is

probably due to the very small number of questions that are typically asked within essays assessments which makes it difficult – if not impossible – to apply reliability coefficient calculations to traditional essays.

It has been advocated that reliability could be increased by applying a more structured format (Verma, Chatwal and Singh, 1997), decreasing the length of the questions (Nendaz and Tekian, 1999) and increasing the number of questions (Feletti and Smith, 1986; Nendaz and Tekian, 1999). By adopting these changes, there is potential for greater objective marking and assessment of a broader spectrum of course content within a similar timeframe - increasing reliability and efficiency. However, the design and format will be changed to the point where it no longer resembles the traditional essay and thus no longer has the advantages associated with asking longer open-ended questions. Instead, adopting these suggested changes would signify that a different form of assessment is being used (e.g., SAQs (see section 1.4.5)).

Other possible means for improving the reliability of essays - without drastically changing their design and format – include adopting double marking (Williams et al., 2015; Jolly and Dalton, 2019) and providing assessors with marking schedules/model answers (Jolly and Dalton, 2019). The former of these approaches could help reduce assessor bias by presenting an opportunity for assessors to reach a consensus on scoring, whereas the latter aims to reduce variability between assessors and assessor bias by promoting more objective scoring. However, whilst the latter approach could improve reliability, it may significantly reduce validity since marking against standardised marking schedules trivialises the essay format (Schuwirth and van der Vleuten, 2003).

Lastly, there appears to be very little literature focused on investigating the validity of traditional essays within medical education subjects. However, Hift's (2014) extensive review of the available literature concluded there was little evidence to suggest that traditional essay formats had good validity, especially when compared to MCQs (see section 1.3.2). The collective lack of favourable evidence on the psychometrics of traditional essays - coupled with their other disadvantages - has led to some authors suggesting that the format should not be used for high stake assessments (Hift, 2014; Williams et al., 2015). However, it is

currently unclear from the available literature if traditional essays are being used formatively and/or summatively within dental education.

1.4.8 Oral examinations

Oral examinations (Vivas/Viva Voces) have been widely used within medical education subjects (Wass et al., 2003; Davis and Karunathilake, 2005) - including dentistry. In general, they involve candidates entering discussions with one or more assessors during which they are asked a series of questions on single or multiple topics. There are various formats of the assessment ranging from completely unstructured to highly structured (Schuwirth and van der Vleuten, 2019). Unstructured formats give assessors greater freedom over the questions they can ask and are a less standardised form of the assessment as a different line of questioning can be pursued for each candidate. In contrast, structured formats (Morrell, 1984; Davis and Karunathilake, 2005) increase standardisation since assessors have a pre-defined list of questions to ask each candidate.

Some formats require candidates to discuss a clinical case (or cases) they have previously seen and/or treated. Other formats present candidates with an unseen clinical case. The former approach can lend itself to a more unstructured format whereas the latter tends to facilitate a more structured format. The degree of structuring will ultimately influence the advantages, disadvantages, and psychometric properties of the assessment (see below).

Oral examinations are used to assess knowledge and application of knowledge (Cox, 1982; Gibbs, Habeshaw and Habeshaw, 1988; Anastakis, Cohen and Reznick, 1991; Jolly and Grant, 1997) - therefore they can be aligned with the “Know” and “Knows How” level of Miller’s triangle of clinical competence (Williams et al., 2015). They are typically used to evaluate clinical reasoning (Ryding and Murphy, 1999; Petrusa, 2002) and decision making (Wass et al., 2003) but can also assess other traits and attributes, such as oral communication skills (Ryding and Murphy, 1999), professionalism (Ryding and Murphy, 1999; Wass et al., 2003), hypothesis generation and the transfer of principles through various contexts. The potential to test these traits and attributes – especially regarding the diagnosis, treatment, and management of authentic clinical situations - is the main advantage of oral examinations over written forms of

assessment. Another advantage is that they can facilitate face-to-face feedback for candidates (Colton and Peterson, 1967; Kearney et al., 2002; Williams et al., 2015).

A disadvantage of oral examinations is that they are time consuming and resource intensive (Wass et al., 2003). They are also prone to the "halo effect", whereby assessor judgments may be affected by their impression of the candidate or through comparing the performance of the candidate they are assessing with the performance of previous candidates (Williams et al., 2015). Another disadvantage is the potential for examiner bias (Colton and Peterson, 1967; Foster et al., 1969). Not only could this be problematic in terms of marking, but candidates may also try to take advantage of assessor bias by identifying topics which individual assessors prefer to ask questions on and then preparing strategically for the examination (Schuwirth and van der Vleuten, 2019). Furthermore, the reliability (Colton and Peterson, 1967; Foster et al., 1969; Muzzin, 1995; Turnbull, Danoff and Norman, 1996; Williams et al., 2015; Schuwirth and van der Vleuten, 2019) and validity (Colton and Peterson, 1967; Foster et al., 1969; Davis and Karunathilake, 2005) of oral examinations have been questioned. However, it should be remembered that these properties can be significantly influenced by the design and format of the assessment.

Authors have proposed that the reliability of oral examinations can be improved by:

- increasing the number of patient cases discussed within the assessment (Daelmans et al., 2001; Williams et al., 2015).

- increasing the number of questions asked per case. One study has advised that at least five set questions should be asked per case (Amiel et al., 1997).

- covering a range of topics (Amiel et al., 1997; Wass et al., 2003; Schuwirth and van der Vleuten, 2019) (i.e., assessing as much course content as possible).

- asking the same questions to each student (Wass et al., 2003).

- having multiple assessors (Wass et al., 2003; Williams et al., 2015). Wass et al. (2003) proposed two assessors (per oral examination).

- using different examiners to assess different cases (Norman, 2000), i.e., avoid having the same examiner for different cases.

- adopting a rotational system, whereby candidates move between assessors, each of whom addresses a different, pre-defined case or topic (Schuwirth and van der Vleuten, 2019).

- using structured marking schedules (Yang and Laube, 1983; Anastakis, Cohen and Reznick, 1991; Wass et al., 2003).

- adopting a "global judgment" scale whereby assessors make a subjective judgment on how they think the candidate performed (Daelmans et al., 2001).

- training and calibrator assessors (Des Marchais and Jean, 1993; Wakeford, Southgate and Wass, 1995; Ryding and Murphy, 1999; Wass et al., 2003).

- training assessors to ask questions which cover a breadth of topics instead of just asking questions on their own areas of interest (Schuwirth and van der Vleuten, 2019).

- increasing the testing time (Daelmans et al., 2001; Wass et al., 2003; Williams et al., 2015). Wass et al. (2003) proposed a total testing time of 80-minutes (four 20-minute assessments).

Most of these claims on improving reliability are supported through studies demonstrating an increase in reliability coefficients following implementation of the suggested changes (Yang and Laube, 1983; Anastakis, Cohen and Reznick, 1991; Amiel et al., 1997; Daelmans et al., 2001; Wass et al., 2003). However, within some publications – especially those which provide summaries of various assessment methods used in medical and dental education (Williams et al., 2015; Schuwirth and van der Vleuten, 2019) - it is not clear where the basis of their claims stem from. In addition, some studies (such as Wass et al. (2003)) present

an improvement in reliability through the application of several of the design features listed above, making it difficult to determine which feature(s) had the greatest impact on the reliability coefficient – either individually or in combination with one or more other design features.

Compared to reliability, there are very few studies which discuss the validity of oral assessments in detail. One study which has investigated their validity is Anastakis, Cohen and Reznick (1991), who compared a structured oral assessment with MCQ and OSCE. The study showed there was some evidence to suggest structured oral assessments had criterion validity for the assessment of clinical knowledge and problem-solving since there was significant correlation between the outcomes of structured oral assessment and MCQ and OSCE scores. However, these results were based on assessment outcomes produced by only twenty-three candidates. A subsequent larger study, which compared the outcomes of 441 structured oral examinations with the results of written "in-training" examinations for anaesthetists, also concluded there was evidence for their criterion validity in describing clinical competence - including the assessment of clinical knowledge and problem-solving (Schubert et al., 1999). However, unlike Anastakis, Cohen and Reznick (1991), it was not clear which format(s) of written assessment were used for the "in-training" examinations.

Despite the lack of evidence for their validity, some authors have proposed there is still a role for oral assessments within medical education subjects providing they are used to test traits and abilities that cannot be measured through other formats (e.g., hypothesis generation and explanation) (Schuwirth and van der Vleuten, 2019). Using them to assess simple factual recall could be counter intuitive since this can be accomplished using methods which are less time consuming and resource intensive.

1.4.9 Objective Structured Clinical Examinations (OSCEs)

Introduced in 1975, OSCEs were designed to provide a standardised, objective and reliable method for assessing clinical skills (Harden et al., 1975) - such as history taking, examination of a patient or performance of a practical procedure (Boursicot, Roberts and Burdick, 2019). They have gained widespread popularity within medical education subjects (Cohen et al., 1990; van der Vleuten, 1996;

Davis, 2003; Newble, 2004; Harden, 2016; Schuwirth and van der Vleuten, 2019) and are now one of the most heavily researched assessment methods with over 1600 publications discussing their use (Harden, 2016).

The typical format of an OSCE involves candidates entering multiple stations where they are asked to perform a task (or tasks) within a predetermined timeframe. Candidate performances are marked against a set list of "objective" criteria and, once the allotted time for task has lapsed, an alarm (e.g., bell or buzzer) sounds to notify candidates to move onto the next station (Williams et al., 2015; Boursicot, Roberts and Burdick, 2019). Although this describes the basic format of OSCEs, they can be implemented in different ways, which can affect their psychometric properties (Harden, 2016) (see below). Since they assess practical clinical skills in a staged/simulated/mock environment, they are aligned with the "Shows How" level of Miller's triangle of clinical competence (Albino et al., 2008; Williams et al., 2015) (see section 1.4.1).

OSCEs have been described as a fair and - as indicated by their name - objective form of assessment (Watson et al., 2002; Williams et al., 2015; Boursicot, Roberts and Burdick, 2019) since all candidates undertake the same clinical scenarios and are marked using the same assessment criteria (Boursicot, Roberts and Burdick, 2019). Since their format requires candidates to demonstrate their proficiency in performing clinical skills, OSCEs may encourage students to adopt learning strategies which ensure they gain the required competencies (i.e., they practise and develop the necessary clinical skills) (Schoonheim-Klein et al., 2009). In contrast, if students are presented with assessments designed to test knowledge (e.g., MCQs and SAQs), they adopt learning strategies which focus of knowledge acquisition and recall (Boursicot, Roberts and Burdick, 2019). If used formatively, OSCEs present further opportunities for student learning through provision of detailed feedback on performance (Hattie and Timperley, 2007; Williams et al., 2015).

Despite these potential advantages, OSCEs have been criticised of having underlying issues with authenticity and case specificity (Swanson, 1995; van der Vleuten, 1996; Norman et al., 2006). Lee and Wimmers (2011) even concluded - from a comprehensive study involving 686 student participants - that OSCEs may not be able to assess proficiency in a single domain of clinical competence.

However, the findings of this study were based on assessment results obtained from a single institution and no subsequent studies appear to have echoed these remarks.

According to several authors, the most prominent drawback of OSCEs is that they are expensive and time consuming to establish, set up and run due to the amount of resources and logistical planning required (Carpenter, 1995; Albanese and Dast, 2014; Brown et al., 2015; Williams et al., 2015; Boursicot, Roberts and Burdick, 2019). However, like with any assessment method, the additional costs and effort may be worthwhile if the assessment proves to be reliable and valid.

As mentioned above, the psychometric properties of an OSCE are determined by its design and implementation. Reliability coefficients for OSCEs have been shown to be influenced by:

- Number of stations (Newble and Swanson, 1988; Schoonheim-Klein et al., 2008; Brannick, Erol-Korkmaz and Prewett, 2011).

- Increased testing time/test length (Newble and Swanson, 1988; Roberts et al., 2006).

- Wider sampling of skills (Watson et al., 2002; Roberts et al., 2006; Schoonheim-Klein et al., 2008).

- Number of examiners (Brannick, Erol-Korkmaz and Prewett, 2011).

Some publications have suggested that structured marking schedules (checklists) can increase reliability (Boursicot, Roberts and Burdick, 2019); however, the evidence which supports this proposal is not apparent within the literature. Other have cast doubts on their influence and have instead suggested that the use of “global scores” (i.e., subjective judgments made on candidate performance by assessors) results in similar (Cunnington, Neville and Norman, 1996) or greater (Regehr et al., 1998; Moineau et al., 2011; Ilgen et al., 2015) reliability compared to structured marking schedules. However, Regehr et al. (1998) also showed that the combined use of structured marking schedules and global rating scores elevated reliability furthest. In addition, Homer and Pell

(2009) demonstrated that the inclusion of simulated patient ratings (i.e., subjective judgments from actors who pretend to be patients in a mock clinical consultation) in marking schedules can potentially enhance the reliability of an OSCE. However, it should be remembering that introducing more subjective elements – such as global scores and simulated patient ratings – may compromise the objectivity of an OSCE.

More subjective judgments could result a greater range of scores being awarded by assessors. Potential assessor bias which may be inherent within a wider range of scores could potentially be reduced with assessor training and calibration. Indeed, several publications concerned with assessment in medical education subjects (such as Boursicot et al. (2011) and Monti et al. (2020)) advocate that assessor training is necessary. However, previous research has shown that assessor training appears to have little impact on improving inter-assessor reliability (Newble, Hoare and Sheldrake, 1980; Boursicot, Roberts and Pell, 2007; Cook et al., 2009) and only reduces the range of assessor scoring (Holmboe, Hawkins and Huot, 2004).

Regardless, there is plenty of support within the literature to suggest that OSCEs can be reliable assessment method within medical education subjects (Brown, Manogue and Martin, 1999; Nickbakht, Amiri and Latifi, 2013; Setyonugroho, Kennedy and Kropmans, 2015) and good reliability coefficients ($\geq$0.7) have been demonstrated in numerous studies (Eva et al., 2004; Park et al., 2004; Roberts et al., 2006; Taghva et al., 2010; Brannick, Erol-Korkmaz and Prewett, 2011; Eberhard et al., 2011; Pascual Ramos et al., 2015; Rahayu et al., 2016; Trejo-Mejia et al., 2016).

In terms of validity, the length of the stations and (like for reliability) the use of structured marking schedules (checklists) and/or global scores are key influential factors. More course content can be covered with longer stations and consideration also needs to be given to the length of time required to assess the skills being tested within each station. Therefore, the length of each station should be determined by their content (Harden and Gleeson, 1979; Cizek, 2001; Hodges, 2003; Newble, 2004; Varkey et al., 2008). For example, a station assessing prescription writing may only need to be five minutes long, whereas a

station assessing the preparation of tooth for a restoration may be 20-minutes long.

Like for reliability, some studies have proposed that using global scores instead of structured marking schedules increases the validity of an OSCE (Cunnington, Neville and Norman, 1996; Regehr et al., 1998; Hodges, 2003; Daniels and Harley, 2017). Structured marking schedules can lead to candidates performing "monkey tricks" within OSCEs rather than demonstrating they acquired the skills being tested (Cizek, 2001), whereas global scores may capture elements of performance that may be overlooked by marking schedules (Govaerts, van der Vleuten and Schuwirth, 2002).

However, Regehr et al. (1998) have shown that using a combination of both improves OSCE validity further and this approach has also been recommended by subsequent publications (Park et al., 2004; Rushforth, 2007; Monti et al., 2020).

Several studies have presented evidence for the face validity of OSCEs. The findings of these studies were based on evaluations from faculty (Macluskey et al., 2011; Barry, Bradshaw and Noonan, 2013; Nickbakht, Amiri and Latifi, 2013) or a combination of both faculty and students who were assessed (Brown, Manogue and Martin, 1999; Walters, Osborn and Raven, 2005). All these studies concluded that OSCEs have face validity.

The studies by Brown, Manogue and Martin (1999), Walters, Osborn and Raven (2005), Macluskey et al. (2011), and Barry, Bradshaw and Noonan (2013) also proposed that OSCEs had content validity. Again, the findings of these studies were based on the opinions of faculty and students. Studies by Varkey et al. (2008) and Hodges et al. (1998) also concluded that OSCEs have content validity. The former's findings were based on the student evaluations, whereas the latter's findings were based on evaluations from residents (i.e., clinicians) in psychiatry. Taghva et al. (2010) also proposed OSCEs had both face and content validity, but it is unclear what the evidence for these claims was within this study.

Although there appears to be a consensus that OSCEs have face and content validity, there was notable heterogenicity between the studies listed above. For

example, the number of OSCE stations which were evaluated ranged from as little as one (Macluskey et al., 2011) to as many as 18 (Walters, Osborn and Raven, 2005) and the number of student participants (i.e., the candidates who sat the exam) varied from as few as 14 (Nickbakht, Amiri and Latifi, 2013) to as many as 498 (Macluskey et al., 2011). The number of faculty who evaluated an OSCE also varied significantly (from "at least" four (Walters, Osborn and Raven, 2005) to twenty-one (Brown, Manogue and Martin, 1999)) or, in the case of some studies, wasn't described (Taghva et al., 2010; Barry, Bradshaw and Noonan, 2013).

In addition to face and content validity, Brown, Manogue and Martin (1999), Hodges et al. (1998) and Taghva et al. (2010) also investigated the criterion validity of OSCEs by comparing their results against other assessments. Brown, Manogue and Martin (1999) found OSCE results correlated poorly with A-levels and the "Final" examinations in medicine - which consisted of written papers, a 'long case' examination, a presentation case and vivas. Hodges et al. (1998) compared OSCE scores with lists of candidate rankings submitted by faculty and found a moderate correlation – but only if global scoring was used for the OSCE. Taghva et al. (2010) demonstrated a moderate correlation between OSCEs and oral examinations, but a weak correlation with MCQs. The latter of these findings echoed previous studies which had also investigated the criterion validity of OSCEs (Ross et al., 1988; Cunnington, Neville and Norman, 1996; Dennehy, Susarla and Karimbux, 2008), however, it should be remembered that the intended purposes of OSCEs and MCQs may differ. MCQs focus on testing knowledge and its application in problem solving (see section 1.4.3) and although OSCEs can also be used to test these traits, they are primarily used to assess clinical skills.

Studies by Park et al. (2004) and Eberhard et al. (2011) have also compared OSCE scores with other forms of assessment. Park et al. (2004) compared results of a nine station OSCE from two-hundred and eighty-six students with the outcomes of the National Board of Medical Examiners Psychiatry Subject Examination and five clinical skills examinations. Eberhard et al. (2011) correlated the scores of an eleven station OSCE from sixty-two students with the results of a "clinical skills examination". Both studies claimed that their findings

demonstrated that OSCEs had adequate construct validity. However, it could be argued that larger studies and a variety of other research approaches are required before such claims can be made (see section 1.3.3.2).

Overall, although evidence for the validity of OSCEs doesn't appear as strong as some studies have suggested, it is an assessment format which is widely accepted and supported within medical education subjects (Brown, Manogue and Martin, 1999; Hodges, 2003; Park et al., 2004; Varkey et al., 2008; Taghva et al., 2010; Barry, Bradshaw and Noonan, 2013; Nickbakht, Amiri and Latifi, 2013). Providing they are well designed and implemented, high reliability and validity can be achieved within the OSCE format (Rushforth, 2007).

1.4.10 Direct observation of procedural skills (DOPS)/Competency tests

Developed by the Royal College of Physicians (Wilkinson et al., 2008; Cohen, Farrant and Taibjee, 2009), DOPS/competency tests are used to assess practical clinical skills in workplace settings (Cohen, Farrant and Taibjee, 2009; Barton et al., 2012; Naeem, 2013; Williams et al., 2015). Students are closely observed by supervising clinical faculty whilst performing a clinical procedure on a real patient. Supervising faculty then score students against a list of predetermined criteria and determine whether they performed the procedure competently (or not). Once the assessment has been completed, students are given feedback on their performance (Wragg et al., 2003; Wilkinson et al., 2008; Williams et al., 2015; Erfani Khanghahi and Ebadi Fard Azar, 2018).

Depending on institutional guidelines, students may be required to demonstrate they are able to perform the procedure competently a set number of times. Alternatively, some institutions invite students to judge when they think they are competent and then arrange to be assessed. Students who satisfy the criteria are then "signed off" as competent in performing the procedure, whereas unsuccessful students are invited to reattempt the assessment later. There is currently no evidence to suggest which institutes use either of these approaches, or whether DOPS are used for formative or summative assessment (or a combination of both).

Since DOPS assess clinical skills in real clinical environments and scenarios, they are aligned with the "Does" level of Miller's triangle of clinical competence (Albino et al., 2008; Williams et al., 2015) (see section 1.4.1). Many publications advocate that the major strength of DOPS is their allocation of time for feedback on performance (Wiles et al., 2007; Wilkinson et al., 2008; Cohen, Farrant and Taibjee, 2009; McLeod, Mires and Ker, 2012; Cobb et al., 2013; Dabhadkar et al., 2014; Erfani Khanghahi and Ebadi Fard Azar, 2018; Norcini and Zaidi, 2019). Feedback provides students who don't pass the assessment with a learning opportunity (Wilkinson et al., 2008; Cohen, Farrant and Taibjee, 2009; McLeod, Mires and Ker, 2012) which highlights areas of performance requiring improvement. This process can facilitate student learning and the development of clinical skills (Erfani Khanghahi and Ebadi Fard Azar, 2018; Tenzin et al., 2019). For students who pass the assessment, feedback can provide reassurance that they are performing to the required standards (Cohen, Farrant and Taibjee, 2009).

Other advantages of DOPS which have been proposed within the literature include:

- The potential to have students assessed by multiple assessors (Norcini and Zaidi, 2019), reducing the risk of assessor bias.

- Close supervision and observation from assessors (Cohen, Farrant and Taibjee, 2009).

- Promotion of student autonomy during the assessment (Dhole, 2017; Erfani Khanghahi and Ebadi Fard Azar, 2018; Tenzin et al., 2019).

- Promotion of deep student reflection (Cobb et al., 2013).

- The use of real patients (Norcini and Zaidi, 2019), increasing authenticity (and validity - see below).

- Acceptability among both students and faculty (Erfani Khanghahi and Ebadi Fard Azar, 2018).

Disadvantages of DOPS include:

- Time consuming and difficult to arrange and/or organise (Bradley and Huseman, 2003; Wilkinson et al., 2008; Cohen, Farrant and Taibjee, 2009; Erfani Khanghahi and Ebadi Fard Azar, 2018).

- Viewed as stressful and artificial (Cohen, Farrant and Taibjee, 2009; Akbari and Mahavelati Shamsabadi, 2013; Cobb et al., 2013; Erfani Khanghahi and Ebadi Fard Azar, 2018) and unsettling by some candidates, resulting in impaired performance (Hamilton et al., 2007).

- Possible bias (Akbari and Mahavelati Shamsabadi, 2013; Amini et al., 2015; Erfani Khanghahi and Ebadi Fard Azar, 2018) or variability among assessors (Erfani Khanghahi and Ebadi Fard Azar, 2018). However, the risk of these issues could be reduced through use of multiple assessors (Norcini and Zaidi, 2019).

- Disagreement between assessors and the student and the assessor(s) on the correct procedural technique (Cohen, Farrant and Taibjee, 2009).

- Difficulty in identifying suitable patient cases for assessment (Cohen, Farrant and Taibjee, 2009).

- Potential to become a “tick-box” exercise which doesn’t provide proof of attainment of competence (Bindal et al., 2013).

The advantages and disadvantages listed above stem from feedback and evaluations submitted by undergraduate students, postgraduate specialist medical subject trainees (i.e., Specialist Registrars), faculty and/or postgraduate trainers.

In terms of psychometric properties, multiple studies have presented high (>0.70) reliability coefficients for DOPS (Hamdy et al., 2003; Marriott et al., 2011; Asadi et al., 2012; Barton et al., 2012; Sahebalzamani and Jahantigh, 2012; Delfino et al., 2013; Tsui et al., 2013; Kuhpayehzade et al., 2014). Factors

which have been associated with improved reliability include the use of structured marking schemes/checklists (Tennant and Scriva, 2000; Scott et al., 2001) and assessor training (Wilkinson et al., 2008).

Studies by (Wilkinson et al., 2008; Marriott et al., 2011; Barton et al., 2012) have proposed that DOPS hold high face validity. These findings were based on the opinions of students (who had been assessed with DOPS) and assessors. Various studies have investigated the content validity of DOPS by calculating their content validity index (CVI) and/or content validity ratio (CVR) (Erfani Khanghahi and Ebadi Fard Azar, 2018). Assessments with a CVI >0.78 are said to be reliable (Polit, Beck and Owen, 2007) as are those with a CVR >0.78 (Zamanzadeh et al., 2015; Kovacic, 2018). Kuhpayehzade et al. (2014) suggested DOPS had low reliability since both the CVI and CVR were <0.78. Hengameh et al. (2015) produced a mixed set of results where DOPS were found to have a low CVR (0.62) but a higher CVI (0.79), the latter of which coincides with findings by Delfino et al. (2013) (0.90) and Amini et al. (2015) (0.95). Other studies have investigated the criterion validity of DOPS by comparing them against other forms of assessment. Hamdy et al. (2003) found DOPS had good Pearson's correlation coefficients (r) (>0.70) with patient management problems, SAQs and an OSCE and a moderate correlation with MCQs (0.67). Barton et al. (2012) found a weak Pearson's correlation (0.28) between DOPS and MCQs.

Pearson's correlation coefficient was also used by Marriott et al. (2011) to investigate the construct validity of DOPS through comparisons with measures of surgical training and experience. Although this study concluded construct validity was demonstrated, the correlation between DOPS and measures of surgical training and experience was moderate and further evidence using a variety of other methodological approaches is required before construct validity can be established.

Like for the reliability, it has been advocated that the validity of DOPS can be improved by incorporating structured marking schemes/checklists into their design (Tennant and Scriva, 2000; Scott et al., 2001). Additionally, the use of global ratings has also been said to improve their validity (Winckel et al., 1994; Larson et al., 2005).

Overall, there is some good evidence to support DOPS as a reliable and valid assessment method (Erfani Khanghahi and Ebadi Fard Azar, 2018); however, there remains a need to conduct further research to strengthen the findings presented by the publications referenced above.

1.4.11 Longitudinal clinical assessment

Like DOPS/competency tests, longitudinal clinical assessment involves assessing and recording student performance in real clinical environments. However, instead of being limited to a single encounter, student performance is evaluated by multiple assessors, over an extended period and within multiple contexts. Assessments can cover a variety of skills and attributes, such as technical clinical skill, communication, and professionalism. Evaluations are collated to provide a rich data source on student performance, as opposed to single encounter evaluations which may only record "best day" or "worst day" performances (Albino et al., 2008). This form of assessment has also been referred to as "continuous assessment" - particularly within nursing education (Neary, 2000; Stuart, 2007; Royal College of Nursing, 2017) - and "observation on clinics" (Kramer et al., 2009; Williams et al., 2015). However, for the purposes of this thesis, the format will be referred to as "*longitudinal clinical assessment*" (as previously mentioned in section 1.4.1).

Longitudinal clinical assessment is aligned with the "Does" level of Miller's triangle of clinical competence (Albino et al., 2008; Williams et al., 2015) (see section 1.4.1). Individual assessments are well suited to formative assessment since assessors are expected to provide students with detailed feedback on their performance. However, several authors have suggested that longitudinal assessment could also be used for summative assessment (Prescott-Clements et al., 2008; Williams et al., 2015; Dawson et al., 2017). This proposal is based on the concept that, since faculty are provided with an extensive pool of information, they are better equipped to make judgements on whether students can synthesise the fundamental knowledge, skills and behaviours needed to treat a range of patients who require a range of treatments of varying difficulty (Dawson et al., 2017). However, there are currently no publications which verify if longitudinal clinical assessment has been used in this manner within dental education. Some publications indicate that longitudinal clinical performance

data have been used for summative assessment within nursing (Neary, 2000) but there is little detail on whether this is common practice across nursing education or if it has been successful.

As discussed above, the main advantage of longitudinal clinical assessment is that it produces an extensive data base on student performance in real clinic environments over a prolonged period through multiple assessors. However, the involvement of multiple assessors means there is potential for discrepancies to arise between individual assessors as marking can be very subjective (Kramer et al., 2009; Crossley et al., 2011; Williams et al., 2015), lack standardisation and may be influenced by the "halo effect" (Williams et al., 2015) (see section 1.4.8).

As a result, longitudinal clinical assessment has been said to have low reliability (Williams et al., 2015) – but there is currently no evidence to support this claim or demonstrate the degree of disparity between assessors within longitudinal clinical assessment and how its reliability is affected. A study by van der Vleuten et al. (2010) has proposed that subjectivity among assessors should be counterbalanced by sampling across a wide range of assessors. However, some students may only be assessed by a select number of faculty over the duration of the course and, therefore, may be marked more predominantly by assessors who are stricter or more lenient. In theory, discrepancies between assessors (and therefore reliability) could be improved with the use of structured marking schemes/checklists (Williams et al., 2015) and through assessor training and calibration, but there are currently no studies which have tested these proposals for longitudinal clinical assessment.

Prescott-Clements et al. (2008) previously investigated the validity of Longitudinal Evaluations of Performance (LEPs) - a form of longitudinal clinical assessment used within Scottish postgraduate DVT schemes (see chapter 3, section 3.5.3.2 for further details). Based on trajectories generated from two Vocational Dental Practitioner (VDP) cohort's (n = 201) LEP data - which demonstrated an increase in VDP performance over the duration of a DVT year - and the options of approximately one-hundred DVT trainers on longitudinal assessment, the study concluded LEPs were a valid form of assessment within the context of DVT (Prescott-Clements et al., 2008).

A more recent study by Dawson et al. (2021) used longitudinal clinical assessment data to explore the validity of using numerical requirements to determine development of clinical competence among dental students. The study - which was based on over 50,000 longitudinal clinical assessment data points for direct restorations across two student cohorts (n = 139) at the University of Liverpool - concluded that the number of direct restorations completed by students should not be exclusively relied upon to determine whether students had developed into competent practitioners, as it did not necessary signify that students have obtained a sufficient breadth of clinical experience. Instead, it was suggested that there should now be a shift towards determining the role of consistency of performance in competence assessment, which was also calculated and explored as part of study. Whilst Dawson et al. (2021) have contributed some early evidence for the validity of longitudinal clinical assessment, there is a need to conduct further meaningful studies on its application within undergraduate dental education and accumulate more evidence on its validity.

1.5 Rationale for study

1.5.1 Current gap in the existing literature and dental education research priorities

Although many of the assessment methods discussed in the previous section (1.4) are well established within dental education, there remains an ongoing debate on which method(s) are best for measuring the development of competence (Dawson et al., 2017). The need for further research into assessment within dental education has been recognised by the Scottish Oral Health Research Collaboration (SOHRC) who, in December 2014, completed a priority setting exercise (Delphi) to establish a basis for a focussed dental education research strategy within and between the Scottish Dental Schools and NHS Boards (Ajjawi et al., 2017). The results of the exercise were presented at the first SOHRC conference in February 2015, where the top three priorities were identified as:

1. The role of assessments in identifying competence.

2. Ensuring that the undergraduate curriculum prepares for practice.

3. Promotion of teamwork within the dental team.

Despite the current lack of published evidence, several authors have proposed that longitudinal assessment (section 1.4.11) is one of the strongest methods for assessment of clinical skills (Albino et al., 2008; Dawson et al., 2017; Patel et al., 2018). From a theoretical perspective, the rationale behind these proposals is understandable since longitudinal clinical assessment should allow patterns of activity and performance to be established from many data points compiled over a prolonged period. Many other assessment methods discussed in section 1.4 are standalone (or "one-off") assessments, which may not be best suited for measuring attainment of some of the GDC's LOs; particularly when evidence of development of ***consistent*** competent clinical performance is necessary to demonstrate that students have developed into "safe beginners".

Some dental schools have incorporated longitudinal clinical assessment systems into their undergraduate assessment repertoire to allow student knowledge and skills to be evaluated at multiple points in time throughout the curriculum. However, although longitudinal clinical assessment can create rich data sets on student performance, there are currently few robust evaluations on the validity of these systems using objective outcome measures within dental education, which has been recognised by the SOHRC. Therefore, this topic merits further investigation.

Determining whether students have achieved the required clinical competencies is a high stakes decision and therefore, if longitudinal clinical assessment were to be used for competence assessment, it requires sufficient supporting evidence. An argument-based approach encourages accumulation of evidence from various sources and the inclusion of investigations on which types of validity can be attributed to the assessment method (see sections 1.3.3.2 and 1.3.3.3) (Kane, 2013).

1.5.2 Previous pilot study on validity of longitudinal clinical assessment

In accordance with the first of the SOHRC's research priorities (see section 1.5.1 above), the lead researcher of this thesis (i.e., the PhD researcher) previously

conducted a pilot study which attempted to contribute early evidence to a validity argument on the use of longitudinal clinical assessment data for competence assessment. A further aim of the study was to generate potential lines of enquiry for further research (Dickie, 2017).

Clinical performance data were obtained from an electronic longitudinal assessment system, known as LIFTUPP© (see chapter 3, section 3.5.3.2 for further details on this assessment method), which has been adopted by several UK dental schools. Thirteen dental students' LIFTUPP© data were analysed and formatted (into barcode graphics and line graphs) to establish and illustrate patterns of development and then compared to outcomes obtained from a simulated standalone clinical competence test and faculty subjective opinion. Qualitative and quantitative evaluations were made to determine if there was any association between longitudinal clinical assessment data patterns and the results obtained from the standalone competence test and faculty subjective opinion.

Overall, the study showed that longitudinal clinical assessment data appeared to offer a richer collection of data on student development compared to the standalone competence test and faculty subjective opinion - both of which yielded several inconsistencies in terms of assessment. Sufficient evidence for the validity of longitudinal data in the assessment of clinical competence could not be determined due to scale and timeframe constraints. As a result, it was suggested further investigation would require several alternative approaches.

Following the completion of the pilot study, it was recognised that further studies should, in accordance with Kane (2013), specify which subtypes of validity (see section 1.3.3.2) are being investigated and attributable to longitudinal clinical assessment. Additional details on these considerations and how they influenced the development of the research questions for this thesis are presented in chapter 2 (section 2.2).

It was also identified that to facilitate further research in this area, a means of summarising longitudinal clinical assessment data using robust statistical methods was needed. This process could be challenging with respect to longitudinal clinical assessment since the data sets may be very large with

assessment data for multiple skills and attributes measured repeatedly over time for individual students.

1.5.3 Modelling longitudinal clinical assessment data

1.5.3.1 Generalised linear modelling

Several approaches for modelling longitudinal data have been suggested within the literature. Tang, He and Tu (2012) proposed the most commonly used techniques are generalised linear mixed-effects modelling (GLMM) and weighted generalised estimating equations (WGEE) – which are both derivatives of generalised linear modelling (GLM) (Lin et al., 2016). These specialised modelling techniques have been recommended since they address two problems typically associated with longitudinal data sets. The first issue is that, since longitudinal data sets generate a series of correlations from multiple assessments conducted on same subjects/participants, traditional cross-sectional data analyses (e.g., linear, and logistical regression) cannot be applied. The second issue is that longitudinal data sets often contain missing data since the studies from which they are generated take place over a long time (Lin et al., 2016). Both GLMM and WGEE take these issues into consideration when interpreting data and allow complex biological, psychological, and behavioural changes to be tracked over time (Tang, He and Tu, 2012; Gunzler et al., 2014).

However, whilst dental student longitudinal clinical assessment data could be modelled using these approaches, it is anticipated that dental students are likely to be following different clinical development patterns. Distinguishing different patterns of dental student progression is desirable since it may identify groups who are not developing as expected (and may require remedial training) and will facilitate comparisons between longitudinal clinical assessment data and the outcomes of other assessment methods to establish criterion validity (see section 1.3.3.2).

GLM techniques do not distinguish different patterns of development as part of model generation process. If GLMs were to be used to investigate different longitudinal clinical development patterns, students would need to be

subjectively categorised into groups prior to data modelling. Each group's data could then be modelled separately and compared. However, this process may be challenging when faced with large data sets and it may be difficult to separate students if any differences between them are subtle. Therefore, it would be preferable to remove any subjectivity and categorise students via an automated procedure.

Automated modelling techniques which can identify different groups and plot their development trajectories more objectivity using the available data are discussed in the following section (1.5.3.2).

1.5.3.2 Latent class analyses

Growth mixture modelling (GMM) (Muthén and Shedden, 1999; Muthén, 2001) (also known as latent growth mixed modelling (LGMM) or latent class growth modelling (LCGM)) and group-based trajectory modelling (GBTM) (Nagin, 2005) are two forms of latent class analysis which were originally developed to track groups of individuals following similar patterns of behaviour or achievement of outcome measures over time within psychology and criminology.

Examples of psychological studies which have used GMM are (Orcutt, Erickson and Wolfe, 2004; Dekker et al., 2007; Mora et al., 2009). The former (Orcutt, Erickson and Wolfe, 2004) tracked post-traumatic syndrome disorder (PTSD) symptoms over time to demonstrate that Gulf War veteran's responses to trauma suffered during conflict were not homogenous. The latter two studies (Dekker et al. (2007) and Mora et al. (2009)) both used GMM to identify different trajectories of depressive symptoms in children and pregnant women, respectively. Both studies concluded that distinguishing different behavioural patterns within their respective populations facilitated identification of those most at risk of depressive illness and, therefore, were in greater need of intervention. This information is valuable in determining where resources and services should be concentrated.

Psychological studies which have used GBTM to track aggression in children include Jester et al. (2008) and Girard et al. (2019). The former identified three different trajectory groups of aggressive behaviour, which were proposed as

means of highlighting which children may be at increased risk of drug abuse in later life. The latter generated five trajectory groups and investigated which factors may have influenced the risk of group membership for each group.

In criminology, GBTM has been used to identify urban areas where crime is more concentrated and, therefore, which areas should be targeted for intervention. It has also been suggested that GBTM could be used to track how levels of crime may change following intervention and/or which areas appear to be "stable" (Weisburd et al., 2004). GBTM has also been used to track various criminal behaviours over time – such as individual aggression, gang membership and arrest history (Nagin and Piquero, 2010).

Both GMM and GBTM are specialised forms of finite mixture modelling and share a common analytical objective: To explain the difference in developmental trajectories across a population. They both typically use the polynomial functions of age or time as part of their modelling process, similar to another form of latent class analysis known as growth curve modelling (GCM) (Bollen and Curran, 2006). However, GCM assumes that all individuals within the population follow a similar trajectory (Nagin and Odgers, 2010) and therefore does not use the available data to distinguish different groups.

The major difference between GMM and GBTM is that the former assumes there are two or more group trajectories within the population, and there may be random effects within them (Jung and Wickrama, 2008; Nagin and Odgers, 2010). In contrast, GBTM does not take random effects within the trajectories into consideration, nor does it assume the shape or number of groups present. Instead, the trajectory groups are used as a statistical means of approximating the unknown distribution of trajectories across population members (Nagin and Odgers, 2010). The statistical processes and output provided by GBTM also allows the shape and number of group trajectories which fit the data best to be evaluated, and the probability of trajectory group membership for individuals to be estimated (Nagin and Odgers, 2010). As a result, GBTM is a more flexible method of latent trajectory modelling.

Although originally developed for psychology and criminology, GBTM is now increasingly applied within other fields of research - including primary and

secondary education (Melhuish et al., 2011; Sutcliffe, Gardiner and Melhuish, 2017) and clinical dentistry (Broadbent, Thomson and Poulton, 2008; Thomson et al., 2013; Shearer, 2016). Like the studies in psychology and criminology cited above, studies in primary and secondary education and clinical dentistry have also proposed GBTM was an effective method for identifying groups of individuals following similar development trajectory patterns and investigating possible determinants for trajectory group membership.

Melhuish et al. (2011) modelled and tracked educational progress across primary and secondary education and found that pre-school attendance had a positive influence on attainment and progression throughout primary and secondary education – especially in children who came from disadvantaged backgrounds. Sutcliffe, Gardiner and Melhuish (2017) also analysed educational progress but for "looked-after children" (i.e., children who have been in the care of a local authority). They advocated GBTM highlighted factors associated with temporal changes in educational progress trajectory patterns (e.g., the age at which a child first enters care and the duration they were cared for). In clinical dentistry, GBTM has been used to follow caries (tooth decay) experience (Broadbent, Thomson and Poulton, 2008) and periodontal (gum) disease experience (Thomson et al., 2013) over time, as well as associations between periodontitis and blood sugar levels (Shearer, 2016). Broadbent, Thomson and Poulton (2008) used GBTM to map and describe caries developmental patterns in permanent teeth up to the age of 32, whilst Thomson et al. (2013) and Shearer (2016) used their resulting trajectories to identify that smokers were at greater risk of periodontal disease and periodontitis was not associated with elevated blood sugar levels (dysglycaemia) between the ages of 26 and 38, respectively.

Whilst each of these studies supported the use of GBTM in their respective fields, they may have been limited by the data available for their study participants. The number and appearance of trajectories generated by GBTM depends on multiple factors - including the prevalence of observations (Frankfurt et al., 2016), sample size (Nagin and Tremblay, 2001), length of follow-up (Eggleston, Laub and Sampson, 2004) – which, in turn, will influence which trajectory group participants are assigned to (and therefore the conclusions drawn).

Although the trajectories generated by Sutcliffe, Gardiner and Melhuish (2017), Thomson et al. (2013) and Shearer (2016) were based on data from many participants (47,500 school pupils, 1,037 and 893 patients, respectively), data were only available for three time points per participant. Sutcliffe, Gardiner and Melhuish (2017) used assessment scores from English national school tests in literacy and numeracy, which are taken at ages 7, 11 and 16. Similarly, Thomson et al. (2013) and Shearer (2016) used periodontal examination data recorded at ages 26, 32 and 38. The latter also used blood sugar level (HbA1c) data recorded at these ages. If more observations per participant were available, additional trajectory groups and shapes may have been found. A minimum of two time points can be used to produce linear trajectories, however if non-linear trajectories (e.g., quadratic) were to be identified, a minimum of four time points is required (Frankfurt et al., 2016).

Broadbent, Thomson and Poulton (2008) generated trajectories from six time points (dental examination data recorded at ages 5, 9, 15, 18, 26, and 32). However, the findings may have been different if data beyond the age of 32 were incorporated (i.e., the study was prolonged). Collectively, the trajectory groups generated by this study, and those cited above, serve as examples on how the models generated by GBTM depend on the amount of data available for analysis – although the same could be said for non-GBTM modelling approaches. These studies analysed secondary data (i.e., data which had been originally collected for other purposes) retrospectively and, therefore, were limited to analysing what data were available. In general, more participants, more observations and (where desirable/appropriate) longer studies should result in more optimal and accurate trajectory models. These data-related features should be considered in the design of future GBTM studies where possible – especially for prospective approaches.

Like other data modelling techniques, it is also worth noting that GBTM has functional limitations (Frankfurt et al., 2016) – particularly since it serves as a method for conveniently analysing complex data sets. These limitations are discussed in greater detail in chapter 9 (section 9.7) at the end of this thesis.

However, despite its limitations, GBTM may facilitate early explorations on the validity of longitudinal clinical assessment data by providing a more flexible

form of latent class analysis (Nagin, 2005) which can summarise large quantities of longitudinal clinical assessment data into simple - but meaningful - graphics and tabulations. Furthermore, since participants' trajectory group memberships can be regarded as developmental outcomes (Nagin and Piquero, 2010), comparisons with other educational assessment outcomes (such as undergraduate examination and postgraduate clinical assessment data) will be facilitated.

Longitudinal clinical assessment data generated by dental students have previously been modelled using Bayes' theorem by Roudsari (2017). Bayes' theorem serves as another example of latent class analysis which tracks the probability of an outcome occurring over time based on prior information that can influence the outcome (Lindley, 1958; Stanford Encyclopedia of Philosophy, 2003). In Roudsari's (2017) study, the outcome was competent clinical performance for extraction of a tooth, which were recorded via a longitudinal assessment system (LIFTUPP© - chapter 3, section 3.5.3.2). Using Bayes' theorem, the study was able to model longitudinal clinical assessment data at cohort level and demonstrated a general upward trend in student clinical performance. Whilst this showed that, collectively, student clinical performance (for tooth extractions) appears to improve over the BDS programme, there remains a need to model longitudinal clinical assessment data in a manner which distinguishes different patterns of student progression within a cohort.

1.5.4 Additional research opportunity

Processing undergraduate examination outcomes so they can be compared with longitudinal clinical assessment would also present an opportunity to explore how yearly examinations function as a form of longitudinal assessment across the duration of the BDS course. This is a line of investigation which has not previously been covered in the literature.

1.6 Summary

This chapter has highlighted how valid clinical assessment within dental education is essential to ensure UK dental students have developed into "safe beginners", which is the standard expected by the GDC at the point of

graduation. The GDC delegate the responsibility of assessment method selection to UK dental schools, who have adopted a range of assessments to demonstrate their graduates have achieved the GDC's LOs and are eligible for entry onto the professional register.

Following a narrative review of the existing literature on assessment methods commonly used by UK dental schools, there was evidently a lack of research on the use of longitudinal clinical assessment within dental education. Even though longitudinal assessment has been advocated as a strong method for clinical subjects, there are currently few readily available publications which have investigated its validity for assessing competent clinical performance in undergraduate dental students using robust statistical methods.

To generate evidence there is a need to test a validity argument. Through testing a validity argument, the weaknesses of some previous assessment validity studies should be avoided. This approach allows sufficient evidence for the various types of validity attributable to longitudinal clinical assessment to be determined and developed, which will give dental schools the information needed to make a balanced and informed judgement as to whether longitudinal clinical assessment is a "good" assessment method (i.e., fit for purpose) – or not.

Chapter 2 - Aim, objectives, and research questions

This chapter presents the overall aim and objectives of this study before concluding with specific research questions.

2.1 Aims and objectives

In order to contribute to a validity argument on the use of longitudinal data for assessing the development of clinical competence in undergraduate dental students, this study aims to investigate the content and criterion validity and reliability of longitudinal clinical data. To meet these aims, the study has the following objectives:

- Compare outcomes of early undergraduate BDS examinations with those from the final BDS examinations.

- Establish the usefulness of GBTMs at modelling longitudinal clinical data and determine patterns of clinical performance over time (content validity).

- Compare patterns of undergraduate clinical performance from longitudinal clinical assessment with the outcomes of undergraduate examinations (criterion validity - concurrent).

- Compare patterns of undergraduate clinical performance from longitudinal clinical assessment with postgraduate clinical performance (criterion validity – predictive).

- Test the reliability for the panel of assessments used within undergraduate assessment (longitudinal and examinations).

- Ascertain views of key stakeholders on the findings from the above analyses and make recommendations on how to enhance the assessment process.

2.2 Research questions

To achieve the above aims and objectives, the following research questions were designed.

Question 1: How does student performance in early BDS year examinations relate to their performance in the final professional BDS degree examinations?

Question 2a: What are the main patterns of longitudinal clinical assessment over time within a BDS year, across BDS years and across cohorts? (Content validity)

Question 2b: What is the association between undergraduate longitudinal clinical assessment and standalone assessment methods? (Criterion validity – concurrent)

Question 2c: What is the association between undergraduate longitudinal clinical assessment and postgraduate assessment? (Criterion validity – predictive)

Question 3: According to key stakeholders in dental education, how might the findings of research questions 2a, 2b and 2c be used to enhance assessment in dentistry?

2.3 Summary

This chapter has outlined the aims and objectives of the study and presented the research questions to be answered. The following chapter describes and justifies the methodological approach and specific methods adopted to address the research questions.

Chapter 3 - Research study design

3.1 Introduction

The following chapter describes the methodological approach taken for this study. It provides details on the philosophical theory (epistemology), methodological frameworks and methodology chosen to address the research questions proposed in chapter 2 (section 2.2). Techniques (methods) applied uniformly to the data sets in preparation for answering the research questions are also described. Subsequent chapters provide further details on methods adopted to specifically answer each research question.

The overall purpose of the project was to contribute evidence to a validity argument on the use of longitudinal clinical performance data for the assessment of dental student competence. To ensure priority was given to the research questions, a pragmatic epistemology was adopted, as it prevented the study from being confined to the rules and methods of a single philosophical paradigm. Instead, the researcher was free to choose the methods and/or procedures that were most appropriate for answering the research questions (Patton, 1990; Morgan, 2007; Tashakkori and Teddlie, 2010; Morgan, 2014; Bryman, 2016; Cohen, Manion and Morrison, 2018).

Quantitative methods were chosen to address research questions 1, 2a, 2b and 2c and a qualitative approach was chosen for question 3. Therefore, the study serves as an example of mixed methods research. Data obtained through each of the quantitative and qualitative components were influenced either solely by the methodological frameworks of interpretivism (Robson and McCartan, 2016) and post-positivism (Carson et al., 2001), or through a combination of both.

A diagrammatic summary of the research methodology is shown in Figure 3.1. Each component of the methodology in this Figure (i.e., the chosen epistemology, methodological frameworks, methodologies, and methods) are discussed in further detail with reference to the literature over the following sections.

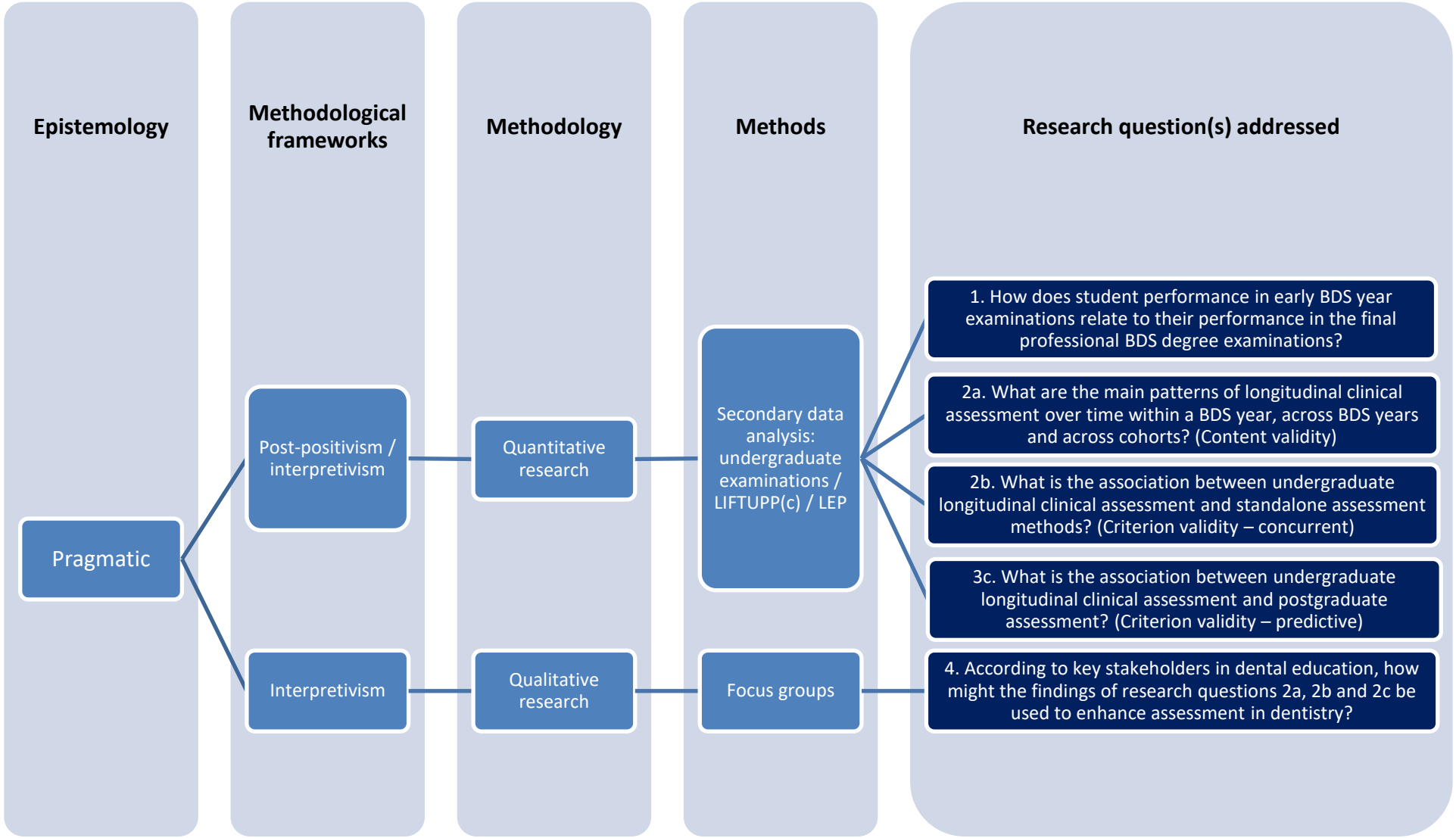


Figure 3.1 – Schematic overview of study design. BDS = Bachelor of Dental Surgery. LEP = Longitudinal evaluation of performance.

3.2 Epistemology

A pragmatic epistemology was adopted to give priority to the study research questions. Pragmatism is a philosophical viewpoint that does not conform to any one system of philosophy or reality. Instead, pragmatic researchers focus their attention on the research problem(s) at hand (Patton, 1990; Morgan, 2007; Tashakkori and Teddlie, 2010; Morgan, 2014) and draw upon a variety of data collection and analysis methods to understand and address the problem(s) (Rossman and Wilson, 1985; Creswell and Creswell, 2018). This ensures the researchers choose methods which are most likely to answer and/or provide insight to the questions raised in the study rather than being confined to using methods that are associated with certain paradigms (Johnson and Onwuegbuzie, 2004).

As a result, pragmatism can provide an appropriate philosophical framework for mixed methods research (Tashakkori and Teddlie, 1998; Johnson and Onwuegbuzie, 2004; Morgan, 2007). Some authors stipulate pragmatism is the paradigm which potentially provides the best philosophical framework for mixed methods (Greene, 2008) – however, it should be noted that mixed methods can be aligned with other paradigms (e.g., transformatism) (Mertens, 2005).

The use of both quantitative and qualitative data to build the best understanding of a problem within mixed methods research coincides with the pragmatist view that the nature of the real world cannot be revealed solely by a single scientific method (Mertens, 2005). Instead, “truth” is discovered through application of the most practical solutions and with “what works at the time” (Creswell and Creswell, 2018) - a perception that indicates “truth” is context dependent.

The dependence of context in relation to “truth” is further apparent through pragmatism’s rejection of “objective truths” that result from actions. This is because pragmatists believe that actions themselves cannot be disconnected from the circumstances in which they occur (Morgan, 2014). Meaning can only be derived from observing outcomes that have resulted from the same actions occurring across different contexts. However, since contexts will continually change (across studies, between institutions, in different environments etc.),

some prominent pragmatists have stopped short of proclaiming that knowledge obtained through the observation of actions can never be labelled as a definitive picture of reality, i.e., the findings can only be "temporal" as subsequent discoveries may lead to changes in our understanding. An example of this implication comes from Dewey (1941), who proposed that, due to the transitional nature of context, knowledge obtained through a pragmatic approach should be referred to as "warranted assertions" (Biesta, 2010). Ultimately, this means the same outcome(s) resulting from an action are not guaranteed to be consistent across different contexts. The same outcome(s) might occur - but this will only be known if the same action is observed across different contexts.

Constantly changing context was relevant to this study as it involved comparing assessment methods used by specific organisations at a point in time. How these assessment systems operate, are used, and applied may change over time and/or between organisations. They may also eventually be replaced with alternatives. However, by studying these assessment systems within one context, it allows them to be compared across differing contexts and determine if there is any consensus. Collectively, this will improve our understanding of the underlining research problem(s).

Another key feature of pragmatism, in relation to this study, is it recognises other philosophical stances may be antagonistic to one another (e.g., positivism versus interpretivism) (Biesta, 2010). This issue can (but doesn't always) occur within mixed methods research since the quantitative and qualitative components of the study can be influenced by different methodological frameworks (e.g., post-positivism and interpretivism – see section 3.3). Rather than accept that research should be restricted to using only tools associated with a methodological framework (e.g., qualitative methods for interpretivist studies and quantitative methods for post-positivist studies), pragmatism advocates quantitative and qualitative research are mutually beneficial and, when combined, can lead to a more profound understanding of the proposed research problems (Greene, 2008; Biesta, 2010; Tashakkori and Teddlie, 2010). This approach ensures the methods best suited to answering the research questions are matched with one another (Mackenzie and Knipe, 2006).

3.3 Methodological frameworks

Research questions 1, 2a, 2b and 2c were addressed through quantitative analysis of data obtained through undergraduate longitudinal clinical assessment, undergraduate BDS examinations and postgraduate longitudinal clinical assessment (see section 3.5.3). These analyses were predominantly influenced by post-positivism – a methodological framework commonly associated with quantitative studies (Creswell and Creswell, 2018).

Post-positivism incorporates the view that evidence obtained via research is always imperfect, as we cannot be certain that knowledge acquired through studying human behaviour and action is correct. The conjectural nature of knowledge means that we can only be guided by the best evidence available at the time of study, and our views may change or evolve as better evidence becomes available. Building the best evidence involves using a "scientific approach", whereby researchers start with a theory, collect data that either supports or contests the theory, and then offer suggestions for revisions that can be applied in further testing (Robson and McCartan, 2016).

Qualitative analysis was chosen to answer research question 3 through obtaining the opinions of key stakeholders within dental education on the results produced by the quantitative component of the study (see section 3.5.4). Deriving meaning from subjective opinions meant this element of the study was conducted under the influence of interpretivism (Crotty, 1998).

Methodologically, interpretivism (also referred to as constructivism) avoids rigid structural research frameworks, permitting a more flexible approach to determine perceptions of reality (Carson et al., 2001). Although interpretivism largely relies on qualitative data collection, it can also be used for mixed methods studies. In such cases, quantitative data are typically used to support or expand upon the qualitative findings and help deepen their description (Mackenzie and Knipe, 2006). However, in the case of this study, quantitative data will not be used to support the qualitative findings - they will be used to prompt discussions that generate qualitative data. This approach was originally described by Greene, Valerie and Caracelli (1989) and has evolved to become known as "explanatory sequential design" – where the quantitative component is

conducted first and is then followed by the qualitative component which builds on and/or helps to explain the results obtained by quantitative component (Morse, 1991; Creswell, 2003; Byrne and Humble, 2006). This research design has been application in a variety of fields, including social and behavioural science (Janz et al., 1996), health care (Neri and Kroll, 2003), education (Buchwitz et al., 2012) and engineering (Wipulanusat et al., 2020).

3.4 Methodology

Research questions 1, 2a, 2b and 2c are concerned with the evaluation and/or comparison of assessment information accumulated by undergraduate longitudinal clinical assessment, undergraduate examinations, and postgraduate longitudinal clinical assessment. All these sources produce numerical data in the form of results/marks and grades which (at the time of this study) have not been previously compared using robust statistical methods. Accordingly, quantitative analysis methods were used to address research questions 1, 2a, 2b and 2c using quantitative analysis methods (see section 3.5.3).

On the other hand, research question 3 relates to the collection, synthesis, and analysis of individuals perceptions of the results from research questions 1, 2a, 2b and 2c and, therefore, is aligned with a qualitative approach.

Mixed methods studies intentionally use both quantitative and qualitative input within a single study to reflect various aspects of the issue in question (Leech and Onwuegbuzie, 2009). Instead of favouring one approach, mixed methods researchers consider quantitative and qualitative methods as complementary to one another (Krathwohl, 1993; Creswell, 2003; Thomas, 2003). The quantitative aspect uses statistical, mathematical, or computational techniques to conduct a systematic, empirical investigation of observable phenomena (Babbie, 2010; Muijs, 2010), whereas the qualitative aspect is used to gain an understanding of people's thoughts and feelings (Sutton and Austin, 2015). Both approaches have advantages and disadvantages, but, when used in conjunction, the weaknesses of each can be potentially mitigated, which consequently gives more comprehensive support when attempting to establish validity (Creswell and Plano Clark, 2011).

Furthermore, mixed methods may be more suited to increasing the reliability and applicability of findings from research that investigates complex themes, concepts, contexts, and initiatives with medical education (Schifferdecker and Reed, 2009).

By using the research questions as the basis for adopting a mixed methods methodology, this study was designed in a "bottom-up" fashion – an approach previously described by Tashakkori (2006). This opposes a "top-down" approach, where the research questions do not drive the use of mixed methods and instead it is the researcher's intention to use mixed methods from the outset of their study design (Mertens, 2003). However, Johnson, Onwuegbuzie and Turner (2007) claimed that each mixed methods study sits somewhere along a spectrum polarised by the "bottom-up" and "top-down" approaches.

In addition to discussing how some studies have been defined by their design approach, Johnson, Onwuegbuzie and Turner (2007) also discussed how other studies have defined the point at which differing data sets (i.e., the quantitative and qualitative data sets) were mixed within the study. They described a variety of ways in which studies have mixed their data sets, ranging from studies that mix either during the data collection or data analysis stages, to those where mixing occurs during at all stages of the study.

For this study, no mixing occurs during the data collection or analysis. Instead, quantitative data were first collated and analysed independently. The results of the analyses were then presented for discussion among key stakeholders to generate qualitative data (see section 3.5.4).

3.5 Methods

3.5.1 Types of validity to be investigated

As outlined in chapter 1 (section 1.5), there is currently a need to build a validity argument on using longitudinal clinical assessment to assess the development of clinical competence in undergraduate dental students. Building a validity argument is a systematic process of accumulating evidence to support the intended interpretation of assessment data (in this case – longitudinal

assessment data) for a proposed purpose (in this case – assessment of clinical competence) (Kane, 2013).

The process should obtain evidence on whether longitudinal clinical assessment is valid in measuring competence using various types of validity (and therefore the development of undergraduate dental students into "safe beginners" (GDC, 2015a)). Out of the four main types of validity (face, content, criterion, and construct - see chapter 1, section 1.3.3.2), this study will focus on generating evidence for content and criterion validity in measuring competent clinical performance. Although this study does not primarily seek to create evidence for construct validity, the results generated may contribute to future studies focused on establishing construct validity for longitudinal clinical assessment since building evidence for this type of validity requires consolidation of information from a variety of sources and perspectives.

Since longitudinal assessment of clinical performance has already been integrated into the undergraduate curricula of some UK dental schools (chapter 1, section 1.5.1), it is reasonable to assume it has face validity. Senior academics within dental schools will have deliberated on whether longitudinal clinical assessment appears to measure what it intends to measure and compared it against standards and quality assurance guidelines provided by the GDC (2015b; 2019a) before adopting it as an assessment tool. If subject experts were involved in these evaluations and agreed that longitudinal assessment could be used to assess competent clinical performance, then content validity could also be attributable. However, this is a bigger assumption to make as there is currently little information readily available within the public domain which outlines how each UK dental school selects and integrates new assessment methods, nor how longitudinal assessment is used within each dental school. Therefore, there remains a need to build more evidence for content validity (on assessment of clinical competence) using robust methods. Criterion validity (including the concurrent and predictive subtypes) also requires investigation.

3.5.1.1 Content validity

To investigate content validity (which was previously defined in chapter 1, section 1.3.3.2), longitudinal patterns of clinical performance for undergraduate

dental students over time will be established. By determining what patterns of clinical development exist within longitudinal data sets, we will gain insight on whether longitudinal assessment can demonstrate development of competent performance and distinguish differences in performance among students.

Patterns of clinical performance could be established through statistical modelling techniques. As previously discussed in chapter 1, this process may be challenging with respect to longitudinal assessment, however GBTM could provide a promising approach (see chapter 1, section 1.5.3.2).

Investigating the evidence for content validity is aligned with research question 2a (see chapter 2, section 2.2).

3.5.1.2 Criterion validity

To investigate criterion validity (which was also defined in chapter 1, section 1.3.3.2), the association between longitudinal and current dental education assessment methods will be explored. Comparisons with well-established undergraduate assessment methods (e.g., OSCEs and written examinations (MSA and MCQ papers)) will determine if longitudinal clinical assessment data have concurrent validity (a criterion validity subtype – see chapter 1, section 1.3.3.2). Comparisons with postgraduate assessment will evaluate whether longitudinal clinical assessment data have predictive validity (another criterion validity subtype – see chapter 1, section 1.3.3.2), which would be of value since the aim of any assessment at the point of student graduation must be to predict future performance in practice.

Investigating the evidence for criterion (concurrent and predictive) validity is aligned with research questions 2b and 2c (see chapter 2, section 2.2).

3.5.1.3 Reliability

The reliability of longitudinal clinical assessment will be investigated and compared against established undergraduate assessment methods. This enquiry is necessary since valid assessments must display a degree of reliability (Beanland et al., 1999; Polit and Hungler, 1999) (see chapter 1, section 1.3.3.2). Therefore, the results of this enquiry will contribute to the investigation of

criterion validity – specifically, the concurrent subtype (research question 2b – see chapter 2, section 2.2).

3.5.1.4 Additional investigations

As per chapter 1 (section 1.5.4), processing and analysing undergraduate examination outcomes as part of the investigations described above also presents an opportunity to explore how yearly examinations function as a form of longitudinal assessment across the duration of the BDS course. Therefore, early BDS year (i.e., BDS1-3) examinations outcomes will be compared to the outcomes of the final BDS professional examinations (held in BDS4/5).

These addition investigations are aligned with research question 1 (see chapter 2, section 2.2).

3.5.2 Key stakeholder opinion

The lines of enquiry described above will contribute quantitative evidence to a validity argument. However, qualitative input will also be of value since it will provide complementary data through triangulation. Discussions with key stakeholders could confirm whether the findings generated from quantitative investigations reflect their own experiences of clinical assessment within dental education, which will contribute further evidence for content validity since "expert" opinions are typically used to investigate this type of validity (see chapter 1, section 1.3.3.2). Therefore, their opinions will also contribute towards answering research question 2a (see chapter 2, section 2.2).

Furthermore, key stakeholder opinion will be used to identify areas for enhancement of assessment in dental education and generate future lines of enquiry for subsequent research. These investigations are aligned with research question 3 (see chapter 2, section 2.2).

3.5.3 Methods for addressing research questions 1, 2a, 2b and 2c

3.5.3.1 Design, setting and eligible participants

Data from three longitudinal cohorts of students were available. Cohorts 1, 2 and 3 each comprised of individuals who graduated BDS from the University of Glasgow in 2017, 2018 and 2019, respectively.

Longitudinal assessment of clinical skills was progressively "rolled out" during the 2014/15 academic term in Glasgow, meaning the first cohort of students to complete a full three years of assessment were the graduating class of 2017. Therefore, cohorts 1, 2 and 3 started BDS1 during academic terms 2012-13, 2013-14 and 2014-15 (respectively), and completed BDS5 (the final year of the curriculum) at the end of terms 2016-17, 2017-18 and 2018-19, respectively.

This study is exploratory in nature, as LIFTUPP© was only adopted at the University of Glasgow Dental School in 2014, resulting in only three full cohorts of students experiencing this form of assessment, and no prior data from which to estimate a minimum sample size to test specific hypotheses. Consequently, there was no requirement for a specific sample size calculation.

Due to the limited number of individuals who were eligible at the time of study, detection of statistically significant results was not expected. The study will still perform appropriate statistical tests (see section 3.5.3.5) and note p-values, however, with small sample sizes, a p-value >0.05 may not necessarily indicate "no difference" in the population, rather it may be that the sample size was too small to detect the observed "difference" to be statistically significant (i.e., have a p-value <0.05). It was more important to focus on the educational significance of the results, and, therefore, more attention was paid to effect sizes than p-values to ensure the study conclusions were appropriate, actionable, and based on the study results - regardless of their statistical significance.

Results where n<5 will not be reported to maintain participant anonymity.

3.5.3.2 Data sources

Assessment data were collected from three sources, the details of which are provided below. All data were secondary data collected as part of routine under- and postgraduate dental training.

1. ***LIFTUPP©*** - a digital longitudinal assessment system used to record student clinical activity and performance throughout undergraduate dental training. The system is owned by ExamSoft®, who were recently acquired by Turnitin®. All uses of the term "LIFTUPP©" from this point onwards refer to the assessment system itself. It should, however, be noted that the system has become known as "Develop©" in some institutions and across other disciplines, its application having been expanded to medical, veterinary, nursing, and allied healthcare education.

   Originally developed by the University of Liverpool, LIFTUPP© functions as a computer application (app) accessed by supervising clinical staff (i.e., assessors) through iPads (or other web interface technology) to record various aspects of student clinical work (technical skill, professionalism, knowledge and understanding etc.) in real-time. For each clinical activity, students are assigned performance indicators based on a 6-point scale by assessors. The descriptors aligned with the performance indicators in the original 6-point scale – which was used for the first year of use of LIFTUPP© in Glasgow (academic term 2014-15) – were amended for the beginning of academic term 2015-16. Both the original and update 6-point scales are shown in Table 3.1.

   Performance indicators and details of the clinical activity assessed are subsequently uploaded into a database which, by the end of the Glasgow's BDS curriculum, typically contains over 100,000 data points of clinical assessment per student cohort and over 1,500 data points per student.

   The choice of skills and/or behaviours that can be assessed within the system is comprehensive, covering all four key domains of competent clinical practice outlined in the GDC's PfP (GDC, 2015a), i.e., clinical, communication, management and leadership, and professionalism (see

chapter 1, section 1.1.2). However, the clinical domain data entries can be subcategorised into two further levels: dental subject area and procedure. Figure 3.2 provides a summary of the levels of LIFTUPP© data categorisation.

Table 3.1- Original and updated scale of LIFTUPP© performance indicators and their interpretation. Original scale used during academic term 2014-15 at the University of Glasgow Dental School. Updated scale used from academic term 2015-16 onwards and still in use at the time of completion of this study (June 2021).

LIFTUPP© performance indicator	Performance descriptor (Original scale)	Performance descriptor (Updated scale)
1	Currently UNABLE to meet the outcome with the required quality. Has caused harm or does not seek essential guidance.	UNABLE to do this. Has caused harm or does not seek essential guidance.
2	Currently UNABLE to meet the outcome with the required quality. Requires major corrective (procedural) intervention from the tutor.	UNABLE to do this independently at present. Largely demonstrated by tutor.
3	ABLE to meet the outcome at the required quality. Minor corrective intervention (procedural) from the tutor.	UNABLE to do this independently at present but able to complete, to the required quality, with significant help, either procedural or by instruction.
4	ABLE to meet the outcome at the required quality. Minor corrective intervention (verbal) from the tutor.	ABLE to do this partially independently at the required quality, but requires minor help with aspects of the skill, either procedural or through discussion.
5	ABLE to meet the outcome independently at the required quality. Confirmatory advice from the tutor.	ABLE to do this independently at the required quality. This may include confirmatory advice from the tutor where the student seeks appropriate assurance.
6	ABLE to meet the outcome independently, exceeding the required quality. Confirmatory advice from the tutor.	ABLE to meet the outcome independently, exceeding the required quality.

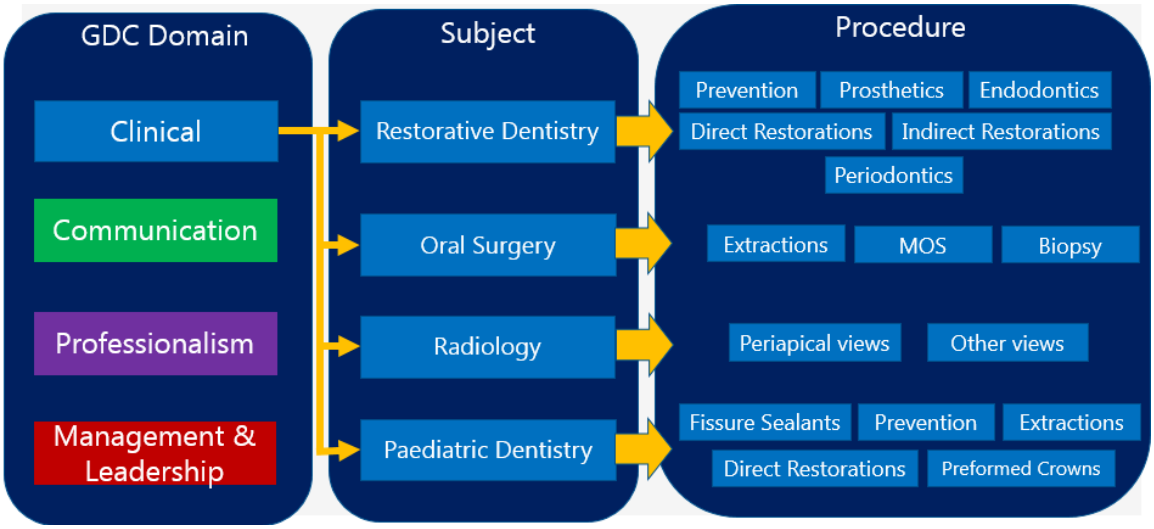


Figure 3.2 – Levels of LIFTUPP© data categorisation.

A full list of variables which can be recorded by LIFTUPP© is provided in appendix 1.

For this study, LIFTUPP© data were only sourced from BDS3, BDS4 and BDS5 for each cohort since these are the years of the course in which Glasgow dental students deliver the majority of their patient care. Therefore, LIFTUPP© data were obtained for the three cohorts between academic terms 2014-15 to 2018-19. These data had previously been collected by the University of Glasgow (the data controller) to formatively monitor student activity and performance with respect to progression and satisfactory completion of the BDS course. As of the 2019-20 academic term, LIFTUPP© data are also being used summatively in Glasgow.

The data are stored on an external server by LIFTUPP© Limited (Ltd) (the company owned by ExamSoft®) and were formatted in Excel spreadsheets (Microsoft Office: 2016. Redmond, WA: Microsoft Corporation). Each cohort's LIFTUPP© data set contains the following participant demographics: forename, surname, student matriculation number, sex, and date of birth.

2. ***Undergraduate professional examinations*** - the results of assessments undertaken by University of Glasgow dental students. Glasgow Dental School currently uses a combination of MCQ and MSA examinations, summative essay assessment and OSCE (see chapter 1, sections 1.4.3, 1.4.5, 1.4.7 and 1.4.9, respectively) to assess the early year groups (i.e., BDS1, BDS2 and BDS3). "Finals" are assessed with an MSA examination and Comprehensive Care

Clinical Case Presentation examination (a form of oral examination – see chapter 1, section 1.4.8) (held in BDS4) and an OSCE (held in BDS5). For this study, the results of these examinations were obtained for the three cohorts between academic terms 2012-13 to 2018-19 and only assessments that were numerically scored were included.

Students who were unsuccessful in passing at the first attempt were eligible to re-sit the examination(s) during additional assessment diets. Although both first attempt and re-sit results were sourced, the latter were not included in the analysis (see section 3.5.3.4).

All examination data had previously been collected by the University of Glasgow (the data controller) as a requirement for progression and satisfactory completion of the BDS course. The data sets were managed by BDS year group administrators and were stored in Excel spreadsheets on University of Glasgow secure password protected servers. Each undergraduate examination data set contained the following participant demographics: forename, surname, student matriculation number, sex, and date of birth.

3. ***Longitudinal evaluations of performance (LEPs)*** – a longitudinal assessment tool used to assess clinical performance in Scottish postgraduate vocational dental training (VDT) schemes.

   VDT is a period of postgraduate training that introduces new graduates to general dental practice in a protected environment, where they are paired with an experienced dentist (trainer) in the same practice. The trainer provides the vocational dental practitioner (VDP) with supervision and help whenever necessary but is also responsible for assessing the VDP's clinical judgement, technical ability, management and leadership skills, professionalism and communication skills which correlate well with the GDC's four domains of clinical competence (GDC, 2015a).

   LEPs serve as work-based assessments that require VDT trainers to directly observe VDP performance in clinical practice and rate their performance against a nine-point scale (Prescott, McKinlay and Rennie, 2001). Guidance

provided to VDT assessors by NHS Education for Scotland (NES) – who are the governing body of Scottish VDT - categorised LEP scores 1-3 as “needs improvement”, 4-6 as “satisfactory” and 7-9 as “superior” (NHS Education for Scotland, Accessed June 2021) (Table 3.2).

Table 3.2 - Scottish vocational dental practice (VDP) longitudinal evaluation of performance (LEP) rating system.

LEP score	Performance descriptor
1, 2, 3	Needs improvement
4, 5, 6	Satisfactory
7, 8, 9	Superior

Within each LEP, VDPs can be rated against one or more of the following eight variables:

- Examination and consultation skills.
- Clinical judgement and diagnosis.
- Technical ability and manual dexterity.
- Communication skills.
- Professionalism.
- Knowledge (level and application – both given as one overall score).
- Organisation.
- Trainee’s insight into performance.

Each VDP must complete forty-two LEPs within the year-long training period across three “blocks”. Sixteen, fourteen and twelve LEPs must be completed

in blocks 1 (August – November), 2 (November – February) and 3 (February – May), respectively. VDPs must also address any “needs improvement” LEP scores by demonstrating improved performance in a similar patient encounter. For example, if a VDP received a “needs improvement score” in performing root canal treatment in a multi-rooted tooth, then they must record another LEP which focuses on this procedure and obtain “satisfactory” or “superior” scores. All recorded LEPs are submitted to a National Review Panel to determine if the VDP has satisfactory completed VDT and is safe and prepared to enter the NHS dental workforce as a fully independent practitioner.

For this study, LEP data were obtained for VDT years 2017-18, 2018-19 and 2019-20. These data had previously been collected and stored by NES (the data controller) to monitor VDP clinical performance and determine if they met the criteria for satisfactory completion of VDT.

LEP data were stored in Excel spreadsheets by NES (the data controller) and contained the following participant demographics: forename, surname, sex, and date of birth.

Table 3.3 provides a summary of all data sources collected for each cohort.

Table 3.3 – Summary of data sources obtained for each cohort per academic year.

Cohort	Academic years							
	2012-13	***2013-14***	***2014-15***	***2015-16***	***2016-17***	***2017-18***	***2018-19***	***2019-20***
1	BDS1 examinations: • Summative written assessment • 2x MCQ • MSA • OSCE • Resit diet results	BDS2 examinations: • MCQ • MSA • OSCE • Resit diet results	BDS3 examinations: • Anatomy examination • MCQ • MSA • OSCE • Resit diet results	BDS4 examinations: • MSA • Case presentation examination • Resit diet results	BDS5 examination: • OSCE	VDT assessment: • LEP	*NA*	*NA*
			LIFTUPP©					
2	*NA*	BDS1 examinations: • Summative written assessment • 2x MCQ • MSA • OSCE • Resit diet results	BDS2 examinations: • MCQ • MSA • OSCE • Resit diet results	BDS3 examinations: • Anatomy examination • MCQ • MSA • OSCE • Resit diet results	BDS4 examinations: • MSA • Case presentation examination • Resit diet results	BDS5 examination: • OSCE • Resit diet results	VDT assessment: • LEP	*NA*
				LIFTUPP©				
3	*NA*	*NA*	BDS1 examinations: • Summative written assessment • 2x MCQ • MSA • OSCE	BDS2 examinations: • MCQ • MSA • OSCE • Resit diet results	BDS3 examinations: • Anatomy examination • MCQ • MSA • OSCE • Resit diet results	BDS4 examinations: • MSA • Case presentation examination	BDS5 examination: • OSCE	VDT assessment: LEP
					LIFTUPP©			

NOTE: Resist examination diets were not required in BDS1 and BDS4 for cohort 3, and in BDS5 for both cohorts 1 and 3.
MCQ = Multiple-choice question MSA = Multiple-short answer OSCE = Objective structured clinical examination
LEP = Longitudinal evaluation of performance

3.5.3.3 Data linkage and pseudonymisation

Raw LIFTUPP©, BDS1-5 (undergraduate) examination results and LEP data were transferred to a third-party analyst based at the University of Glasgow via a secure *nhs.net* email account by data management staff at LIFTUPP© Limited, Glasgow BDS year group administrators and NES administrative staff (respectively) using password protection. All data sets were sent as Excel spreadsheets and subsequently uploaded into SAS® statistical software (SAS Institute. 2008. SAS® software: Release 9.2. Cary, NC: SAS Institute Inc.).

Undergraduate examination data for each student in each cohort across all five BDS years were linked using student matriculation numbers. Errors in the matriculation numbers were identified and fixed during this process. Only three errors were found and appeared to have been caused by typos. Correcting matriculation number errors was achieved by the third-party analyst cross checking participant demographic data (forename, surname, sex, and date of birth).

For BDS1 data (in each cohort), a dummy identification (ID) called "STUDENT" was calculated simply as "1" to "n" (where "n" = the total number of BDS1 students in the cohort). Another ID was calculated from the students' names in the first of the five years of study. If students were not part of a cohort in BDS1 (e.g., the joined the cohort in a subsequent year), their dummy IDs were hard coded into the merged file. LIFTUPP© data were then linked by matriculation number, and LEP data were linked by the name-based ID.

In total, 20, 20 and 17 Excel spreadsheets were linked for cohorts 1, 2 and 3, respectively. The difference in the number of files merged between cohorts was explained by the need for resit examination diets in some academic years.

Participant forename, surname and matriculation number data were subsequently removed from all linked files by the third-party data analyst to provide pseudonymisation. Linked files were then saved as Excel spreadsheets. Sex and date of birth variables were retained and used by the PhD researcher to check for data linkage errors (see section 3.5.3.4).

Linked (and pseudonymised) data sets were securely transferred and stored on a secure networked folder at the University of Glasgow which was accessible only to the researcher and one member of the supervisory team. Copies of these files were made and transferred to a separate secure networked folder to provide additional security through creation of master files. Subsequent data processing (see section 3.5.3.4) and analyses (see section 3.5.3.5) were conducted using the copied files to preserve masters of each cohort's linked pseudonymised data.

Figure 3.3 presents a diagrammatic summary of the data linkage process.

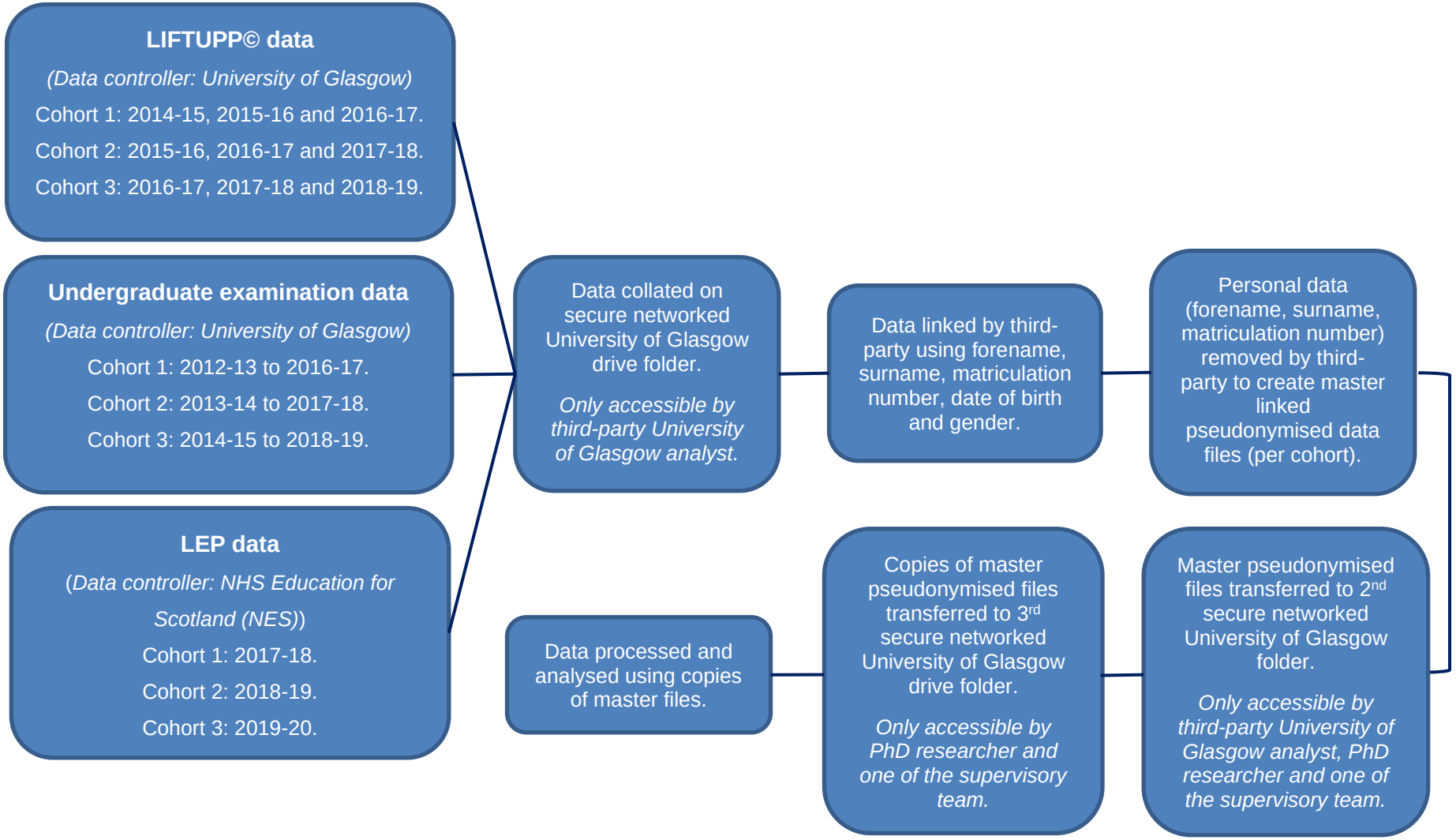


Figure 3.3 - LIFTUPP©, undergraduate examination and Longitudinal Evaluation of Performance (LEP) data linkage process.

3.5.3.4 Quality assurance, data cleaning, and data management

Linked LIFTUPP©, undergraduate examination and LEP assessment data were imported into Stata® 15 statistical software (StataCorp. 2017. Stata Statistical Software: Release 15. College Station, TX: StataCorp LLC).

Checking data linkage

For each cohort, participant ID numbers, sex and date of birth variables were compared across LIFTUPP©, undergraduate examination and LEP data sets to check for linkage errors. Successful data linkage was indicated if the data combination for all three variables was consistent across each of the compared data sets. No data linkage errors were found (see Table 3.4). Dates of birth and sex were then removed from all data sets to fully anonymise the data.

Table 3.4 - Data linkage check findings. ID = Identification. LEP = Longitudinal evaluations of performance.

	Data set comparisons					
	Undergraduate exams *vs* LIFTUPP©		**LIFTUPP© *vs* LEP**		**Undergraduate exams *vs* LEP**	
Cohort	Participant ID number, gender, and date of birth **MATCH** [n (%)]	Participant ID number, gender, and date of birth **DO NOT MATCH** [n (%)]	Participant ID number, gender, and date of birth **MATCH** [n (%)]	Participant ID number, gender, and date of birth **DO NOT MATCH** [n (%)]	Participant ID number, gender, and date of birth **MATCH** [n (%)]	Participant ID number, gender, and date of birth **DO NOT MATCH** [n (%)]
1	91 (100)	0 (0)	80 (100)	0 (0)	91 (100)	0 (0)
2	104 (100)	0 (0)	86 (100)	0 (0)	104 (100)	0 (0)
3	75 (100)	0 (0)	68 (100)	0 (0)	75 (100)	0 (0)

Accounting for and management of missing data

Missing LIFTUPP©, undergraduate examination and LEP data were observed in all three cohorts. Potential reasons for these missing data were investigated by tracking student progression through the BDS curriculum using the examination outcomes data available for each student.

These investigations revealed the missing data for 7/91, 5/104 and 2/75 students in cohorts 1, 2 and 3 (respectively) were explained by unsuccessful outcomes obtained during initial and resit examination diets. For example, students who received failing grades (E, F, G or H) (University of Glasgow, 2019) for both the initial and resit BDS3 examinations had no subsequent BDS4 and BDS5 data as they were either excluded from the course or were required to repeat the academic year.

Since repeating students "dropped" into a different cohort, their subsequent examination results were recorded within the cohort they had joined instead of their original cohort. This also explained missing LIFTUPP© and LEP data sets for these students since these were collated under the years in which students graduated and completed VDT, respectively. For example, if a student who was initially in cohort 1 was required to repeat a BDS year, their LIFTUPP© and LEP data would instead be found in the cohort 2 data sets. If they were required to repeat a further BDS year, their LIFTUPP© and LEP data would have been found in cohort 3.

Missing examination data for 5/91, 8/104 and 6/75 students in cohorts 1, 2 and 3 (respectively) could not be explained through the available examination results, however, they could still be accounted for. These students either i) left a cohort for reasons other than unsuccessful examination outcomes; or ii) had joined a "later" cohort. In the case of the former, students were missing progressive BDS examination results within cohort data sets, whereas the latter were missing earlier results.

Potential reasons for students dropping out a cohort (other than unsuccessful examination outcomes) include:

a) Temporarily left BDS course to pursue an intercalated degree.

b) Suspension of studies.

c) No longer wished to pursue BDS degree.

d) Required to repeat academic year for reasons other than examination results (e.g., unsatisfactory attendance).

e) Exemption from sitting an examination due to special circumstances.

f) Expulsion from BDS course (e.g., due to gross misconduct).

Potential reasons for students joining a cohort included:

a) Repeating an academic year.

b) Returning from intercalated degree.

c) Returning from suspension of studies.

Missing LEP data could be explained by unsuccessful BDS5 examination results for two students. The remaining missing LEP data were assumed to be due to graduating students not enrolling in a Scottish VDT scheme as it is not mandatory for students who graduate from Scottish dental schools to complete VDT in Scotland. Some may choose to enrol in other UK countries' VDT schemes or practise outside the UK. Others may choose not to enrol in VDT at all (see chapter 1, section 1.1.2).

Identification of key variables available with LIFTUPP©, undergraduate examination and LEP data sets

Of the 38 variables available for LIFTUPP© data, 19 were identified as "key" for data cleaning, creation of additional variables or the subsequent statistical analyses. A complete list of all 38 LIFTUPP© variables is provided in appendix 1.

Variables presenting the undergraduate examination percentage scores and grades were also considered as key. Percentage scores and grades were recorded as numerical and categorical data, respectively. A complete list of all variables in the undergraduate examination data sets is also provided in appendix 1.

For the LEP data sets, 16 out of the 18 available variables were initially regarded as key. However, this number was subsequently reduced to nine after it became

apparent that communication, management and leadership, and professionalism data could not be compared with equivalent LIFTUPP© data (see LEP data cleaning section below). Appendix 1 also provides and a full list of all LEP variables.

Key variables within the LIFTUPP©, undergraduate and LEP data sets are highlighted within appendix 1.

Data cleaning and deriving variables

In general, frequency tabulations and histograms were produced for all key LIFTUPP©, undergraduate examination and LEP data variables to check ranges and observe any unusual or typographical errors. Cross tabulations and scatter plots were used to detect logic errors.

i) **LIFTUPP©**

Data cleaning

LIFTUPP© data were cleaned by removing any assisted/observed/simulated activity (leaving only procedures that students had undertaken on patients) and entries with missing key variable data (e.g., performance indicators, anonymous patient ID numbers, and clinic dates). Data for students who repeated or re-joined any BDS year were also removed as they were not comparable with the rest of their peers' data since the timeframe on which they progressed through the BDS course differed. Data for procedures which had been assessed for less than ten students in at least one cohort were also removed. These data were identified through cross-tabulations between all procedures and student ID numbers.

As described in section 3.5.3.2, assessors can assign LIFTUPP© performance indicators across four domains: clinical, communication, management and leadership, and professionalism. Data points from the latter three domains were excluded from this study since, between 2014 and 2016, LIFTUPP© did not attribute the assessment of these skills to single patient encounters and were assigned per clinical session instead. Some clinical domain data points had also

been recorded in this manner - two examples of which were the assessment of extra- and intra-oral examination skills and administration of local anaesthetic. These data points were, for the same reason, also excluded.

Therefore, the remaining clinical data points referred to assessment of any hands-on dental procedure recorded as part of a single patient encounter across various disciplines - namely Restorative Dentistry, Paediatric Dentistry, Oral Surgery and Radiology. The list of procedures which were collectively assessed across these disciplines included biopsies, direct restorations, endodontics, extractions, fissure sealants, indirect restorations, minor oral surgery, treatment of temporomandibular disorder (e.g., occlusal splints), periodontal therapy, preventive therapy (e.g., oral hygiene instruction and fluoride applications), removable prosthodontics, radiographs, and suturing.

Further cleaning was undertaken to remove “ineligible” clinical procedures. Procedures were classified as ineligible if an individual student had not performed all key stages of a treatment item on an individual patient. A list of the key stages for each type of procedure is provided in appendix 2. Ineligible procedures were identified through creation of a binary variable, i.e., observations associated with eligible and ineligible procedures were marked with “1s” and “0s”, respectively. Observations marked with a “0” were subsequently removed.

A summary of the LIFTUPP© data cleaning process is shown in Figure 3.4.

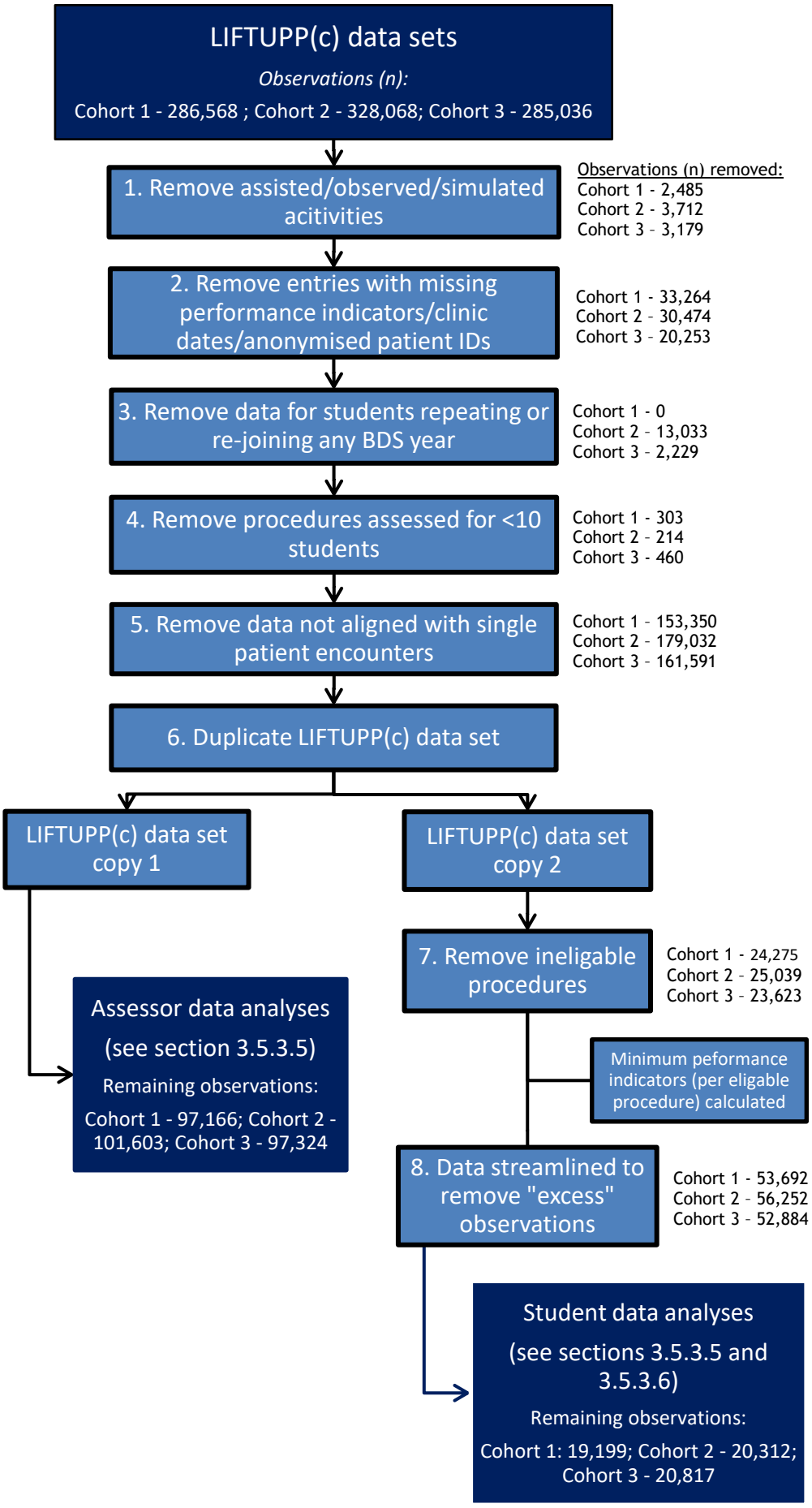


Figure 3.4 – Summary flow chart of LIFTUPP© data cleaning process.

Deriving a minimum performance indicator and thresholds of clinical performance

The minimum performance indicator assigned per eligible clinical procedure was used to summarise each student's performance for a clinical procedure at a certain point in time since it was assumed that a student's overall performance would only be as good as their lowest performance indicator for a single procedural stage. For example, if a student who completed a direct restoration was awarded a performance indicator of 3 for the caries (decay) removal stage but was given 4's for all other stages, their minimum performance indicator was recorded as 3. Observations which were not aligned with the minimum performance indicator awarded for each procedure were removed to streamline the data sets. This process reflects that of Roudsari (2017), who has also investigated LIFTUPP© data patterns through statistical modelling (see chapter 9, section 9.3.7).

Additionally, for each clinical procedure a student undertook, a binary score of "1" was assigned if the minimum performance indicator was above a threshold level and a "zero" was assigned if it was below. Anecdotally, there currently appears to be inconsistency between assessors on whether a LIFTUPP© performance indicator of 4 or 5 constitutes competent clinical performance. For the purposes of this study, performance indicator 4 was used as the baseline threshold for satisfactory clinical performance and to investigate content validity. A threshold performance indicator of 5 was subsequently used to investigate criterion validity, as it allowed more than one clinical performance trajectory per student cohort to be identified (see chapter 5, sections 5.4.2.2 and 5.4.2.3).

Finally, each student's remaining LIFTUPP© observations were labelled "1" to "n", where "n" = the number of remaining LIFTUPP© observations aligned with an individual student. Therefore, the total number of eligible clinical procedures completed by a student was the same as "n".

A list of the additional variables created within the LIFTUPP© datasets is provided alongside the original data set variables in appendix 1.

In summary, all the above steps ensured each remaining LIFTUPP© observation contained serial data on minimum clinical performance indicators attached to a date for all eligible clinical procedures over BDS3-5 and indicated whether the minimum performance indicators had met thresholds of 4 or 5.

ii) **Undergraduate examinations data**

Data cleaning

Data for students who repeated or re-joined any BDS year had already been removed as part of the LIFTUPP© data cleaning process.

Deriving mean aggregated examination performance scores

Average examination performance for individual students was calculated as for each BDS year as a percentage and labelled as the "mean aggregated examination performance". Examination performance (per BDS year) were then categorised into thirds, quarters, and fifths. Binary variables were derived for students scoring in the top third (33%), quarter (25%) and fifth (20%) in each BDS year.

Only assessments that were numerically scored could be included. Therefore, the BDS1 summative essay results were omitted as they did not use numerical scoring. Instead, students had been subjectively awarded alphabetic grades by assessors (based on the University of Glasgow's Schedule B grading scheme (University of Glasgow, 2020)), which were categorical data. This was consistent across all three cohorts.

A similar issue was observed for the BDS4 case presentation results in cohort 1 where the students were only awarded alphabetic grades (based on the University of Glasgow's Schedule B grading scheme (University of Glasgow, 2020)) had been awarded. Numerical scoring was introduced to the BDS4 case presentation examination in 2016-17, meaning the results of this assessment could have contributed to the calculation of BDS4 mean aggregated examination performance in cohorts 2 and 3. However, to ensure consistent statistical analysis across all three cohorts, BDS4 case presentation results were omitted.

iii) LEP data

Data cleaning

Data for participants who repeated or re-joined any BDS year had already been removed as part of the LIFTUPP© data cleaning process.

Variables related to LEP scores could be categorised according to the GDC's four domains of clinical practice (clinical, communication, management and leadership, and professionalism) (GDC, 2015a). Table 3.5 provides a summary of the categorisation.

Table 3.5 - Categorisation of Longitudinal Evaluation of Performance (LEP) data set variables according to the General Dental Council's (GDC's) domains of competent clinical practice.

LEP score variable	GDC domain
Examination and consultation skills	Clinical
Clinical judgement and diagnosis	Clinical
Technical ability and manual dexterity	Clinical
Communication skills	Communication
Professionalism	Professionalism
Knowledge level and application	No equivalent
Organisation	Management and leadership
Trainee's insight into performance	Professionalism

Data categorised under communication, management and leadership, and professionalism were removed since they could not be compared with their equivalents in LIFTUPP©. This data within LIFTUPP© had previously been removed since they were not aligned to single student-patient encounters until Glasgow Dental School's 2016-17 academic year (see above). Knowledge level and application data were also removed.

Although the "examination and consultation skills" and "clinical judgement and diagnosis" variables could be categorised within the "clinical" domain, it was not possible to consistently compare the assessment of these skills with counterpart data recorded by LIFTUPP©. This was because (like communication, management and leadership, and professionalism entries within LIFTUPP©) assessment data for these skills had also not been aligned to single student-

patient encounters within LIFTUPP© until 2016-17. Consequently, “examination and consultation skills” and “clinical judgement and diagnosis” LEP data were excluded from the study. This left “technical ability and manual dexterity” scores for the analysis of the clinical domain LEP data.

In addition, LEP data tabulations revealed there were observations in which scores for both “examination and consultation skills” and “technical ability and manual dexterity” had been awarded despite indicating that only examination skills had been assessed. A mock example of a such a data entry is provided in Table 3.6.

To ensure such entries were not included, all “examination” observations were removed along with those with missing clinical dates and missing “technical ability and manual dexterity” scores.

Figure 3.5 summarises the LEP data cleaning process.

Table 3.6 – Mock example of an examination skills assessment where a score for "technical ability and manual dexterity" has been provided. ID = identification. LEP = Longitudinal evaluation of performance.

Participant ID number	Date	LEP title	Examination and consultation skills	Clinical judgement and diagnosis	Technical ability and manual dexterity	Communication skills	Professionalism	Knowledge level and application	Organisation	Trainee's insight into performance
2	22/8/17	Examination	6	6	5	5	6	6	6	6

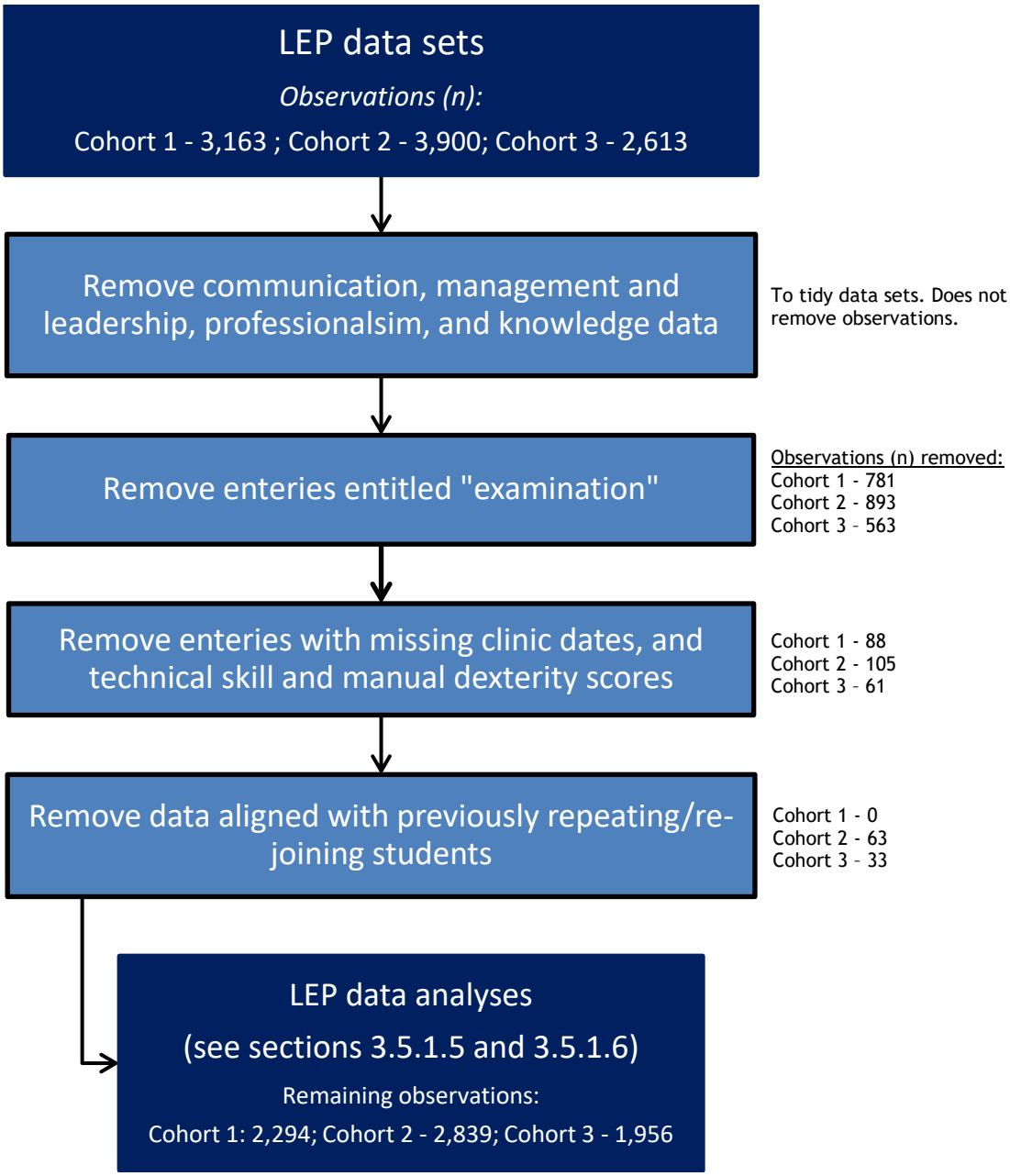


Figure 3.5 – Summary flow chart of longitudinal evaluation of performance (LEP) data cleaning process.

Deriving thresholds of clinical performance

For each LEP assessment a VDP undertook, a binary score of "1" was assigned if the technical skill and manual dexterity score awarded was above a threshold level and a "zero" was assigned if it was below. The guidance for LEP scoring provided by NES (see section 3.5.3.2) appears to suggest a score of 4 would serve as the minimum indicator of competent performance (as any score <4 is classed as "Needs Improvement").

For the purposes of this study, four different threshold scores (4, 5, 6 and 7) were investigated for LEP data and, therefore, four additional variables were

generated with the LEP data sets: one to indicated “success” where students obtained a score ≥4, another to indicate “success” when a score ≥5 was achieved etc.

Finally, each VDP’s remaining LEP observations were labelled “1” to “n”, where “n” = the number of remaining LEP observations aligned with an individual VDP. Therefore, as for the student LIFTUPP© data, the total number of eligible LEPs completed by a VDP was the same as “n”.

3.5.3.5 Statistical analysis

Analysis strategy: Level of analysis

As previously described in section 3.5.3.2, data entries within LIFTUPP© are automatically categorised according to the GDC’s (2015a) four key domains of competent clinical practice (i.e., clinical, communication, management and leadership, and professionalism) and clinical domain data are subcategorised further according to dental subject (e.g., Oral Surgery) and then clinical procedure (e.g., tooth extraction) (see Figure 3.2 in section 3.5.3.2).

Conducting statistical analyses for each domain of clinical competence and at each level of categorisation was initially considered. However, during data cleaning (see above), it became apparent that LIFTUPP© communication, management and leadership, and professionalism data had not been aligned with individual student-patient encounters in both cohorts 1 and 2. This was a manifestation of how LIFTUPP© data entries for these domains were made during the early years of the system’s implementation. Performance indicators for communication, management and leadership, and professionalism were previously awarded per clinical session, in which students may have treated more than one patient. In contrast, LEP assessment data for all four domains were linked with individual patient encounters.

The LIFTUPP© system was updated in 2016-17 to permit linkage of communication, management and leadership, and professionalism data with individual patient encounters. Therefore, these data were aligned to single student-patient encounters in cohort 3. However, it was decided that only

clinical domain LIFTUPP© and LEP data were to be analysed and compared to maintain consistency across all three cohorts.

It also became apparent during the data cleaning phase that consistent comparisons between LIFTUPP© and LEP data sets could not be conducted at the subject or procedural levels. This issue stemmed from VDT trainers being able to describe the clinical procedures being assessed within LEPs in their own words (free text). In many cases, this made it difficult to determine what procedure or dental subject was being assessed. For example, there were entries described as "emergency treatment" or "emergency appointment" – terms which could constitute a variety of dental problems and subsequent treatment modalities.

Furthermore, some LEP entries recorded assessments of multiple procedures with a single LEP, e.g., a restoration (Restorative Dentistry) and an extraction (Oral Surgery). In such cases, it was not possible to determine precisely how scores for clinical performance had been awarded (e.g., an average score based on the performance of all procedures? The lowest score awarded across all procedures?). However, it was still clear that "clinical" performance had been assessed.

Based on these observations, it was decided to conduct data analysis at the domain level and for only clinical domain entries.

Analysis strategy: LIFTUPP© data - assessor and student perspectives

LIFTUPP© data could be investigated from two perspectives: assessors and students. The former would involve exploring the number of procedural stage assessments conducted per assessor, the number of individual students assessed per assessor, and the frequencies of LIFTUPP© performance indicators awarded by assessors within and across BDS academic years. The latter would involve exploring the number of "eligible" clinical procedures completed per student, the frequencies of minimum LIFTUPP© performance indicators awarded to students for each eligible procedure and generating trajectory models tracking students' clinical performance over time (see descriptive statistical analysis and group-based trajectory modelling sections below).

It was decided that analysis from the perspective of assessors was not to be conducted on LIFTUPP© data sets in which ineligible procedures had been removed as it may have resulted in significant over- or underestimation of assessor assessment activity. This was because the same assessor did not always assess each stage of a clinical procedure performed by students. Therefore, since assessors would not always assess procedures in their entirety, each procedural stage for which an assessor had provided scores was regarded as a "procedural stage" assessment (when analysing from the assessor perspective). Removing all data aligned with ineligible procedures reduces the number of procedural stages within the data sets (see chapter 5, section 5.4.1.1) and, therefore, would have severely reduced the number of procedural stage assessments performed by assessors.

For this reason, LIFTUPP© data sets were analysed from assessor and student perspectives independently.

Descriptive statistical analysis

All continuous variables within the linked data sets were plotted using histograms to visually assess distributions in order that appropriate summary statistics and analyses were adopted. Summary statistics (mean, standard deviation (SD), minimum, median, maximum, Q1 and Q3 statistics) were used to describe:

- the number of eligible LIFTUPP© clinical assessments completed per student per cohort (within and across BDS years).

- the number of clinical procedural stage assessments completed per assessor (within and across BDS academic years).

- the number of individual students assessed per assessor (within and across BDS academic years).

- all numerical undergraduate examination data.

- the number of eligible LEP clinical assessments completed per VDP.

The number of eligible LIFTUPP© clinical assessments completed (per student per cohort), all numerical undergraduate examination data and the number of eligible LEP clinical assessments completed (per VDP) were also summarised graphically using boxplots.

Frequency tables were produced for all categorical and discrete data, which included:

- the minimum LIFTUPP© performance indicators awarded for eligible clinical procedures (within and across BDS years).
- LIFTUPP© performance indicators awarded per procedural stage assessment by assessors (within and across BDS years).
- grades awarded for each undergraduate examination.
- LEP scores awarded for eligible clinical assessments (within and across LEP blocks).

Bar charts were also produced to graphically illustrate the distribution of minimum LIFTUPP© performance indicators awarded for eligible clinical procedures (within and across BDS years) and LEP scores awarded for eligible clinical assessments.

Additional statistical analyses for undergraduate examination data

Undergraduate assessment data were considered in several ways: as continuous data; and as categorical data split into equal thirds, quarters, and fifths.

Further details on additional statistical tests undertaken for undergraduate examination data are provided in chapter 4 (section 4.3).

3.5.3.6 Group-based trajectory modelling

GBTM was used to detect latent trajectory groups for LIFTUPP© (minimum score) and LEP (score) data to provide a means of summarising longitudinal clinical

assessment data for each student/VDP. LIFTUPP© data modelling was only conducted from the student perspective.

Models were generated using the Stata® statistical software plugin "*traj*" (Jones and Nagin, 2012; Jones and Nagin, 2013), which calculates: a) the predicted trajectory for each group identified in the data set; and b) the probability of trajectory group membership (for each trajectory) per participant. Individual participants are allocated to the group for which they had the highest probability of membership (i.e., a probability >0.5).

Details for GBTM generation, evaluation and selection are provided in the following sections.

Model generation

i) Measurement of time/independent variable

LIFTUPP© and LEP data trajectories were investigated using calendar dates as the measurement of time, which served as independent variable for the GBTMs.

ii) Data distribution/dependent variable

LIFTUPP© and LEP data were investigated using two data distribution models supported by the traj plugin: *censored normal* and *Bernoulli* (Jones and Nagin, 2012). The former refers to data that typically follow a normal data distribution pattern but are confined by a minimum and/or maximum value (Tobin, 1958). The latter is a discrete distribution of the probability of two possible outcomes occurring ("success" and "failure") (Evans, Hastings and Peacock, 2000), for which a threshold for "success" is to be defined.

As there is no clear cut-off performance indicator for LIFTUPP© in practice, it was decided to consider Bernoulli models for LIFTUPP© data based on the probability of participants obtaining threshold performance indicators of 4 (for content validity) and 5 (for criterion validity). Similarly, for LEP, where clear cut-off performance indicators do not exist in practice, Bernoulli models were

based on the probability of participants obtaining threshold LEP scores of 4, 5, 6 and 7.

Table 3.7 provides a summary of the variations of GBTMs analysed according to the selected data distribution and (where appropriate) threshold scores.

Table 3.7 – Variations of group-based trajectory models (GBTMs) for LIFTUPP© and longitudinal evaluation of performance (LEP) data according to distribution of data, and (for Bernoulli data distributions) the threshold scores for competent performance.

Data set	GBTM variation	Data distribution	Threshold score set
LIFTUPP©	1	Censored normal	NA
	2	Bernoulli	4
	3	Bernoulli	5
LEP	1	Censored normal	NA
	2	Bernoulli	4
	3	Bernoulli	5
	4	Bernoulli	6
	5	Bernoulli	7

iii) Number and shape of trajectory groups

For each variation of GBTM (Table 3.7), the existence of single, two-, three- and four-group models were investigated. All 340 potential combinations of trajectory shape(s) (zero-order (horizontal line), linear, quadratic (one turning point) and cubic (two turning points)) within each model were explored (see appendix 3).

The traj plugin uses numerical codes to depict trajectory shapes. These codes are as follows:

0 = zero order (straight line)

1 = linear

2 = quadratic (one turning point)

3 = cubic (two turning points)

Model evaluation and selection

i) Initial ranking: Bayesian information criterion

In addition to the predicted trajectory for each group and the probability of trajectory group membership, the statistical output produced by the traj plugin also includes two Bayesian information criterions (BICs): one calculated from the number of participants; and another based on the number of observations analysed. The BIC is closely related to the Akaike information criterion (AIC), which is also generated by the traj plugin. Both the BIC and AIC are indices that can be calculated to help choose between two or more alternative statistical models.

The BIC can be defined mathematically as:

$$BIC = -2 \log L + K \log n$$

L = likelihood. K = number of model parameters. n = number of data points.

The formula for the AIC is:

$$AIC = -2 \log L + 2K$$

L = likelihood. K = number of model parameters.

When selecting the model which fits the data best, both the BIC and AIC can be used. However, the BIC is generally stricter in penalising models with intricate parametrisation (i.e., models based on many rules and conditions) compared to the AIC and therefore tends to favour more parsimonious models (Schwarz, 1978; Kass and Raftery, 1995; Neath and Cavanaugh, 2012; Shearer, 2016). The BIC always has a negative value and models which fit the data best typically produce the *least* negative BIC (i.e., the value closest to zero).

Within this study, model BIC's were imported into Excel and ranked from highest (least negative) to lowest (most negative) BIC based on the number of

participants. In accordance with recommendations made by Wagenmakers (2007) and Hox (2010), the BIC based on the number of participants was favoured for ranking and choosing the most suitable models, instead of the BIC based on the number of observations.

ii) Models with statistical errors

At this stage, models which had returned statistical errors were discarded. Statistical errors which could be identified by the traj plugin may include:

1. the standard error for the co-efficient not being calculated due a non-symmetric or highly singular variance matrix.

2. False convergence (which indicates failure to estimate model parameters).

iii) Trajectory group membership

At the time of study, there were no specific guidelines on the minimum percentage (or number) of individuals required in the smallest trajectory group for different sample sizes. Klijn et al. (2017) proposed the percentage of individuals estimated to be assigned to the smallest group should serve as a criterion for model selection. They stated a cut-off point of 1% is often applied but suggested that this threshold may require adjustment depending on study sample size, with smaller samples requiring a larger minimum percentage to ensure there are enough members in the smallest trajectory group.

This study required a larger minimum group number restriction since dental student cohorts typically never exceed 100 and, therefore, a cut of 1% would not equate to a single student. Furthermore, since part of this study was designed to compare longitudinal clinical assessment trajectories with mean undergraduate examination scores, trajectory groups consisting of a single student would be unsatisfactory. As a result, a minimum sample size restriction of at least n = 5 per trajectory group was set. The impact of minimum group number restrictions of 10, 15 and 20 on the data patterns was also investigated.

iv) Testing model adequacy

For each cohort and data distribution, the five GBTMs with BICs closest to zero for each of the minimum group membership restrictions were selected and subjected to further statistical scrutiny to determine their suitability. Nagin (2005) and Nagin and Odgers (2010) advocate that, for models to be statistically adequate, they had to satisfy each of the following criteria:

1) the average posterior probability (AvePP) value is >0.70 for each group.

2) the odds of correct classification based on probabilities of group membership is >5.0 for each group.

3) there is close correspondence between each trajectory group's estimated probability of group membership and the proportion of students classified to that group according to posterior probability of group membership.

4) tight confidence intervals are observed around estimated group probabilities.

GBTMs which did not satisfy these criteria were discarded (see results in chapters 5 and 6).

v) Final model selection(s)

The BICs of the remaining GBTMs were compared. In accordance with Raftery's (1995) approach, if the difference in BIC between models which were compared was greater than 2, then the model with the higher (i.e., less negative) BIC was considered as better fitting, and therefore selected as a representative of each cohort's data (according to the data distribution selected, threshold score and minimum group number restrictions being investigated). More specifically, a BIC difference between 2 and 6 indicates there is "positive" evidence for selecting models with a higher BIC as the best fitting. A difference between 6 and 10 suggests "strong" evidence, and a difference greater than 10 is "very strong".

However, if the difference between BICs was less than 2, Raftery (1995) suggests there is little evidence to support that one model is better than the other. In these instances, the most parsimonious models (i.e., those which were simple but provided a clear explanation of the data) were selected upon reviewing the graphical representation of each model's trajectory groups (see chapters 5 (LIFTUPP©) and 6 (LEP) (e.g., a linear model would be chosen over a quadratic; and a quadratic over a cubic). This approach followed Nagin and Odgers (2010) proposal of selecting models which address the research question(s).

Figure 3.6 summarises the GBTM creation, evaluation and selection processes described above.

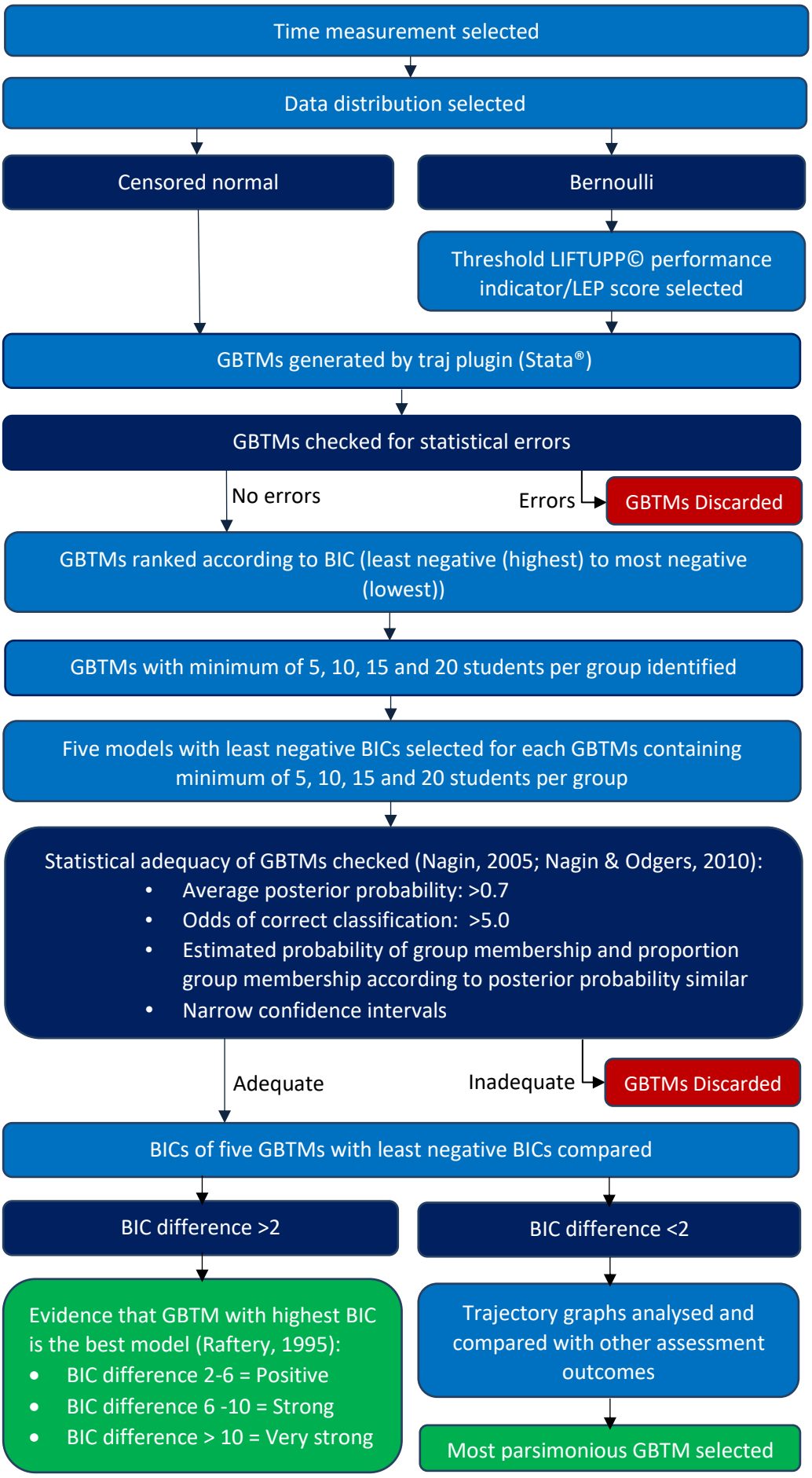


Figure 3.6 – Stages of GBTM generation, evaluation, and selection.

Descriptions of the GBTMs selected at the end of these processes are provided in the results chapters 5 (LIFTUPP© data) and 6 (LEP data). The models selected were used to:

1) Establish patterns of undergraduate clinical performance recorded by LIFTUPP© over time (content validity – research question 2a).

2) Compare LIFTUPP© performance with undergraduate examination outcomes (criterion (concurrent) validity – research question 2b).

3) Compare LIFTUPP© performance with LEP performance (criterion (predictive) validity – research question 2c).

Specific details on the comparisons between assessment data sets are provided in chapter 7 (LIFTUPP© vs undergraduate examinations and LEPs).

3.5.3.7 Reliability

Cronbach's alpha was calculated for the panel of undergraduate assessments between BDS1-5. For LIFTUPP© assessment, the probability of being in the "best" performing group was used for each student. Cronbach's alpha was then recalculated following removal of each assessment item individually and assessments were considered to be reliable if Cronbach's alpha >0.7 (Cronbach, 1951).

Details on selection of a "best" performing trajectory group are provided in chapter 5 (section 5.4.2).

3.5.4 Method for addressing research question 3

3.5.4.1 Design and setting

Two focus groups were conducted to obtain data from key stakeholders within dental education. The data to be collected were the key stakeholders' thoughts and opinions on how assessment within dental education could be enhanced based on results obtained from the quantitative analysis described in section

3.5.3. These data were to be used to address research question 3 (see chapter 2, section 2.2).

Originally, the focus groups were going to be conducted in person. However, due to the Coronavirus disease 2019 (COVID-19) pandemic and to comply with social distancing measures imposed by the Scottish and UK governments during 2020-21, it was not possible to conduct face-to-face focus groups at the time of study. As a contingency and following ethical approval (see section 3.6.1), focus group meetings were conducted online using Microsoft Teams video conferencing technology (version 1.4.00.2781. Microsoft Office 365. Redmond, WA: Microsoft Corporation).

3.5.4.2 Sampling

Key stakeholders who were originally invited to participate in the focus groups included undergraduate dental students (enrolled at the University of Glasgow at the time of study), VDPs (undertaking VDT at the time of study), recently qualified (2017-19) dentists and Scottish VDT trainers. Staff based at the University of Glasgow Dental School who were invited to participate included clinical teaching staff, BDS course co-ordinators, Teaching Leads for specific dental subjects (e.g., Restorative Dentistry, Paediatric Dentistry, Oral Surgery and Oral Medicine), the Clinical Teaching Lead, the Director of Dental Education, and the Head of School. Equivalent assessors from other UK and Irish dental institutions were also invited to participate, as were representatives from NES (the Postgraduate Dental Dean and the Associate Postgraduate Dental Deans).

Further details on those who were recruited (including the resultant number of participants (n) for each focus group) are provided in the chapter 8 (section 8.4.1).

3.5.4.3 Recruitment and consent

Key stakeholders were invited to participate via email. The invitation contained a participant information leaflet detailing the purpose of the study, its methods and how their data would be used, anonymised, and stored (appendix 4). Those who wished to participate were asked to reply to the invitation email. They

were then provided with a privacy notice and written consent form (appendix 4). Those who signed and completed the consent forms were subsequently enlisted in the study.

Recruited participants were split into two groups. One group consisted of undergraduate dental students, VDPs, and recently qualified dentists. These individuals were labelled as "student" focus group participants. The other focus group consisted of staff from both the University of Glasgow Dental School and other UK and Irish dental school's - and a representative from NES. Collectively, dental school staff and the NES representative were labelled as "faculty".

3.5.4.4 Data generation and anonymisation

During each focus group, participants were shown a series of short presentations on key findings produced by the quantitative analyses described in section 3.5.3. After each short presentation, participants were asked to discuss their thoughts and opinions on the quantitative data and how it could be used to enhance assessment within dental education.

Since there were no specific pre-existing parameters or guides on how to deduce meaning from LIFTUPP©, undergraduate examinations and LEP data, key stakeholders had to present their personal views on the situation being studied. This fits with Creswell and Creswell's (2018) description of the interpretivist approach (see section 3.3), where the study participants' views must be relied upon as much as possible. The views presented will have been shaped by the participant's previous experiences, and meaning is constructed through their discussions and interactions with others.

Focus group discussions were moderated by the PhD researcher. A "medium" level (Kitzinger, 1994; Bryman, 2016) of moderation was adopted to ensure the discussions generated data for answering the qualitative research question and the opinions and perceptions of key stakeholders were not influenced by the views and interpretations of the PhD researcher. To help with moderation of the focus groups, a topic guide (appendix 5) was developed by the PhD researcher using guidelines proposed by Krueger (1998). The topic guide consisted of a series of questions which were categorised as follows:

- Opening question – allows participants to introduce themselves with one another and settle into the focus group.

- Introduction question – encourages participants to start thinking about the topic at hand and focuses the conversation.

- Transition questions – provide links between the quantitative results presentations and the key questions.

- Refocus questions – prevent discussions from straying too far off-topic.

- Key questions – generate discussions on areas of greatest concern (i.e., those required to answer the qualitative research question).

- End questions – conclude sessions and allow participants to make further comments on topics they felt were not covered.

The format of the topic guide also helped contribute to data quality and minimisation by ensuring conversations were predominately focussed on the research topic and, therefore, data necessary for answering the qualitative research question were obtained. If no topic guide was used, there would be a potential risk of gathering a lot of data that may not contribute to answering the research question. However, the "medium" level of moderation also permitted some flexibility for additional lines of conversation – particularly if participants raised a relevant topic which had been potentially overlooked by the PhD researcher/moderator.

Audio and video recordings of the online focus group discussions were produced using the "record" feature available within Microsoft Teams video conferencing software. These recordings were automatically uploaded to Microsoft Stream (Microsoft Office 365. Redmond, WA: Microsoft Corporation. [Accessed December 2020 and February 2021]) to facilitate transcription of the audio recordings, which were produced using the "transcribe" feature available within the platform. An additional audio recording was made using a digital voice recorder placed next to the speakers of the PhD researcher's laptop computer.

This additional recording served as a backup should any technical difficulties have occurred with Microsoft Teams.

Text from the automatically generated transcripts was imported into Microsoft Word software (Microsoft Office: 2016. Redmond, WA: Microsoft Corporation). The accuracy of the automatically generated transcripts was checked by the PhD researcher, who listened back to the audio/video recordings of the discussions whilst reading through the texts produced by Microsoft Stream. Errors produced by the automated transcription were corrected by the PhD researcher where necessary, however the transcripts were kept fully verbatim, meaning word-for-word records of what was originally spoken were produced and retained filler words such as “umm’s” and “err’s”, poor grammar, false starts, stutters etc.

During the review process, personal identifiable data (such as names, job titles/BDS year of study) were removed and alphanumerical codes were assigned to the participants to initially provide pseudonymisation. Upon completion of the transcript reviews, the original audio/video recordings and automatically generated transcripts were deleted from Microsoft Stream as were backup audio recordings from the digital voice recorder. Deleting the recordings resulted in fully anonymised transcript data sets (in Word documents). Copies of the transcripts were saved to a subfolder on the University of Glasgow’s secure networked J-Drive.

3.5.4.5 Data analysis

Qualitative data generated by focus group discussions were analysed using “thematic analysis”. This approach (described in detail by Seal (2016) and Maguire and Delahunt (2017)) was used to identify themes (i.e., patterns) in the data which were of significant interest and would address the qualitative research question.

In accordance with recommendations by Cohen, Manion and Morrison (2017) and Saldana (2013), coding was used to carefully reduce and refine the volume of data within the focus group transcripts and facilitate identification of categories and key themes. The coding process involved reading (and re-reading) through the transcripts and creating annotations which assigning short names, phrases or

notes to pieces of text. All codes were subsequently listed onto a series of Post-It notes and reviewed. Post-It notes with codes which referred to similar or relatable concepts were arranged into categories which in turn were arranged into overarching themes (see appendix 6). Themes were then reviewed against the transcripts and refined where necessary before they were defined. Finally, summaries of focus group discussions relating to each theme were produced in a Word document.

A short 6-phase overview of this process for thematic analysis has been described by Braun and Clarke (2006):

1. Become familiar with qualitative data.
2. Generate initial codes.
3. Search for themes.
4. Review themes.
5. Define themes.
6. Summarise findings.

Although two separate focus groups were conducted, the transcripts generated by each were analysed together. This approach was adopted to facilitate identification of common categorises and themes across both discussions. The intention of this part of the study was not to compare the opinions of the two groups, but to allow the opinions of all key stakeholders to be used collectively to contribute further data for answering research question 2a (see chapter 2, section 2.2) and to answer research question 3 (see chapter 2, section 2.2). However, it was anticipated that the two groups may present different opinions during their respective discussions, and these are reported in the results (chapter 8).

The rationale for two separate focus group discussions was to prevent potential issues arising from dependent relationships between students and faculty members (see section 3.6.1.1).

3.6 Approvals: Ethics and information governance

3.6.1 Ethical approval

Ethical approval was sought from the University of Glasgow's College of Medical, Veterinary & Life Sciences (MVLS) Ethics Committee for Non-Clinical Research Involving Human Subjects who approved the project in its entirety (reference number: 200170146) (appendix 7).

3.6.1.1 Quantitative study component

The main ethical considerations for the quantitative component of study were i) confidentiality, ii) consent and iii) data security/non-disclosure.

i) Participant confidentially

Participant LIFTUPP©, undergraduate examination and LEP data needed to be linked for comparisons to be made and to allow narratives on the development of competent dental practice to be established. For the datasets to be linked correctly, personal identifying factors (foreman, surname, student matriculation number, sex, and date of birth) needed to be made available. However, to preserve confidentiality, the researcher and project supervisors were not privy to any disclosive participant information during the linkage process. Instead, a third-party University of Glasgow data analyst, who had no links to the study, performed the linkage and provided the study team with a pseudonymised linked data set. Details of this process were described in section 3.5.3.3.

Working with pseudonymised data reduced the possibility of the researchers (i.e., the PhD researcher and project supervisors) - who are also assessors of undergraduate students - being able to identify individuals' LIFTUPP© and undergraduate examination data. LEP data were not identifiable since the researchers had no role in postgraduate VDT assessment. However, there was chance the PhD researcher and one of the supervisory team could recognise

LIFTUPP© and undergraduate examination performance data patterns as their academic responsibilities involved analysing these data sets in a fully disclosive format to contribute to decisions on student progress. The size of the pseudonymised data set for this research project mitigated this problem as it was likely that only data extremes (i.e., exceptionally good, or poor performance compared to peers) could have been attributable to certain individuals. Despite this potential eventuality, student progress outcomes could not be influenced since the assessment data were analysed retrospectively.

ii) Participant consent

Since personal data were required to initially link the multiple data sets without participant consent, the study needed to ensure it was compliant with the European Union's (EU's) General Data Protection Regulations (GDPR) (2016).

The study protocol was subjected to the UK Information Commissioner's Office's (ICO's) "*three-part test*" (ICO, 2018), which determines if there is a lawful basis for processing personal data. The test concluded, in accordance with GDPR, there was a legitimate interest for the study to progress since risk to the participants was minimal and the study would contribute towards improving future assessment within dental education, as the information gathered will help inform dental assessors and regulators on the use of longitudinal data as a method of assessment.

Ultimately, this contributes to public protection by ensuring dental students are being assessed with the best, evidence-based methods available before their entry onto the professional registers.

A copy of the responses to the ICO's three-part test for this study are provided in appendix 7.

iii) Data security/non-disclosure

Assessment data sets were sent to the third-party data analyst's secure *nhs.net* email account. Copies of the raw data sets were saved to a folder on a secure networked drive at the University of Glasgow which was only accessible to the

third-party data analyst. The data sets were then deleted from the third-party analyst's email account.

Identifying data (forename, surname, matriculation number, sex, and date of birth) were used by the third-party analyst to link the assessment data sets. Once data linkage was completed, three of the five personal identifiers (forename, surname, and matriculation number) were removed to produce pseudonymised data sets (see section 3.5.3.3). Sex and date of birth were both initially retained to allow the PhD researcher to check for data linkage errors. Once these checks were completed, date of birth was removed from the data sets (see section 3.5.3.4).

The linked pseudonymised data sets were stored in a folder on a secure networked drive at the University of Glasgow. This folder was only accessible to the third-party analyst, the PhD researcher and one of the supervisory team.

Master files of the linked pseudonymised assessment data were never edited. Instead, copies of the master files were transferred into a subfolder and it was these files versions which were edited as part of the data cleansing and analysis processes (see sections 3.5.3.4 and 3.5.3.5, respectively). This subfolder was only accessible to the PhD researcher and one of the supervisory team.

iv) Data sharing agreement - The University of Glasgow and NES

Ethical approval required a data sharing agreement between the University of Glasgow and NES to permit use of the [quantitative] data sets for research purposes. The agreement outlined the purpose of the study as well as security arrangements for the data, which included how the data sets would be linked (see section 3.5.3.3), stored (see section 3.5.3.3), processed (see section 3.5.3.4) and analysed (see section 3.5.3.5) with relevance to the University of Glasgow's data security protocols. It also provided confirmation that LEP data were to be transferred from NES to the third-party analyst based at the University of Glasgow using secure *nhs.net* email accounts for linkage and pseudonymisation.

Appendix 7 provides a copy of the data sharing agreement.

3.6.1.2 Qualitative study component

The main ethical considerations centred round the qualitative component of study were i) confidentiality, ii) data security and non-disclosive information, iii) privacy and iv) dependent relationships.

i) Confidentiality

Personal data (name, email address and, where appropriate, job title or BDS year of study) were required as part of the focus group recruitment process. Audio and video recordings of human participants are also regarded as personal data under GDPR regulations (European Parliament and Council of European Union, 2016).

Participant focus group responses were anonymised during the transcriptions of the recorded discussions. Alphanumeric codes were used instead of any disclosive information (such as names and job title) (as per section 3.5.4.4). However, whilst the PhD researcher and his supervisory team made every effort to ensure participation was anonymised, due to the nature of focus groups, this assurance could not be made on behalf of the other participants. All participants were informed of this via the study information leaflet provided as part of the recruitment process (appendix 4).

ii) Data security and non-disclosive information

All personal data (surname, forename, email address, job title/BDS year of study and focus group recordings) needed to be processed, stored, and protected appropriately as they are all personal data.

Personal data obtained as part of the recruitment process (i.e., participant names, email address and either job titles or student year of study) were stored on the University of Glasgow's secure networked J-Drive and only accessible to the PhD researcher and the supervisory team.

Audio/video recordings and transcripts from focus groups with key stakeholders were only accessible to the PhD researcher and the supervisory team. Transcriptions contained no identifiable data and instead participants were

assigned an alphanumeric code that was only known to the researchers. Once the transcriptions were completed, checked, and verified, the audio and video recordings of the focus groups were deleted. Transcripts were stored on the University of Glasgow's secure networked J-Drive (only accessible to the PhD researcher and the supervisory team). Deletion of the original audio and video recordings provided greater anonymity to the transcripts (see section 3.5.4.4).

iii) Privacy

Due to the nature of online focus groups, some participants may have joined the discussion from their own homes and therefore their privacy may have been compromised if video communication were used. Other individual's privacy may be compromised due to unintended viewers.

Participants were invited to use the "virtual background" feature within Microsoft Teams, which can increase privacy by blocking out the background of the room from which they are broadcasting.

The focus group moderator (i.e., the PhD researcher) used Microsoft Team's settings to ensure no video footage could be recording until after all participants had entered the online meeting room and chosen to activate their video footage themselves. This also gave participants the option on whether they wish to appear on the video recording of the meeting or not.

iv) Dependent relationships

Undergraduate dental students were invited to participate in the focus groups. Invitations for participation were sent by the PhD researcher, who is a teacher known to the students at the University of Glasgow Dental School. As a result, some students may have felt obliged to volunteer for participation.

Undergraduate students invited for participation were given assurances that:

- Participation was entirely voluntary.

- No prejudice was set between those who chose to participate and those who did not.

- Participants were free to withdraw from the study up to (and including) the date on which focus group conversations were transcribed. After this point, they were no longer be able to withdraw as the transcriptions had been anonymised and it was no longer possible to determine what had been said by individual participants.

- Responses did not impact on academic records and/or student progress.

- Data gathered from the focus groups were only to be used for research purposes.

- Responses were to be anonymised as part of the transcription process (see section 3.5.4.4).

These assurances were reiterated to those who expressed an interest in taking part in the study through one-to-one discussion prior to signing a privacy notice consent form for participation.

Furthermore, to reduce any potential student anxiety caused by dependant relationships, separate student and faculty focus groups were held (i.e., there were no other Glasgow Dental School staff present in the student focus group apart from the researcher) (see section 3.5.4.3).

Data protection impact assessment

Since audio and video recordings of human participants are regarded as personal data under GDPR (European Parliament and Council of European Union, 2016), a Data Protection Impact Assessment (DPIA) was required as part of the ethical approval process. The DPIA described how key stakeholders were recruited, and how the data generated by the focus group discussions were to be collected, stored, processed, used, and deleted. It also provided details on data security risks which had been identified and what steps had been taken to reduce these risks.

These details have previously been described across sections 3.5.4.1-3.5.4.5, however a copy of the completed DPIA is also provided in appendix 7.

3.7 Summary

This chapter has outlined the underlying philosophical considerations that have influenced this study. It has described how priority was given to addressing the research questions, which have driven the adoption of a mixed methods methodology aligned with post-positivism and interpretivism methodological frameworks and a pragmatic epistemology. Details and discussions surrounding the chosen study design and methods were provided as were those concerned with ethics and information governance.

The data sources used for the quantitative component of the study were introduced and described. These included LIFTUPP© - the source of undergraduate longitudinal clinical assessment data – and more established methods of assessment used in under- and postgraduate dental education (undergraduate examinations and LEPs, respectively). The data linkage and processing required in preparation for analysis were detailed, as were the descriptive statistics used across all three quantitative data sources. Descriptions and a summary of the stages of GBTM generation, evaluation, and selection for both LIFTUPP© and LEP data were also provided. Finally, details on the acquisition and processing of qualitative data produced from focus groups discussions with key stakeholders in dental education were given and how these data were then interpreted and analysed was described.

The subsequent chapters (4-8) report the findings generated by the methodological approaches outlined in this chapter.

Chapter 4 - Undergraduate examinations

4.1 Introduction

This chapter presents the results of a series of analyses that explored student performance in the BDS examinations and investigated relationships between the early years of the BDS curriculum (BDS1-3) and the final examinations (BDS4/5). These data will be used to assess criterion validity (particularly the concurrent subtype) for LIFTUPP© in a later chapter (7), therefore it is important to fully describe these data and look at internal relationships within undergraduate examinations in the first instance.

4.2 Aim

To compare student performance in early year examinations (BDS1-3) with performance in the final professional degree examinations (held in BDS4/5).

This aim addresses research question 1: How does student performance in early BDS year examinations relate to their performance in the final professional BDS degree examinations? – (see chapter 2, section 2.2). However, the results of the analyses will also be used to compare undergraduate examination performance with LIFTUPP© performance data (see chapter 7), which will address research question 2b: What is the association between undergraduate longitudinal clinical assessment and standalone assessment methods? (Criterion validity – concurrent) – (see chapter 2, section 2.2).

4.3 Method

In addition to summary statistics, which were used to describe each numerically scored BDS1-5 assessment in each cohort (see chapter 3, section 3.5.3.5), scatter plots with linear regression lines, R^2 values and Pearson correlation coefficients (r) (or Spearman's as appropriate) were produced to assess associations between performance in each of the early BDS examinations (i.e., BDS1-3) and performance in each of the final BDS examinations (held in BDS4 and 5). These analyses were required to answer research question 1 (see chapter 2, section 2.2).

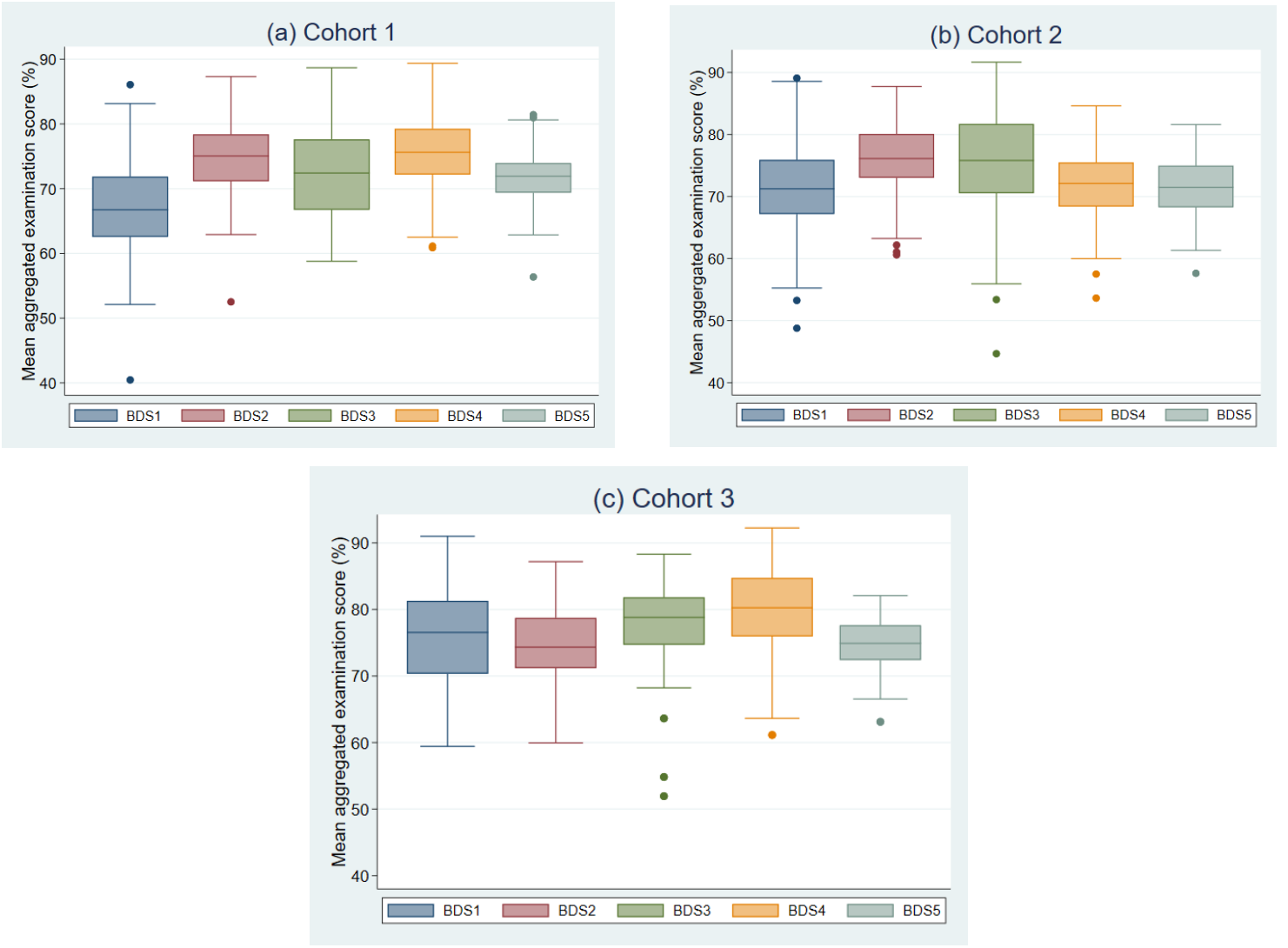
An aggregated performance score was calculated for each student by calculating a mean score from all assessments in each BDS year as a percentage (see chapter 3, section 3.5.3.5). These scores were subsequently categorised into equal groups with an indicator for top fifth performance in the final assessments. Categorical scores in BDS1-3 were cross tabulated with categorical scores in BDS4 and BDS5, and Fisher's Exact tests were used to test the associations.

Finally, c-statistics (Bamber, 1975; Hanley and McNeil, 1982) were derived from receiver operating characteristic (ROC) curve analysis between performance in the BDS1-3 examinations and top fifth performance in each of the final (BDS4/5) examinations. These measured how well early year performances predicted a top fifth performance in the finals. C-statistics <0.5 would indicate a poor relationship and c-statistics = 0.5 would indicate the model was no better classifying the outcomes than random chance.

4.4 Results

4.4.1 Summary statistics

Summary statistics for numerical examination results (for both individual examinations and mean aggregated BDS year examination performance) produced by each cohort are provided in tabular form in the appendix 8. Mean aggregated examination scores (per BDS year) are also represented graphically by boxplots in Figures 4.1a-c.



Figures 4.1 (a-c) - Boxplots displaying minimum, Q1, median, Q3 and maximum statistics for mean aggregated examination results (in percentages) for each BDS year.

It was noted that student examination scores in each cohort were generally high with low variance, which may make it harder to distinguish between groups when comparing these results against LIFTUPP© performance (see chapter 7).

Tables 4.1-4.3 display the frequencies of grades awarded for each examination in cohorts 1, 2 and 3, respectively. These frequencies are also represented graphically in Figures 4.2-4.4. "Written" examination grades are determined by the University of Glasgow from the cumulative totals of MCQ and MSA examinations in BDS1 and BDS2. In BDS3, the anatomy examination also contributes to the calculation for written examination grades. The written examination grade for BDS4 is solely based on one MSA examination.

At this point it is worth acknowledging that Glasgow Dental School uses the Modified-Angoff method (Bellara, 2018) to standard set most of its undergraduate examinations, which establishes the cut score for overall pass/fail. This process requires a panel of at least seven standard setters (across all dental subjects) to review the questions within each assessment and decide what barely passing candidates would score on each question. Total assessment scores are then pooled before mean, median and range of scores are calculated. These are then used to inform discussions on what the pass mark should be.

The borderline regression method (BRM) (Kramer et al., 2003; Hejri et al., 2013) is also used as part of the standard setting procedure for the BDS4 case presentation. It is also used as an additional check for BDS5 OSCE (which is standard set using the Modified-Angoff method). The BRM requires assessors to award each candidate a "global score" for the assessment (BDS4 case presentation) or for each station within the assessment (BDS5 OSCE). Global scores are based on a 4-point scale: "Fail"; "Borderline"; "Pass"; and "Good Pass". Numerical assessment scores are regressed on the awarded global scores to provide a linear equation, and the "Borderline" global score is substituted into the equation to predict the cut-score for pass/fail. For the BDS5 OSCE, the standard error of measurement (SEM) is also calculated to inform decisions on whether candidates with examination scores close to the borderline mark should pass or fail (McManus, 2012), as assessors must be confident that those who pass the examination have reached the level of "safe beginner" (GDC, 2015a). Grades awarded for the BDS5 OSCE are determined by a modification of a method

described by Roberts et al. (2006), whereas all other BDS examinations use the methods outlined by the University of Glasgow (University of Glasgow, 2020; McKerlie, Accessed 2021).

In all three cohorts, few students received failing grades (E, F, G or H (University of Glasgow, 2020)) across the five-year curriculum. Most cohort 1 failures occurred in the BDS1 examinations where 8.89% (n = 8/90) and 7.87% (n =7/89) were unsuccessful in the written examinations and OSCE, respectively. Cohort 3 also had most failures in BDS1 with 8.22% (n = 6/73) and 2.74% (n = 2/73) who did not pass the summative essay and OSCE assessments. However, cohort 3 also produced the highest average results for BDS1 compared to cohorts 1 and 2. The most failures in cohort 2 were seen in BDS3 where 10.87% (n = 10/92) and 3.26% (n = 3/92) did not pass the written and OSCE assessments, respectively.

A “C” was the most awarded grade for every BDS1 examination in both cohorts 1 and 2. It was also the most common grade for the BDS5 OSCE in cohort 1 and the BDS3 OSCE in cohort 3. A “B” was the most frequently awarded grade for the BDS2 written and BDS4 MSA examinations in all cohorts, the BDS3 OSCE, BDS4 case presentation in cohorts 1 and 2, and the BDS5 OSCE in cohorts 2 and 3. It was also the most common grade for the both the BDS1 summative essay and OSCE in cohort 3. An “A” grade was most frequently awarded in the BDS2 OSCE for all three cohorts.

There was no single commonly awarded grade for the BDS3 written and BDS4 case presentation examinations in cohorts 1 and 3, respectively. For the former, an equal percentage (35.63%) of Bs and Cs were awarded, whereas the latter had an equal number (23.61%) of Bs, Cs and Ds.

Table 4.1 – Cohort 1: Frequency of grades awarded for each BDS professional examination. MCQ = Multiple-choice question. MSA = Multiple-short answer. OSCE = Objective structured clinical examination.

Grade	BDS1			BDS2		BDS3		BDS4		BDS5
	Summative Essay (n = 89)	**Written (2 x MCQ)** (n = 90)	**OSCE*** (n = 89)	**Written (MCQ + MSA)** (n = 88)	**OSCE*** (n = 88)	**Written (Anatomy + MCQ + MSA)** (n = 87)	**OSCE** (n = 87)	**MSA (final)*** (n = 84)	**Case presentation (final)*** (n = 84)	**OSCE (final)** (n = 82)
A	19 (21.35%)	9 (10.00%)	7 (7.87%)	7 (7.95%)	**48 (54.54%)**	7 (8.05%)	15 (17.24%)	15 (17.86%)	23 (27.38%)	14 (17.07%)
B	23 (25.84%)	29 (32.22%)	9 (10.11%)	**37 (42.05%)**	32 (36.36%)	**31 (35.63%)**	**36 (41.38%)**	**40 (47.62%)**	**25 (29.76%)**	26 (31.71%)
C	**35 (39.33%)**	**34 (37.78%)**	**41 (46.07%)**	30 (34.09%)	8 (9.09%)	**31 (35.63%)**	22 (25.29%)	15 (17.86%)	23 (27.38%)	**27 (32.93%)**
D	12 (13.48%)	10 (11.11%)	25 (28.09%)	11 (12.50%)		17 (19.54%)	10 (11.49%)	12 (14.29%)	8 (9.52%)	14 (17.07%)
E		7 (7.78%)	7 (7.87%)							
F										
G										
H										

The total of frequency percentages for some assessments (marked by an *) do not equate to 100% since all results have been rounded to two decimal places. Mode grades highlighted in **bold.** Cells where n<5 are greyed out.

Table 4.2 – Cohort 2: Frequency of grades awarded for each BDS professional examination. MCQ = Multiple-choice question. MSA = Multiple-short answer. OSCE = Objective structured clinical examination.

Grade	BDS1			BDS2		BDS3		BDS4		BDS5
	Summative Essay (n = 92)	**Written (2 x MCQ)** (n = 92)	**OSCE** (n = 92)	**Written (MCQ + MSA)** (n = 92)	**OSCE*** (n = 92)	**Written (Anatomy + MCQ + MSA) *** (n = 92)	**OSCE** (n = 92)	**MSA (final)** (n = 91)	**Case presentation (final)*** (n = 91)	**OSCE (final)** (n = 93)
A	15 (16.30%)	9 (9.78%)	8 (8.70%)	21 (22.83%)	**52 (56.52%)**	18 (19.57%)	7 (7.61%)	11 (12.09%)	17 (17.71%)	15 (16.13%)
B	27 (29.35%)	20 (21.74%)	19 (20.65%)	**48 (52.17%)**	25 (27.17%)	**31 (33.70%**	**35 (38.04%)**	**23 (25.27%)**	**31 (32.29%)**	**33 (35.48%)**
C	**37 (40.22%)**	**40 (43.48%)**	**42 (45.65%)**	14 (15.22%)	10 (10.87%)	26 (28.26%)	34 (36.96%)	42 (46.15%)	28 (29.17%)	32 (34.41%)
D	11 (11.96%)	16 (17.39%)	21 (22.83%)	7 (7.61%)	5 (5.43%)	7 (7.61%)	13 (14.13%)	11 (12.09%)	15 (15.63%)	12 (12.90%)
E		7 (7.61%)				8 (8.70%)				
F										
G										
H										

The total frequency percentages for some assessments (marked by an *) do not equate to 100% since all results have been rounded to two decimal places. Mode grades highlighted in **bold.** Cells where n<5 are greyed out.

Table 4.3- – Cohort 3: Frequency of grades awarded for each BDS professional examination. MCQ = Multiple-choice question. MSA = Multiple-short answer. OSCE = Objective structured clinical examination.

Grade	BDS1			BDS2		BDS3		BDS4		BDS5
	Summative Essay (n = 73)	**Written (2 x MCQ)** (n = 73)	**OSCE*** (n = 73)	**Written (MCQ + MSA)** (n = 72)	**OSCE*** (n = 72)	**Written (Anatomy + MCQ + MSA)** (n = 72)	**OSCE*** (n = 72)	**MSA (final)*** (n = 72)	**Case presentation (final)** (n = 72)	**OSCE (final)*** (n = 69)
A	13 (17.80%)	14 (19.18%)	11 (15.07%)	8 (11.11%)	**38 (52.78%)**	27 (37.50%)	7 (9.72%)	40 (55.56%)	18 (25.00%)	23 (33.33%)
B	**21 (28.77%)**	27 (36.99%)	**31 (42.47%)**	**18 (25.00%)**	22 (30.56%)	**37 (51.39%)**	24 (33.33%)	**19 (26.39%)**	**17 (23.61%)**	**36 (52.17%)**
C	17 (23.29%)	**29 (39.72%)**	17 (23.29%)	25 (34.72%)	8 (11.11%)	6 (8.33%)	**32 (44.44%)**	11 (15.28%)	**17 (23.61%)**	9 (13.04%)
D	16 (21.92%)		12 (16.44%)	18 (25.00%)			6 (8.33%)		**17 (23.61%)**	
E	6 (8.22%)									
F										
G										
H										

NOTES: The total of frequency percentages for some assessments (marked by an *) do not equate to 100% since all results have been rounded to two decimal places. Mode grades highlighted in **bold.** Cells where n<5 are greyed out.

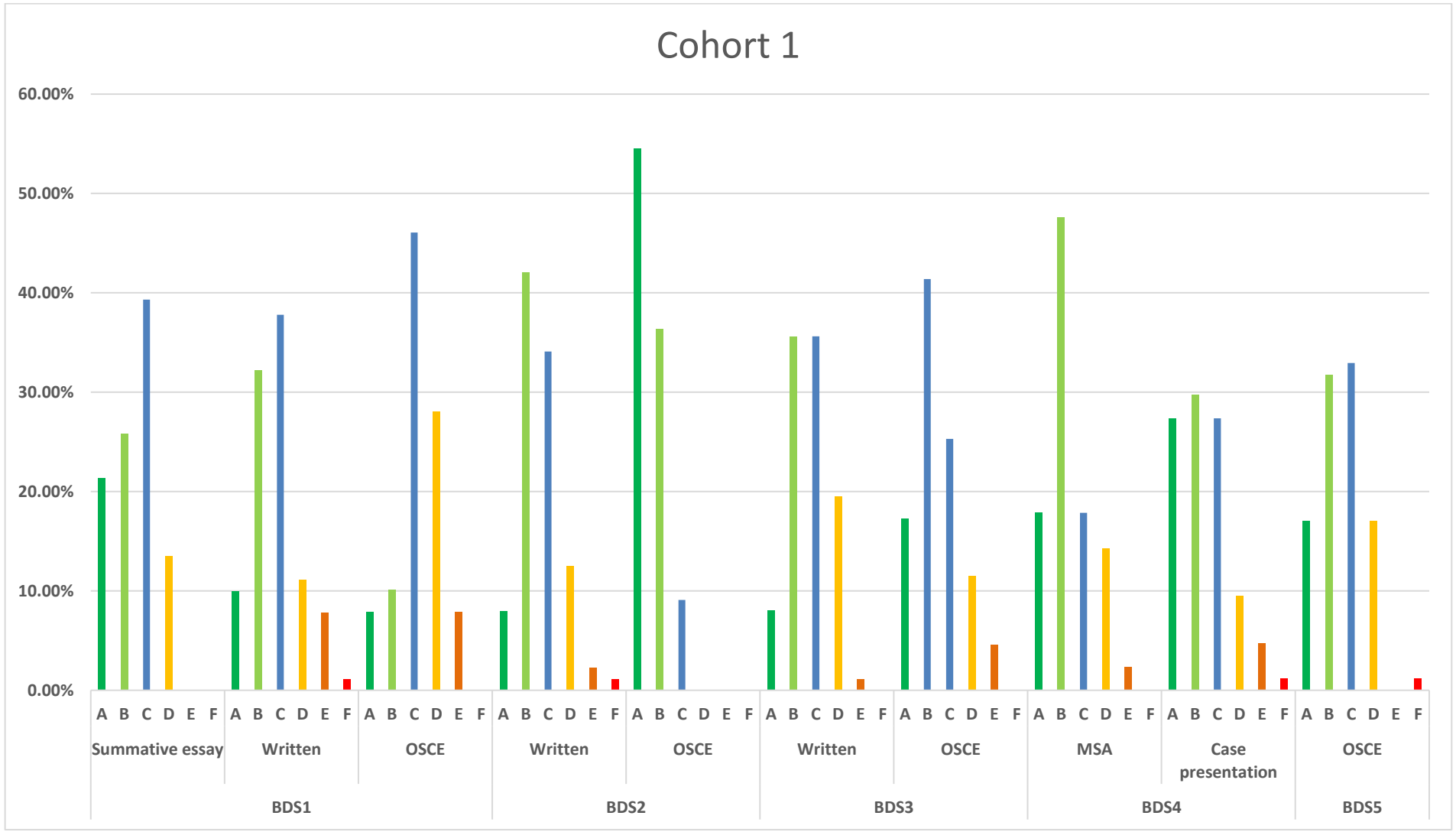


Figure 4.2 **– Cohort 1: Frequency of grades awarded for each BDS professional examination. OSCE = Objective structured clinical examination. MSA = Multiple-short answer.**

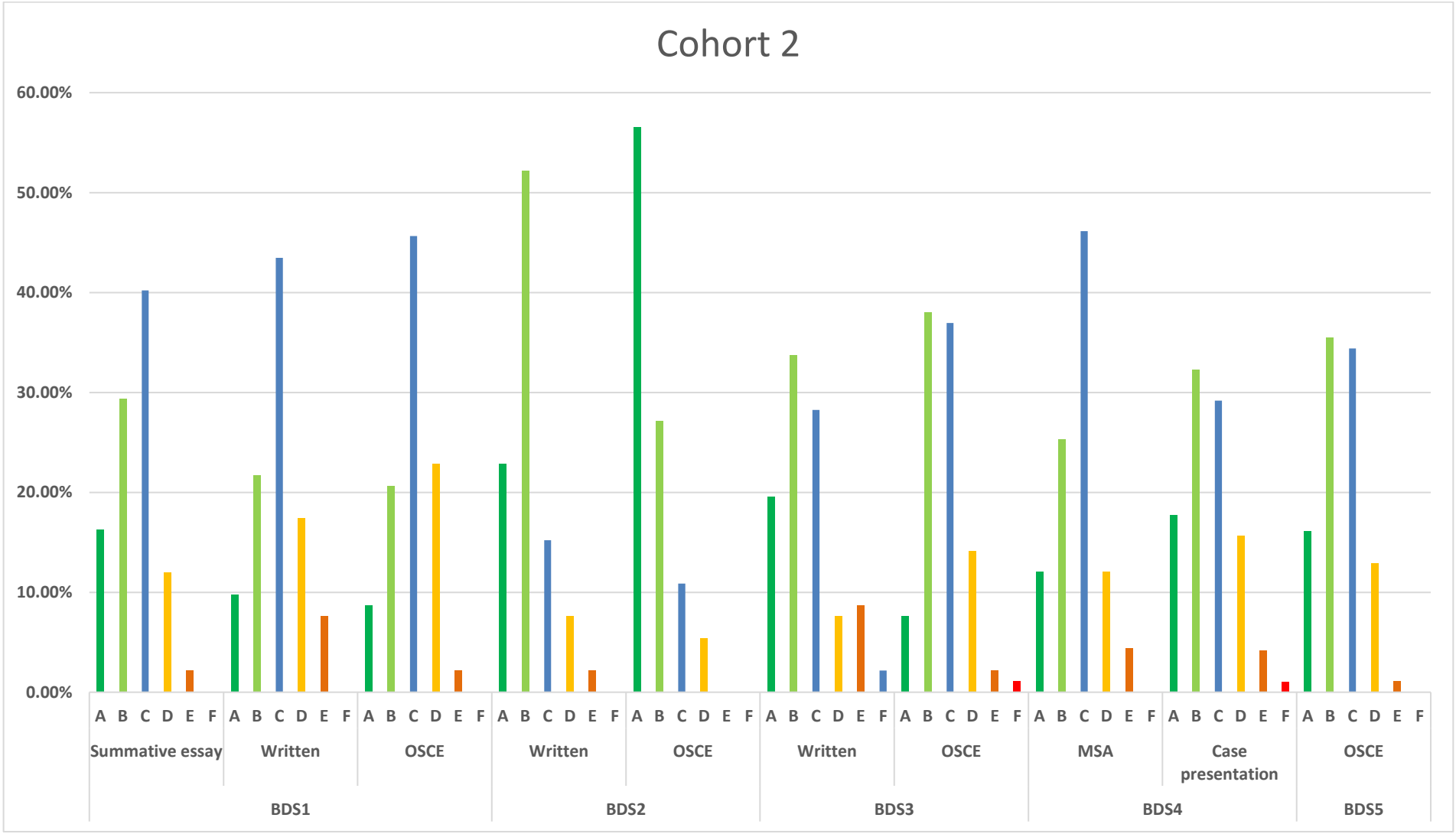


Figure 4.3 – Cohort 2: Frequency of grades awarded for each BDS professional examination. OSCE = Objective structured clinical examination. MSA = Multiple-short answer.

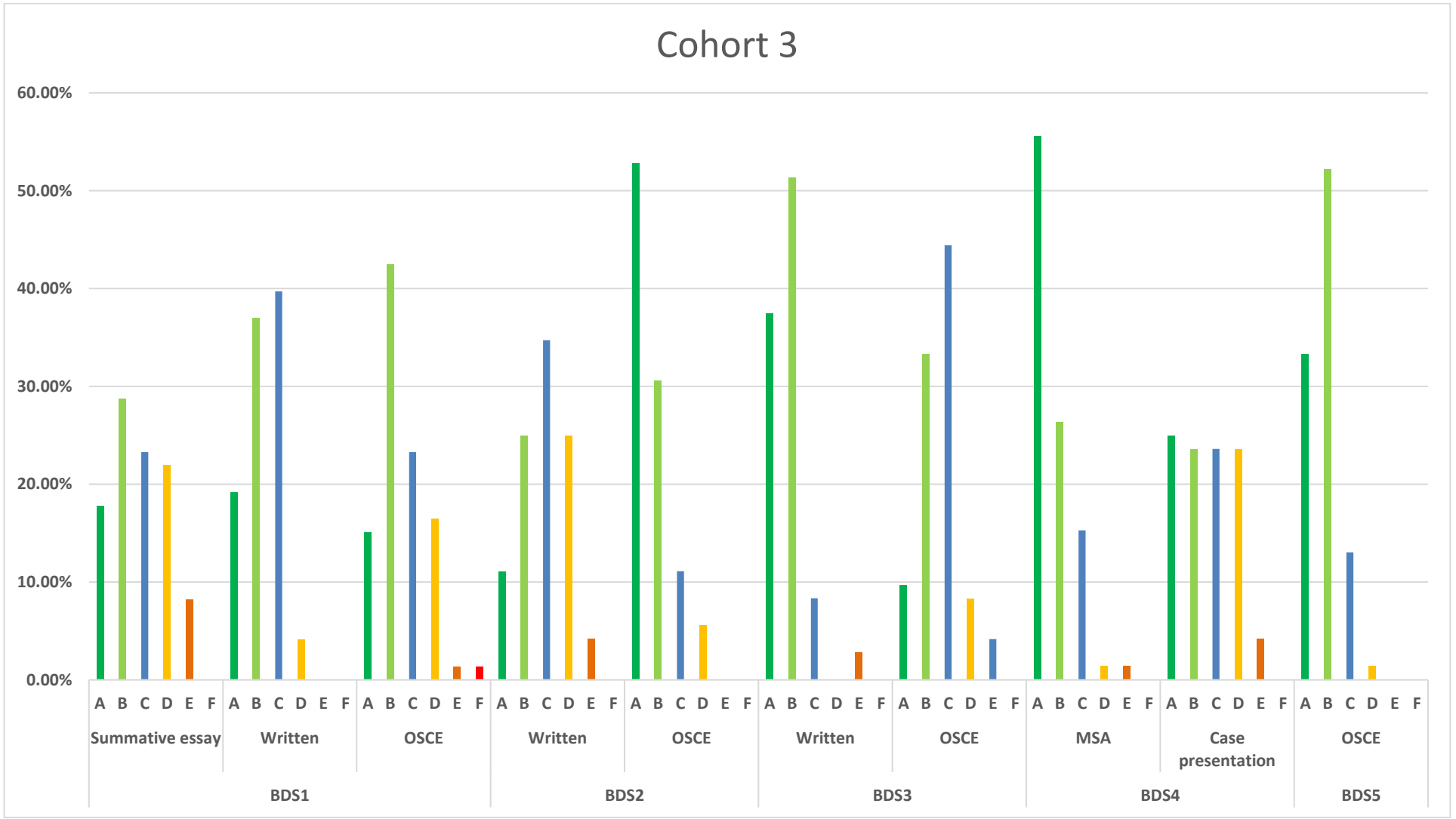
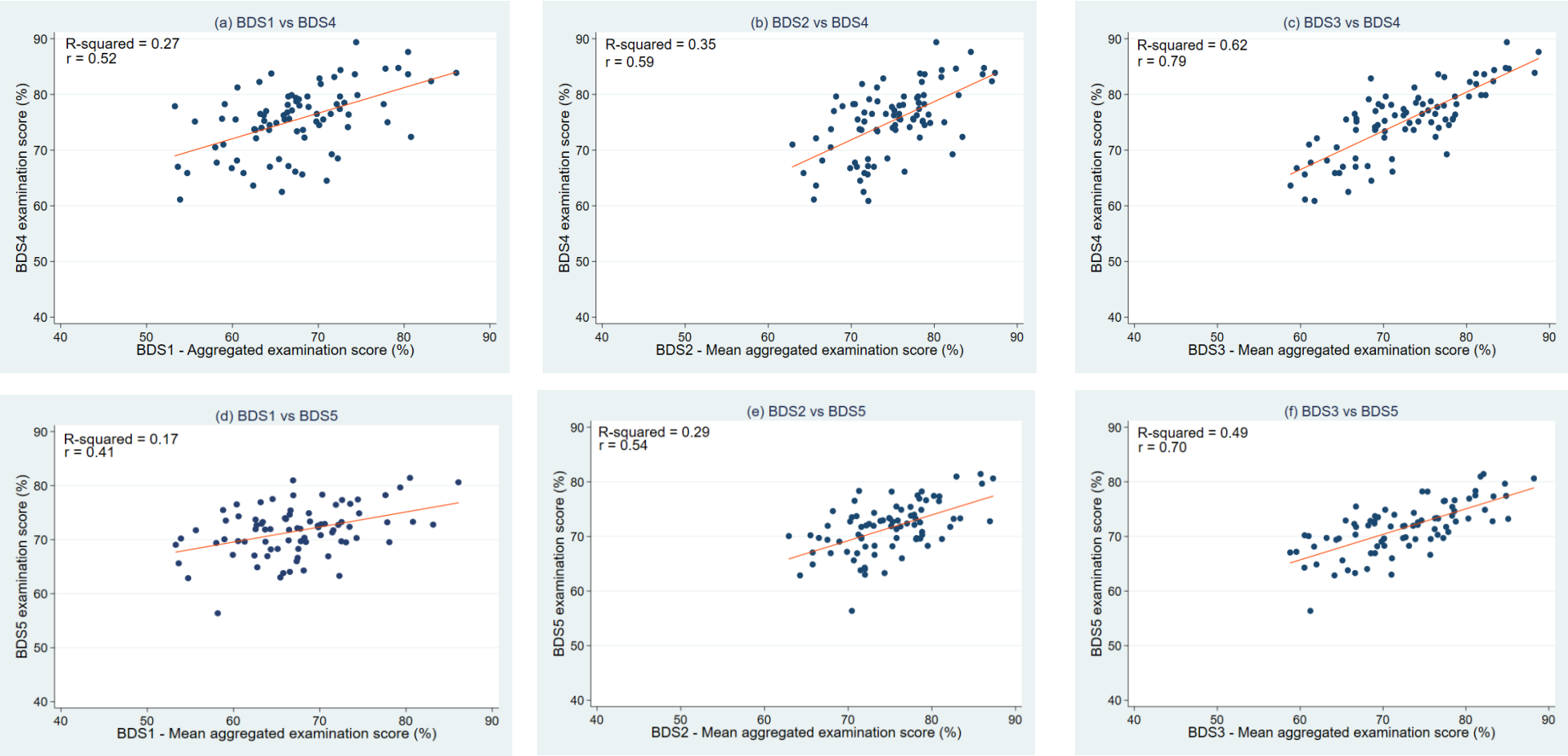


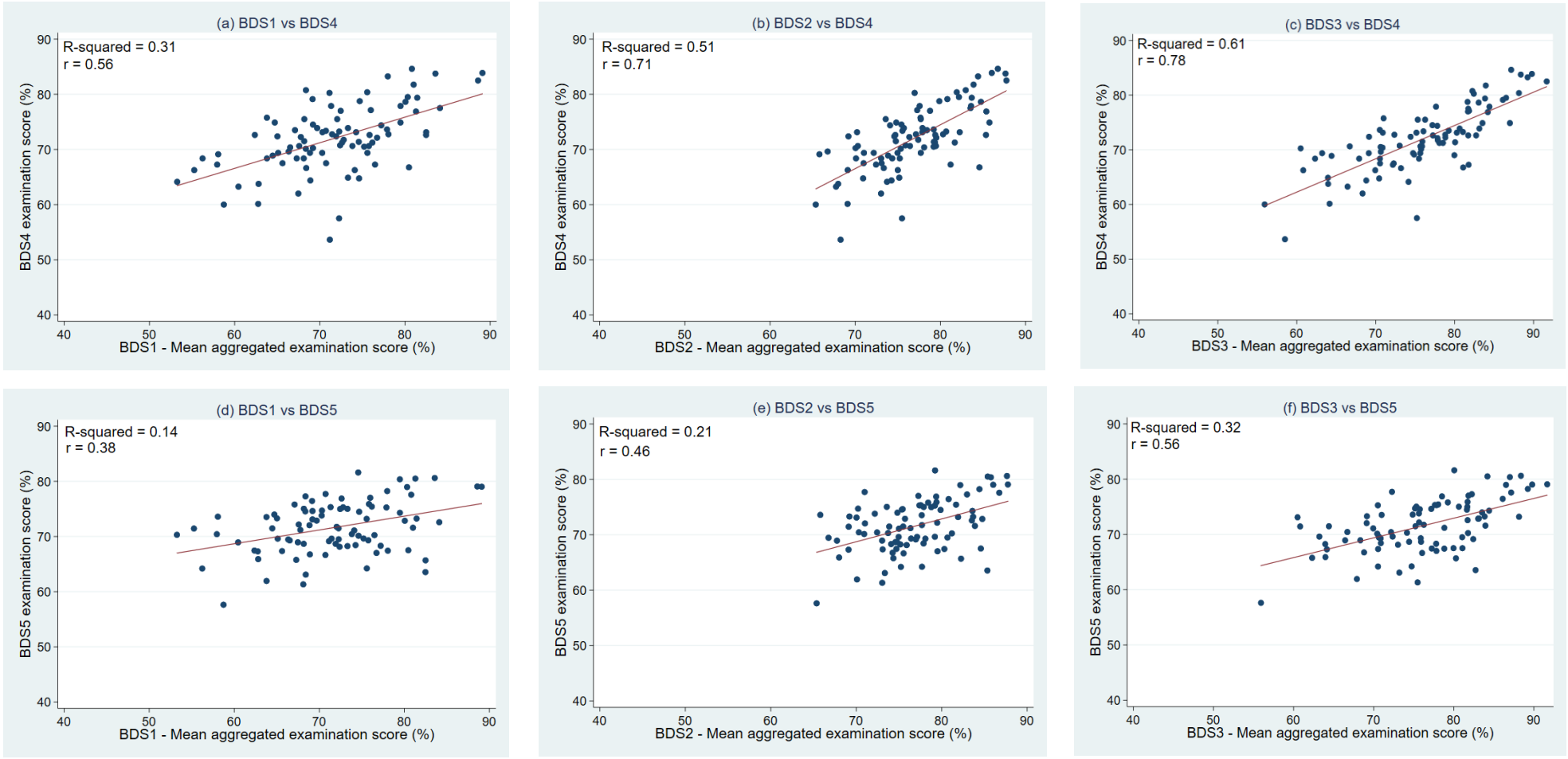
Figure 4.4 **– Cohort 3: Frequency of grades awarded for each BDS professional examination. OSCE = Objective structured clinical examination. MSA = Multiple-short answer.**

4.4.2 Association between early (BDS1-3) and final (BDS4/5) examinations

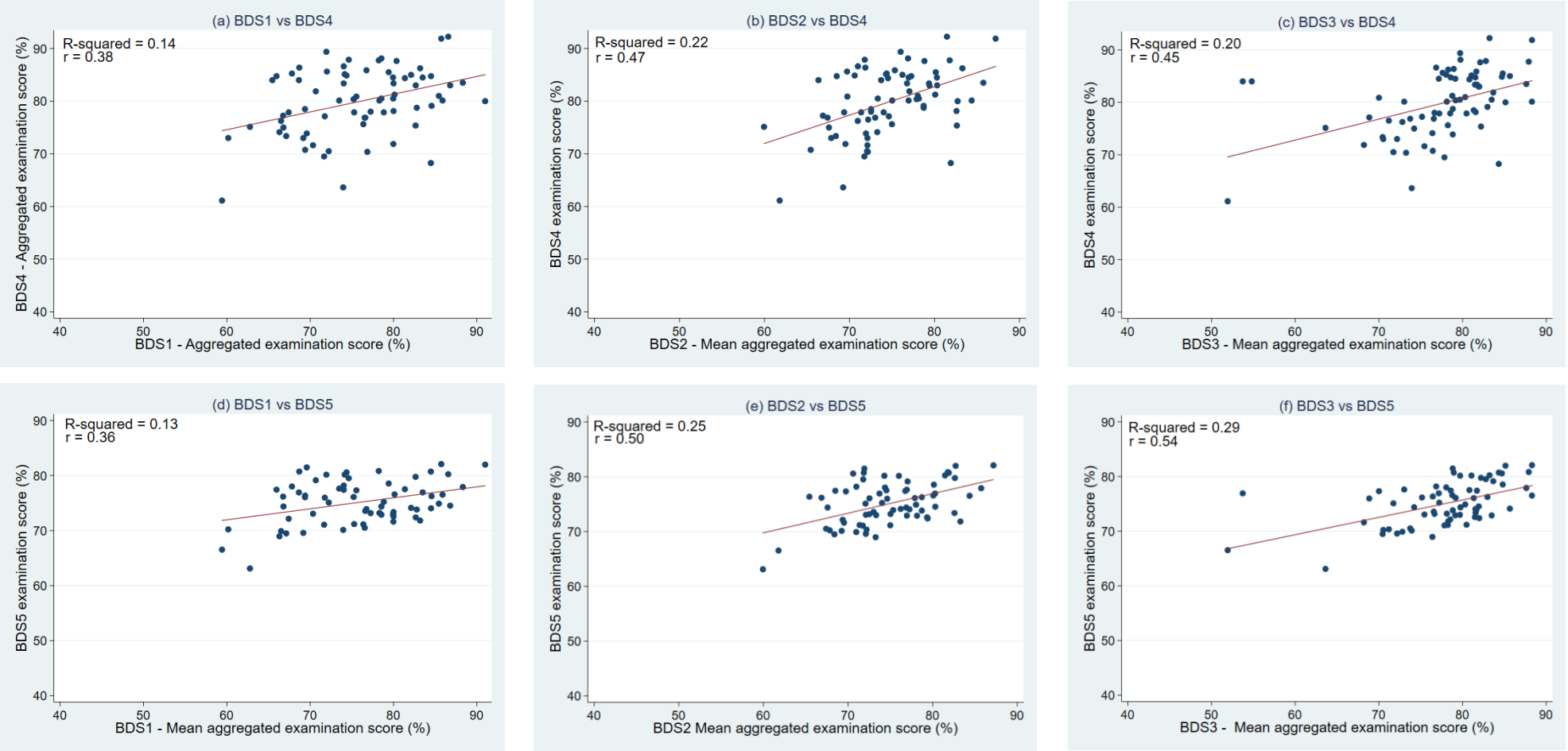
Figures 4.5-4.7(a-f) present scatter plots of BDS1-3 vs BDS4/5 for all three cohorts. Regression lines are presented along with correlation coefficients. Correlation coefficients ranged from 0.36-0.79 – indicating, for all early BDS years (1-3) assessments, there was a moderate to high positive correlation with performance in the finals (BDS4/5). The lowest correlation (0.36) was observed between the BDS1 and BDS5 examinations in cohort 3 (Figure 4.7d), and the highest correlation (0.79) was seen between the BDS3 and BDS4 examinations in cohort 1 (Figure 4.5c).



Figures 4.5 (a-f) - Cohort 1: Scatter plots between BDS1/2/3 and BDS4 (a-c) and BDS5 (d-f) examination performances (with R^2 and r values).



Figures 4.6 (a-f) - Cohort 2: Scatter plots between BDS1/2/3 and BDS4 (a-c) and BDS5 (d-f) examination performances (with R^2 and r values).



Figures 4.7 (a-f) - Cohort 3: Scatter plots between BDS1/2/3 and BDS4 (a-c) and BDS5 (d-f) examination performances (with R^2 and r values).

4.4.3 Association between early (BDS1-3) and top fifth performance in the final (BDS4/5) examinations

The choice of analysing a top-fifth performance in the final assessments (BDS4/5) was based on selecting students who received an "A" grade. To ensure sufficient cell sizes when cross tabulating with early years performances, BDS1-3 aggregated scores were categorised into thirds, as any more groups would lead to a large number of "zero" cells.

Summary statistics for thirds of mean aggregated performance in the early examinations and fifths of performance in the final examinations are shown for cohorts 1, 2 and 3 in appendix 8 as are data for cross tabulations between thirds of mean aggregated performance in the early examinations and fifths of final examination performance in each cohort. Figures 4.8-4.13 show the percentages of students in each third of performance in the early (BDS1-3) examinations who produced a top fifth (20%) performance in the final examinations (BDS4/5).

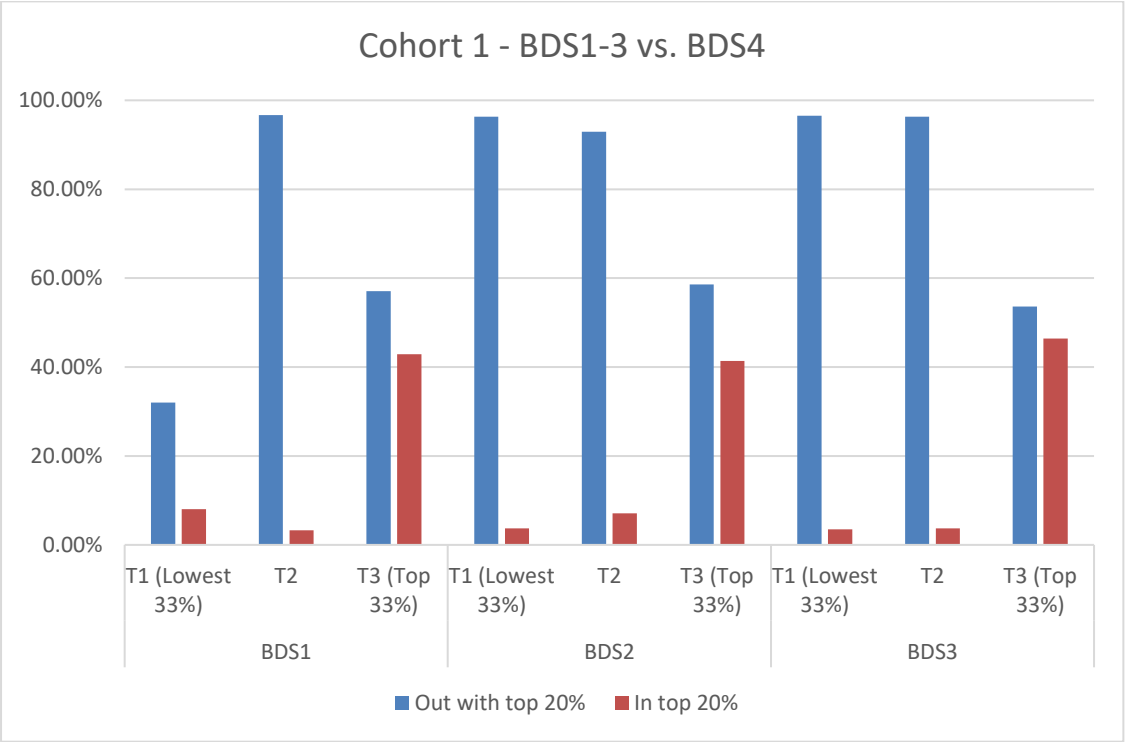


Figure 4.8 – Cohort 1: Percentage of students who achieved a top fifth performance in the final written (multiple-short answer (MSA)) examination (BDS4) according to thirds of examination performance (T) in BDS1-3.

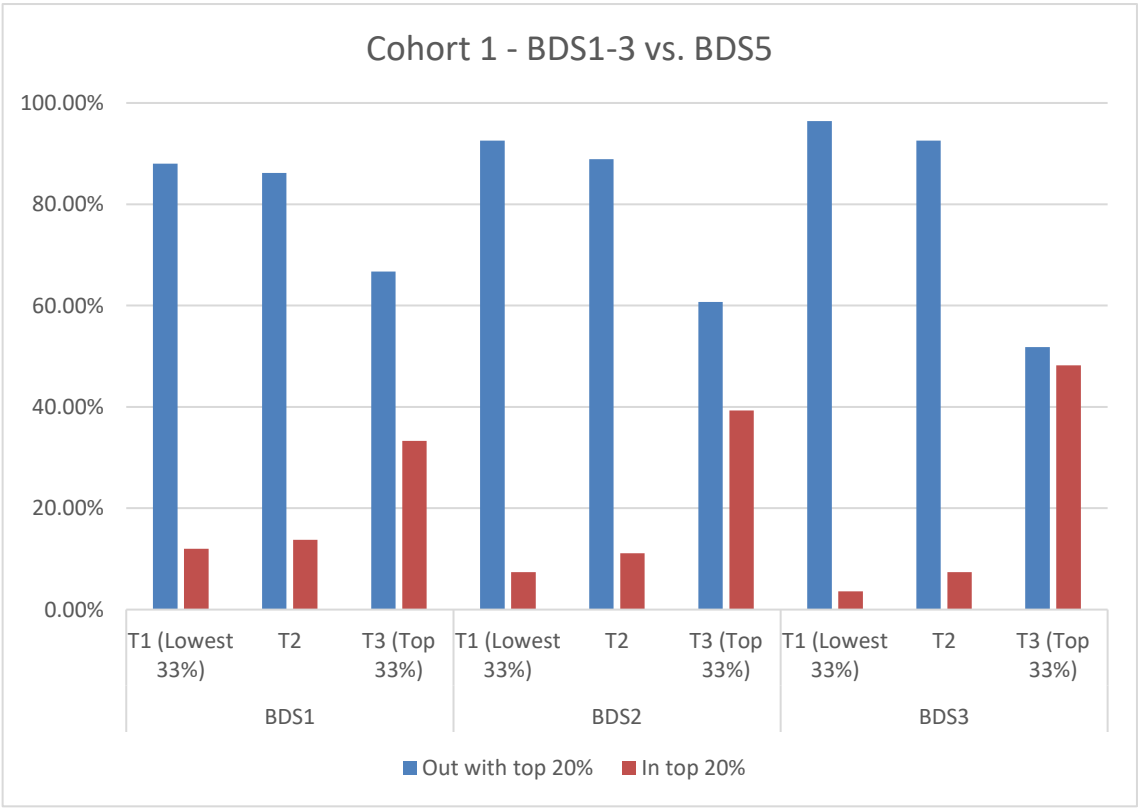


Figure 4.9 – Cohort 1: Percentage of students who achieved a top fifth performance in the final objective structured clinical examination (OSCE) (BDS5) according to thirds of examination performance (T) in BDS1-3.

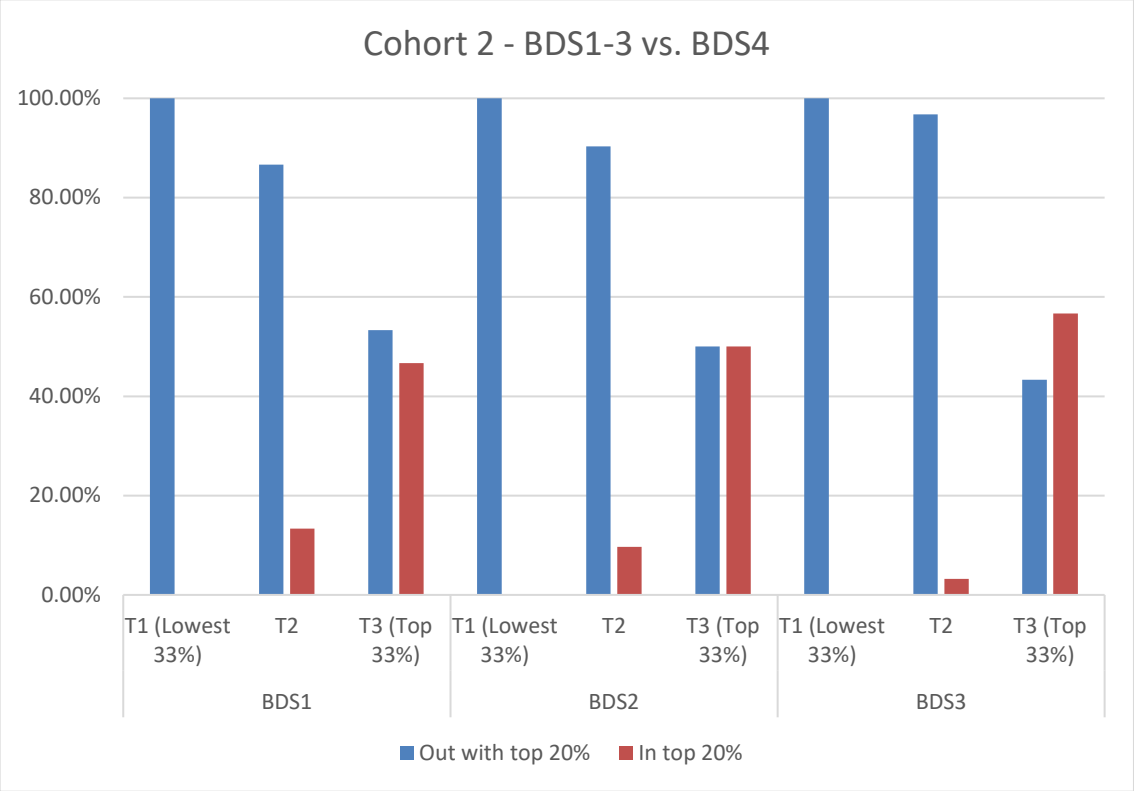


Figure 4.10 – Cohort 2: Percentage of students who achieved a top fifth performance in the final written (multiple-short answer (MSA)) examination (BDS4) according to thirds of examination performance (T) in BDS1-3.

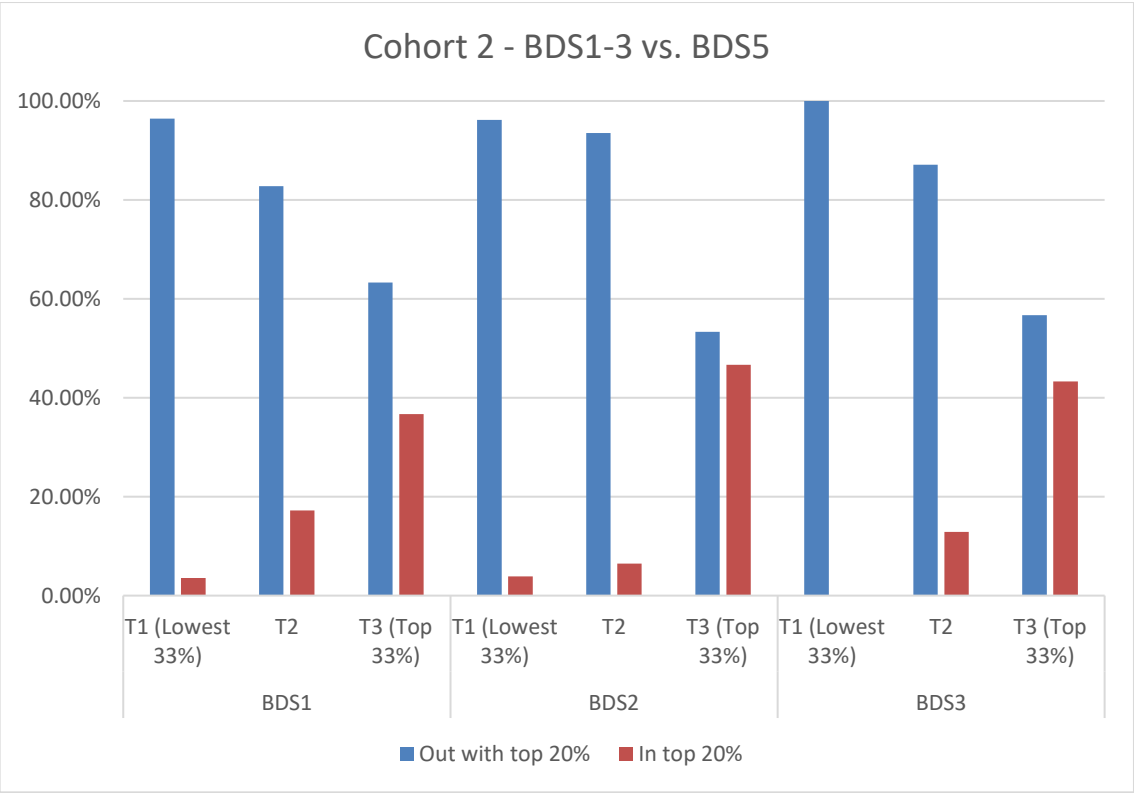


Figure 4.11 – Cohort 2: Percentage of students who achieved a top fifth performance in the final objective structured clinical examination (OSCE) (BDS5) according to thirds of examination performance (T) in BDS1-3.

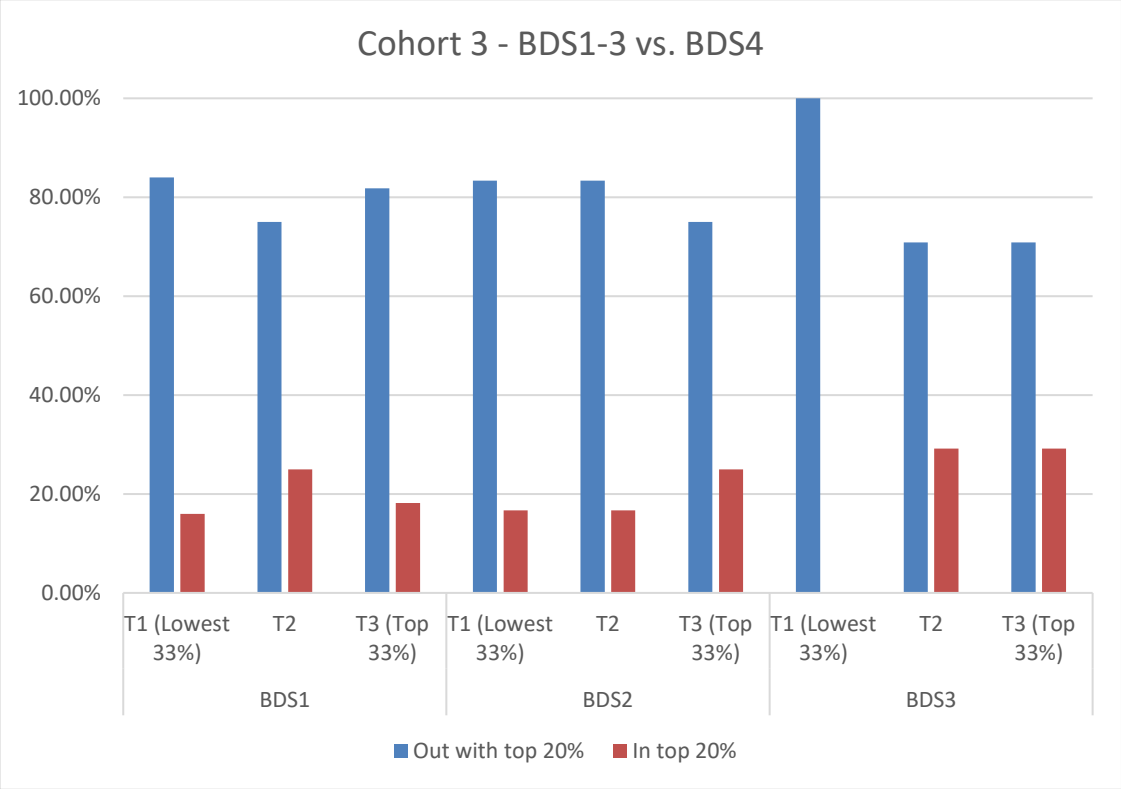


Figure 4.12 – Cohort 3: Percentage of students who achieved a top fifth performance in the final written (multiple-short answer (MSA)) examination (BDS4) according to thirds of examination performance (T) in BDS1-3.

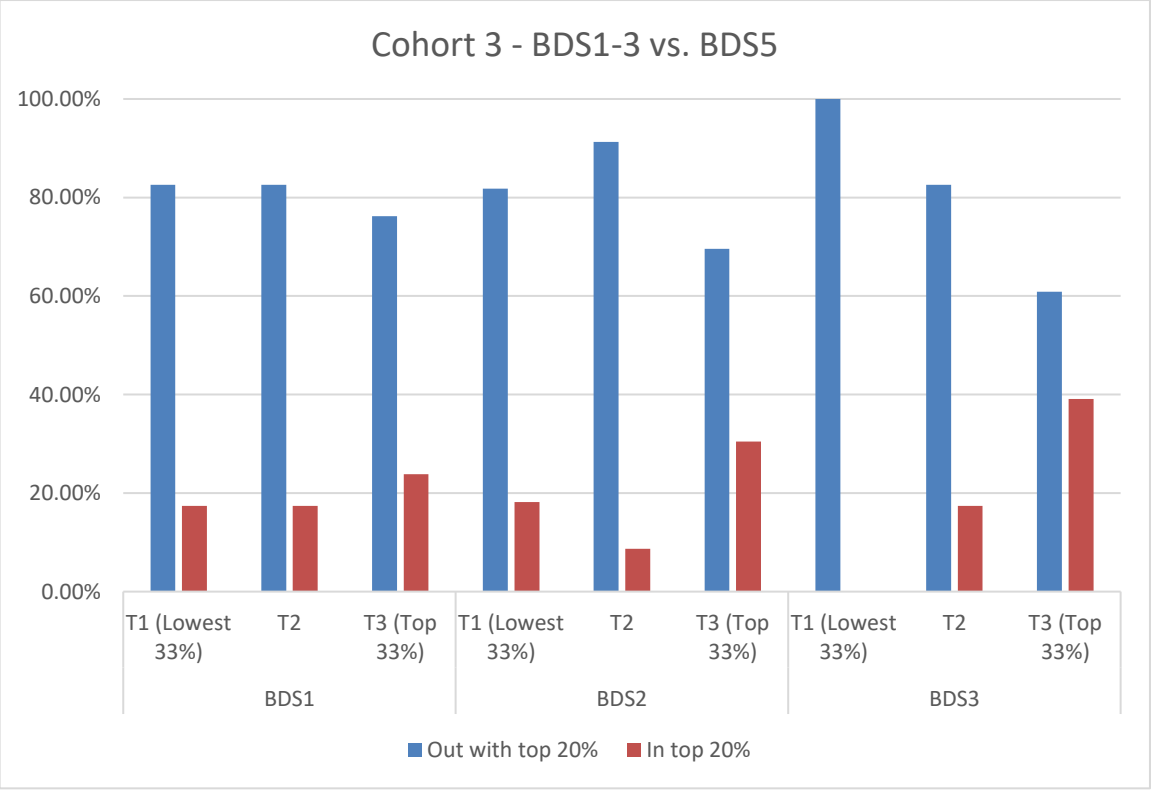


Figure 4.13 – Cohort 3: Percentage of students who achieved a top fifth performance in the final objective structured clinical examination (OSCE) (BDS5) according to thirds of examination performance (T) in BDS1-3.

Across all three cohorts, very few or no students in the lowest performing third of the early examinations produced a top fifth performance in either of the final examinations.

In cohort 1, only 8% (n = 2/25), 3.7% (n = 1/26) and 3.5% (n = 1/29) of the lowest performing third of BDS1/2/3 students (respectively) achieved scores in the top fifth for the BDS4 final written examination, compared to 42.9% (n = 12/28), 41.4% (n = 12/29) and 46.4% (n = 13/28) of students who were in the top performing third of BDS1/2/3 (all $p<0.01$). A top fifth performance in the BDS5 final OSCE was achieved by 12% (n = 3/25), 7.4% (n = 2/27) and 3.6% (n = 1/28) of the lowest performing third of BDS1/2/3 students (respectively), compared to 33.3% (n = 9/27), 39.3% (n = 11/29) and 48.2% (n = 13/27) of the top performing third students (BDS1 vs BDS5 $p=0.37$; both BDS2 vs BDS5 and BDS3 vs BDS5 $p<0.01$).

In cohort 2, no student in the lowest performing third of the BDS1/2/3 examinations produced top fifth scores for the BDS4 final written examination (BDS1: 0.00% (n = 0/28); BDS2: 0.00% (n = 0/27) and 0.00% (n = 0/28)). A top fifth performance in BDS4 was achieved by 46.67% (n = 14/30), 50.00% (n = 15/30) and 56.67% (n = 17/30) of students who were in the top performing third of the BDS1/2/3 examinations (respectively) (all $p<0.01$). Only 3.57% (n = 1/28), 3.85% (n = 1/26) and 0.00% (n = 0/27) of the lowest performing third of BDS1/2/3 students (respectively) were in the top fifth for the BDS5 final OSCE, compared to 36.67% (n = 11/30), 46.67% (n = 14/30) and 43.33% (n = 13/30) of the top performing third in the early examinations (BDS1 vs BDS5 $p=0.11$; both BDS2 vs BDS5 and BDS3 vs BDS5 $p<0.01$).

In cohort 3, only 16.00% (n = 4/25) and 16.67% (n = 4/24) of the lowest performing third of BDS1/2 students (respectively) achieved scores in the top fifth for the BDS4 final written examination. No student in the lowest third of the BDS3 examination produced a top fifth BDS4 performance (0.00%, n = 0/24). A top fifth BDS4 score was achieved by18.18% (n = 4/22), 25.00% (n = 6/24) and 29.17% (n = 7/24) of students who were in the top performing third in early examinations (BDS1 vs BDS4 p=003; both BDS2 vs BDS4 and BDS3 vs BDS4 $p<0.01$). A top fifth performance in the BDS5 final OSCE was achieved by 17.39% (n = 4/23) and 18.18% (n = 4/22) of the lowest performing third of BDS1/2 students

(respectively) and no student in the lowest third of the BDS3 examinations delivered a top fifth BDS5 performance (0.00%, n = 0/22). In comparison, 23.81% (n = 5/21), 30.43% (n = 7/23) and 39.13% (n = 9/23) of the top performing third students in the BDS1/2/3 examinations (respectively) achieved a top fifth BDS5 performance (BDS1 vs BDS5 p=0.16; BDS2 vs BDS5 p=0.03; BDS3 vs BDS5 p<0.01).

Using logistic regression to generate the ROC (and therefore c-statistics) revealed there was a degree of predictive capacity for students delivering a top fifth performance in their final examinations based on their overall performance in each of the early examinations. This predictive capacity was evident since the c-statistic was >0.5 in each logistic regression model – varying from 0.77 to 0.82, 0.69 to 0.82 and 0.52 to 0.74 in cohorts 1, 2 and 3, respectively (see appendix 8).

4.5 Summary

This chapter has reported on investigations undertaken to explore the relationships between sets of undergraduate examination data produced by three recently graduated student cohorts from Glasgow Dental School (2017-19).

There were moderate to high positive correlations between exam performance in the early BDS years (1-3) and performance in the final years (BDS4-5) for all three cohorts - which generally were stronger in BDS3. This result remained consistent when the relationship between early exam performance and achieving a score in the top-fifth at finals was examined, however there was a reasonable amount of variation across the cohorts, suggesting students may recover from poorer performances in the early BDS years and perform well in their final examinations.

The findings of this chapter are discussed further in chapter 9 in conjunction with results reported in other chapters of this thesis.

Chapter 5 - Exploring content validity of undergraduate longitudinal clinical assessment

5.1 Introduction

The following chapter presents the results of the analyses of LIFTUPP© data obtained from three undergraduate student cohorts. Summary statistics are first presented from two perspectives (students and assessors) before the results of GBTM generation, evaluation, and selection for student clinical LIFTUPP© data are presented.

The summary results provide an initial overview of the student clinical performance data and assessment patterns recorded by LIFTUPP©. These are of interest for each cohort and may also facilitate understanding of the trajectory patterns produced through GBTM. The trajectory models themselves will contribute evidence towards testing the content validity of longitudinal clinical assessment data and establish the usefulness of GBTMs for modelling these data.

Since longitudinal data can be modelled differently according to various criteria – such as model data distribution, threshold performance scores (where applicable), and restrictions on the minimum number of participants per trajectory group - multiple models are presented to demonstrate how patterns of LIFTUPP© performance are influenced according to these criteria. These investigations were necessary to determine which GBTM(s) may represent LIFTUPP© data best and could be selected for comparisons with other forms of assessment in dental education – namely undergraduate examinations and LEPs – to investigate the criterion validity of longitudinal assessment (see chapter 7).

5.2 Aims

To investigate the content validity of undergraduate longitudinal assessment, both within and across academic BDS years, by determining if:

a) LIFTUPP© data trajectories over time reflect the expected profile of student clinical development.

b) multiple trajectories exist within LIFTUPP© data sets to distinguish different patterns of student clinical development.

These aims address research question 2a: What are the main patterns of longitudinal assessment over time within a year and across years? – (chapter 2, section 2.2).

5.3 Method

As detailed in chapter 3 (section 3.5.3.5), LIFTUPP© data were first summarised as follows:

- *Student perspective data*
    - Number of eligible procedures per cohort per year were summarised using histograms and, based on the data distributions observed, means, standard deviations, minimums, medians, maximums, Q1 and Q3 statistics.
    - Minimum LIFTUPP© performance indicators (per eligible procedure) were summarised using frequency tables and bar charts - stratified by cohort and BDS year within each cohort.

Eligible clinical procedures were those in which an individual student had completed all key stages of a treatment item on an individual patient - see chapter 3 (section 3.5.3.4) and appendix 2.

- *Assessor perspective data*
    - Number of clinical procedure stages assessed were summarised using histograms and, based on the data distributions observed, means, standard deviations, minimums, medians, maximums, Q1 and Q3 statistics. These data were stratified by cohort and BDS year within each cohort.

- Number of individual students assessed per cohort per year were summarised using means, standard deviations, minimums, medians, maximums, Q1 and Q3 statistics.

- LIFTUPP© performance indicators awarded to students summarised using frequency tables and bar charts - stratified by cohort and BDS year within each cohort.

Once the data were summarised, GBTMs were produced to track undergraduate clinical performance across the BDS course for each cohort using the "traj" plugin for Stata statistical software (Jones and Nagin, 2012; 2013). The processes for generation, evaluation, and selection of GBTMs for LIFTUPP© data have been described in chapter 3 (section 3.5.3.6).

5.4 Results

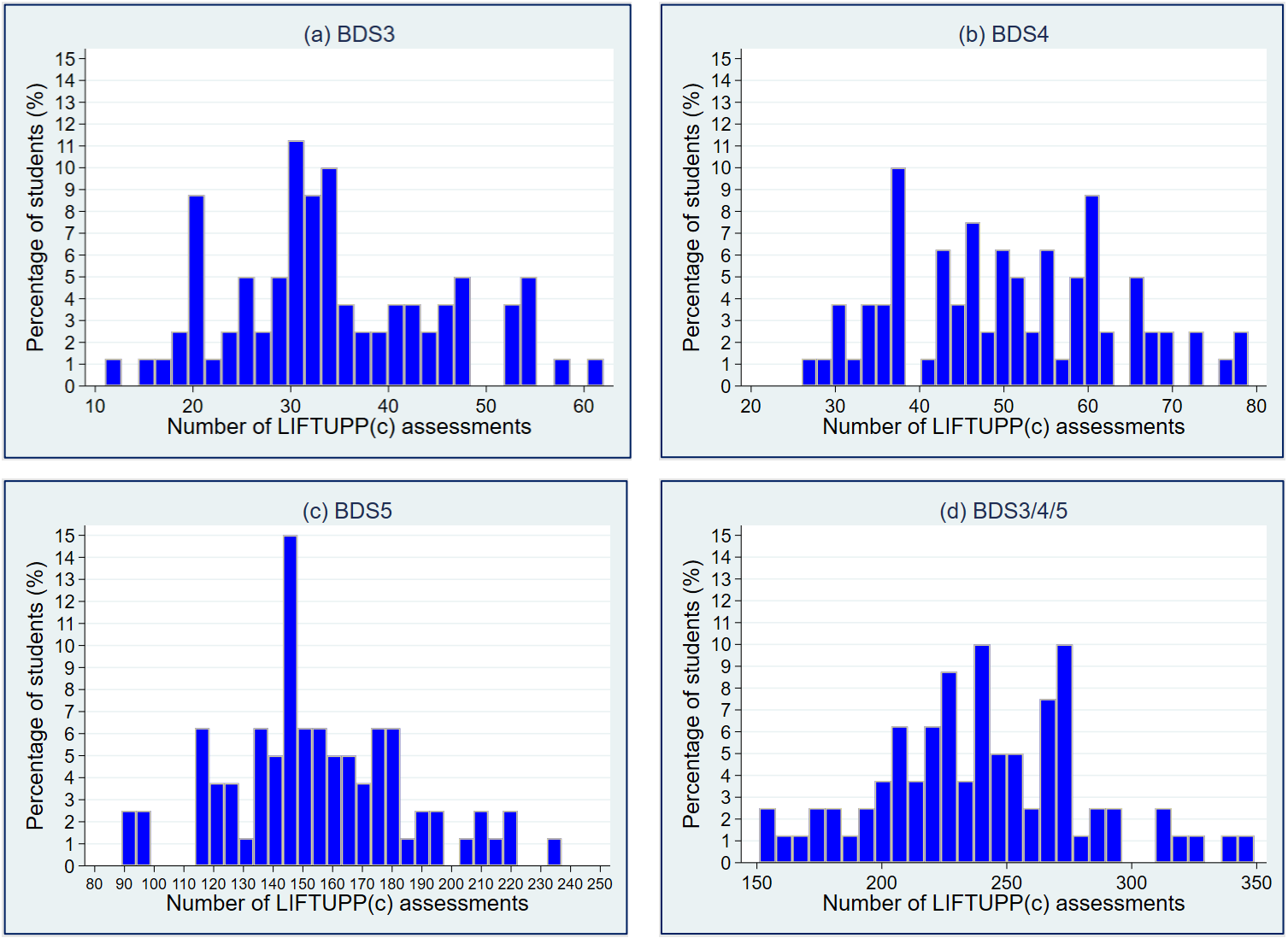
5.4.1 A description of LIFTUPP© data from student and assessor perspectives

5.4.1.1 Number of LIFTUPP© assessments – Student perspective

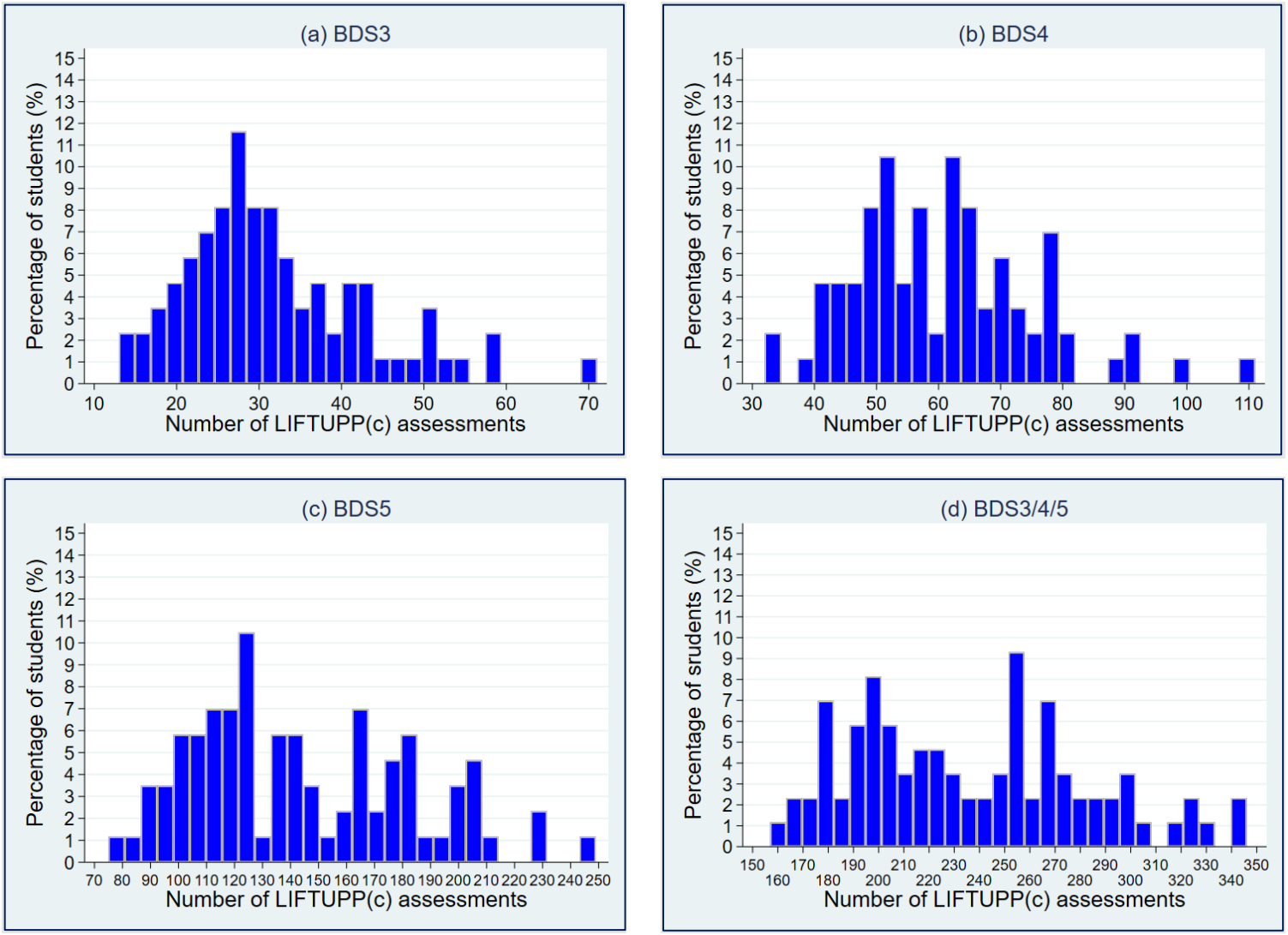
Following completion of data quality assurance, cleaning and management processes outlined in chapter 3 (section 3.5.3.4), the number of clinical LIFTUPP© procedures eligible for inclusion in the study for cohorts 1, 2 and 3 were 19,199/33,056 (60.5%), 20,312/35,573 (57.1%) and 20,817/36,016 (57.8%), respectively.

Histograms for the number of eligible clinical assessments performed per student within and across BDS years displayed either normal or slight positively-skewed distributions (see Figures 5.1–5.3 (a-d)). Summaries for this variable are available in appendix 9 and illustrated using boxplots in Figures 5.4 (a-c). In all three cohorts, the mean number of clinical assessments completed by each student increased year on year as they progressed through the BDS curriculum. The most significant increases were observed in BDS5. For example, in cohort 1 the mean number of complete clinical LIFTUPP© assessments performed per student increased from 34.5 to 50.6 to 155.0 as students went from BDS3 to BDS4 to BDS5, respectively.

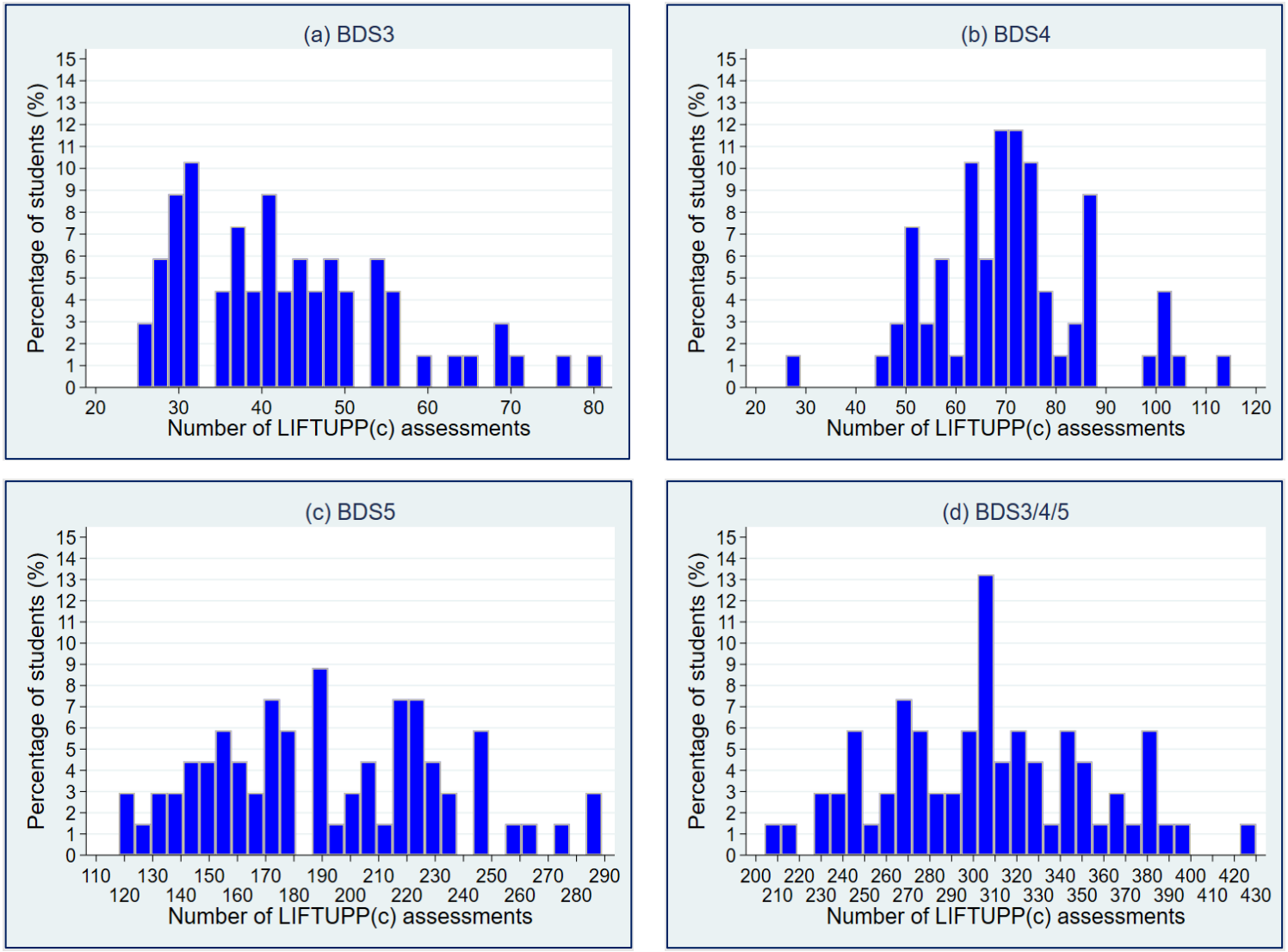
As can be seen in Figures 5.4 (a-c), the total number and range of LIFTUPP© assessments completed also increased through sequential BDS years. Although cohort 1 (collectively) completed fewer clinical assessments over the entire BDS course than cohort 2 (19,199 vs 20,312, respectively), both cohorts displayed similar means (240.0 and 236.2), minimums (151 and 157) and maximums (349 and 346) for the number of procedures per student. Cohort 3 completed the most assessments (20,817) even though it consisted of fewer students than cohorts 1 and 2. The mean (306.1), minimum (204) and maximum (430) number of assessments (per student) were also greater in cohort 3.



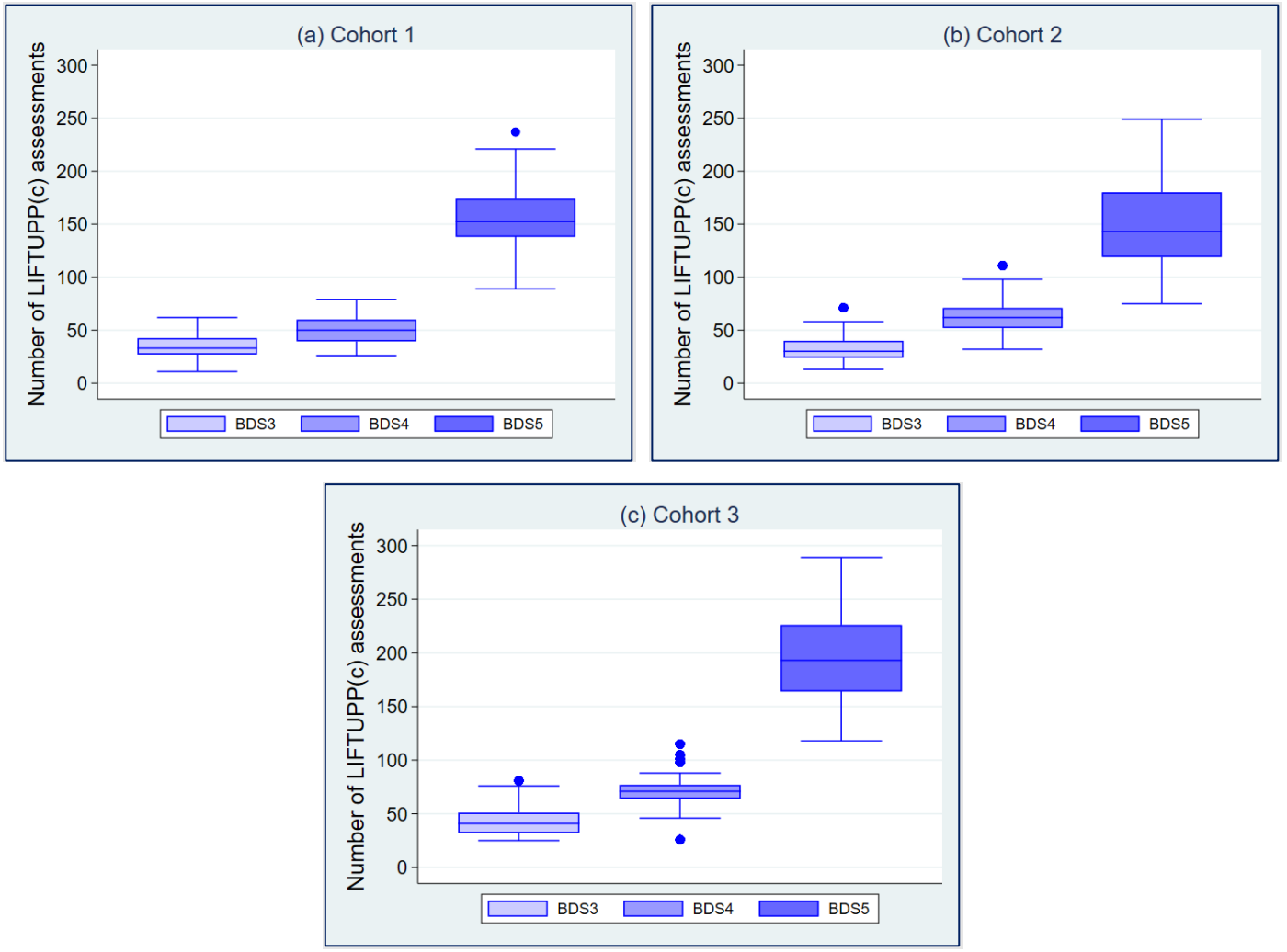
Figures 5.1 (a-d) - Cohort 1: Number of eligible clinical LIFTUPP© assessments completed per student.



Figures 5.2 (a-d) - Cohort 2: Number of eligible clinical LIFTUPP© assessments completed per student.



Figures 5.3 (a-d) - Cohort 3: Number of eligible clinical LIFTUPP© assessments completed per student.



Figures 5.4 (a-c) – Cohorts 1, 2 and 3: Boxplots for the number of eligible clinical assessments completed per student within each BDS academic year.

5.4.1.2 LIFTUPP© performance indicators – Student perspective

Frequencies of minimum LIFTUPP© performance indicators awarded per eligible procedure are displayed for each cohort in Tables 5.1-5.3. In all three cohorts, the most common minimum performance indicator awarded to students for clinical procedures performed in BDS3 and BDS4 was 4, which accounted for 43.7% and 46.1% (cohort 1), 49.4% and 47.0% (cohort 2) and 56.0% and 47.5% (cohort 3) of all minimum performance indicators in each of these year groups, respectively. A minimum performance indicator of 4 was also the most frequently awarded in BDS5 for cohort 3 (46.5%) and the most predominant over the entire BDS curriculum for both cohorts 2 (46.1%) and 3 (48.1%). In cohort 1, a 5 was the most prevalent minimum performance indicator in BDS5 and over the duration of the BDS course, accounting for 54.6% and 47.5%, respectively. A 5 was also the most frequent BDS5 minimum performance indicator for cohort 2 (46.1%).

In all three cohorts, minimum performance indicators of 1 or 2 were rarely awarded. A minimum performance indicator of 1 constituted less than 0.1% of all assessments undertaken both within and across BDS years (in all cohorts). The greatest prevalence of 2s was found in cohort 2 for BDS3 and accounted for >2.5% of all minimum performance indicators awarded throughout this academic year.

The proportion of 1s, 2s, 3s and 4s generally decreased over subsequent years in all three cohorts. However, some exceptions were observed in cohort 1 where no 1s were awarded in BDS2 and there was as increase in 4s from BDS3 to BDS4 followed by a decrease in BDS5. The proportion of 5s awarded increased over subsequent BDS years in all three cohorts (Tables 5.1-5.3; Figures 5.5 (a-c)).

Table 5.1 – Cohort 1: Frequencies of minimum LIFTUPP© performance indicators awarded within and across BDS academic years.

BDS year	Total LIFTUPP© assessments completed	Minimum LIFTUPP© performance indicators [n (%)]					
		1	***2***	***3***	***4***	***5***	***6***
3	2,756		35 (1.27)	425 (15.42)	**1,205 (43.72)**	880 (31.93)	208 (7.55)
4	4,044		54 (1.34)	443 (10.95)	**1,865 (46.12)**	1,478 (36.55)	204 (5.04)
5	12,399		69 (0.56)	680 (5.48)	4,055 (32.70)	**6,766 (54.57)**	828 (6.68)
All (BDS3-5)	19,199		158 (0.82)	1,548 (8.06)	7,125 (37.11)	**9,124 (47.52)**	1,240 (6.46)

Mode [minimum] performance indicators (within and across all BDS year groups) are in **bold.** Cells where n<5 have been greyed out.

Table 5.2 – Cohort 2: Frequencies of minimum LIFTUPP© performance indicators within and across BDS academic years.

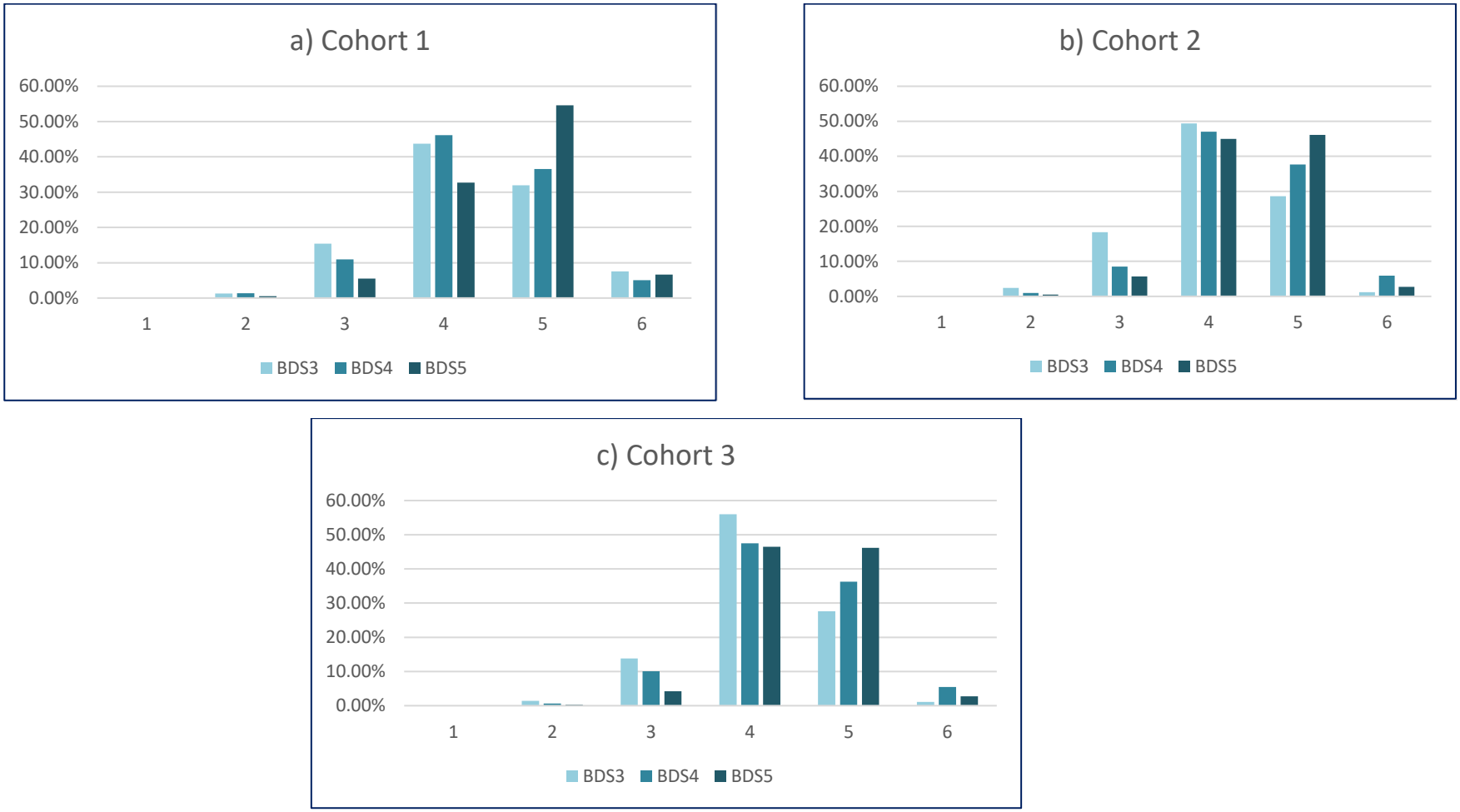
BDS year	Total LIFTUPP© assessments completed	Minimum LIFTUPP© performance indicators [n (%)]					
		1	***2***	***3***	***4***	***5***	***6***
3	2,746		66 (2.40)	505 (18.32)	**1,356 (49.38)**	786 (28.62)	33 (1.20)
4	5,226		51 (0.98)	443 (8.48)	**2,457 (47.02)**	1,966 (37.63)	307 (5.88)
5	12,343		67 (0.54)	700 (5.67)	5,548 (44.95)	**5,692 (46.12)**	332 (2.69)
All (BDS3-5)	20,312		184 (0.91)	1,646 (8.10)	**9,360 (46.08)**	8,444 (41.57)	672 (3.31)

Mode [minimum] performance indicators (within and across all BDS year groups) are in **bold**. Cells where n<5 have been greyed out.

Table 5.3 – Cohort 3: Frequencies of minimum LIFTUPP© performance indicators awarded within and across BDS academic years.

BDS year	Total LIFTUPP© assessments completed	Minimum LIFTUPP© performance in [n (%)]					
		1	***2***	***3***	***4***	***5***	***6***
3	2,938		42 (1.43)	406 (13.82)	**1,645 (55.99)**	811 (27.60)	32 (1.09)
4	4,806		30 (0.62)	484 (10.07)	**2,283 (47.50)**	1,744 (36.29)	263 (5.47)
5	13,073		43 (0.33)	552 (4.22)	**6,074 (46.46)**	6,041 (46.21)	362 (2.77)
All (BDS3-5)	20,817		115 (0.55)	1,442 (6.93)	**10,002 (48.05)**	8,596 (41.29)	657 (3.16)

Mode [minimum] performance indicators (within and across all BDS year groups) are in **bold.** Cells where n<5 have been greyed out.



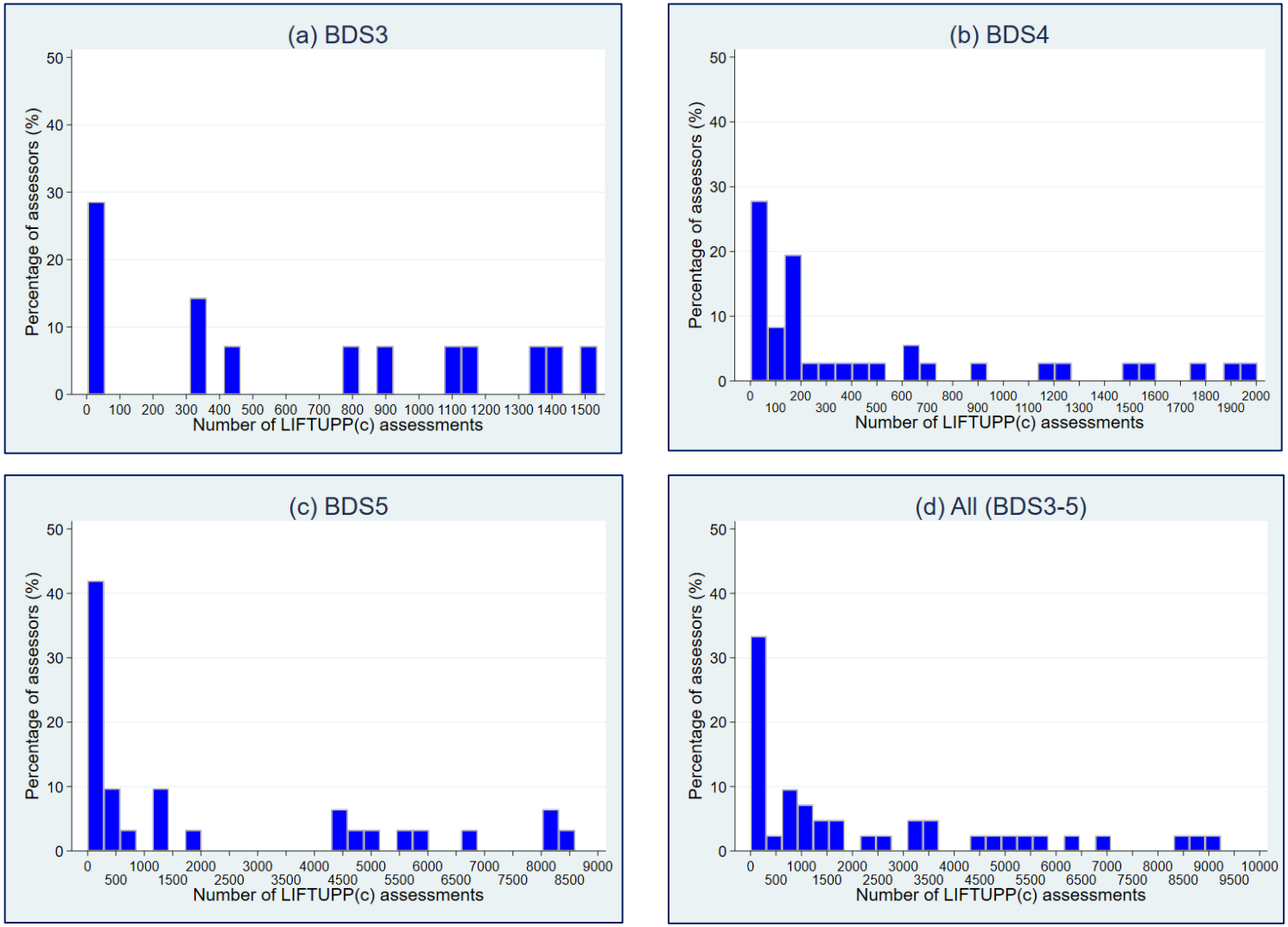
Figures 5.5 (a-c) – Cohorts 1, 2 and 3: Bar chart representations for proportions of minimum LIFTUPP© performance indicators awarded for eligible clinical procedures per student within each BDS academic year.

5.4.1.3 Number of LIFTUPP© assessments – Assessor perspective

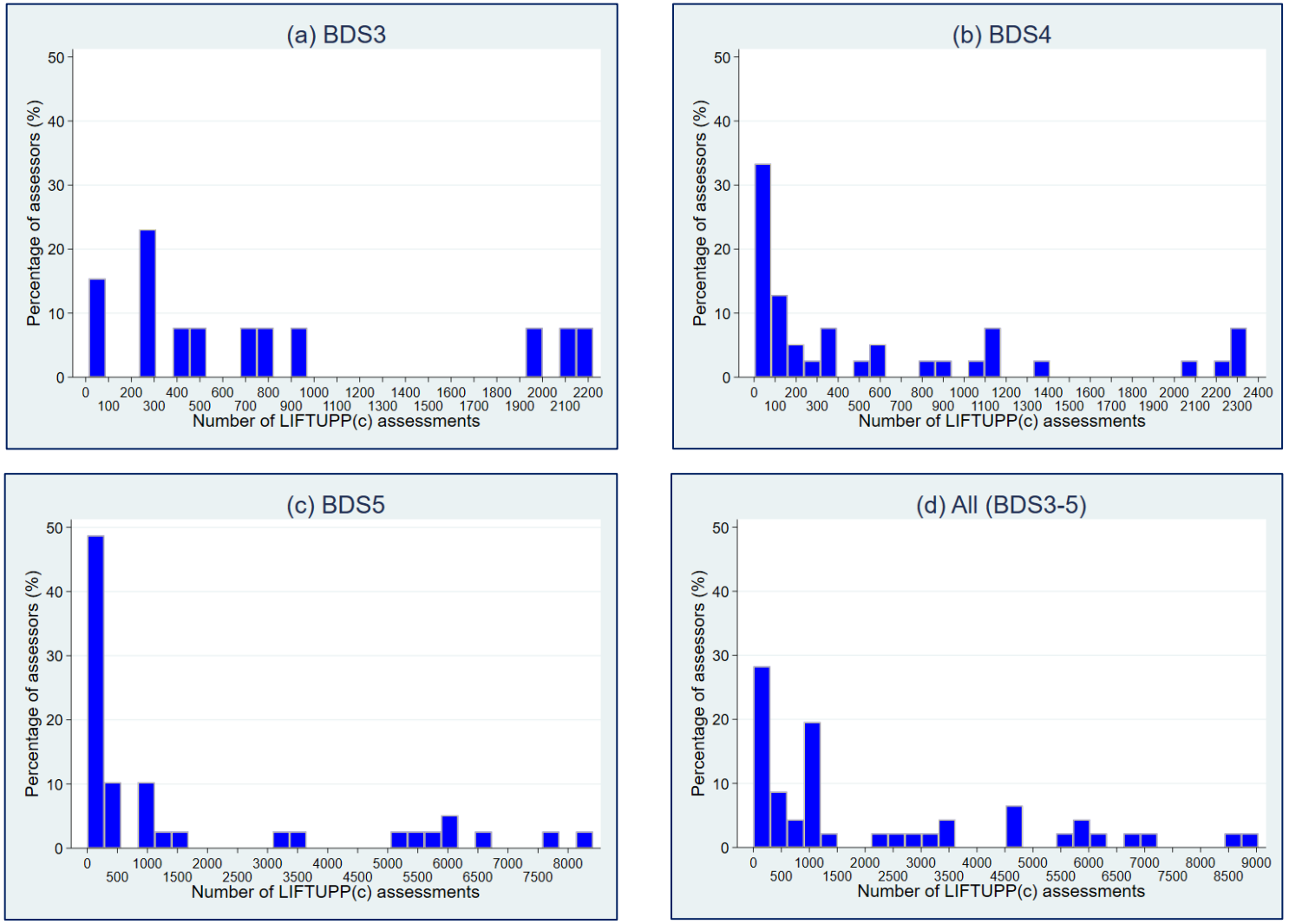
Histograms for the number of procedural stage assessments conducted by assessors within and across BDS years are shown in Figures 5.6–5.8 (a-d). A total of 42 different assessors assessed cohort 1 students across BDS3-5, with 14, 36 and 31 (of the 42) assessing in each of BDS3, BDS4 and BDS5, respectively. In cohort 2, 46 different assessors assessed across BDS3-5, with 13 assessing in BDS3 and 39 assessing in both BDS4 and BDS5. Finally, in cohort 3, a total of 49 different assessors assessed students across BDS3-5, and 15, 38, 39 assessed in BDS3, BDS4 and BDS5, respectively. Data are summarised for each cohort in Tables 5.4-5.6.

The mean number of clinical LIFTUPP© procedural stage assessments completed per assessor decreased from BDS3 to BDS4, before increasing in BDS5. In cohort 1, there was a decrease from 677.8 (BDS3) to 486.4 (BDS4) before an increase to 2263.5 (BDS5) (Table. 5.4). Cohort 2 saw a decrease from 813.5 (BDS3) to 597.3 (BDS4) followed by an increase to 1721.1 (BDS5) (Table 5.5). Finally, cohort 3 displayed a decrease from 681.9 (BDS3) to 540.4 (BDS4) following by an increase to 1687.1 (BDS5) (Table 5.6). Tables 5.4-5.6 also showed significant differences in assessor assessment activity/experience in all three cohorts. Some assessors conducted a much larger number of assessments than others, for example, in cohort 1, at least one assessor only contributed one assessment across the entire BDS course whereas another had performed over 9000 (see Table 5.4).

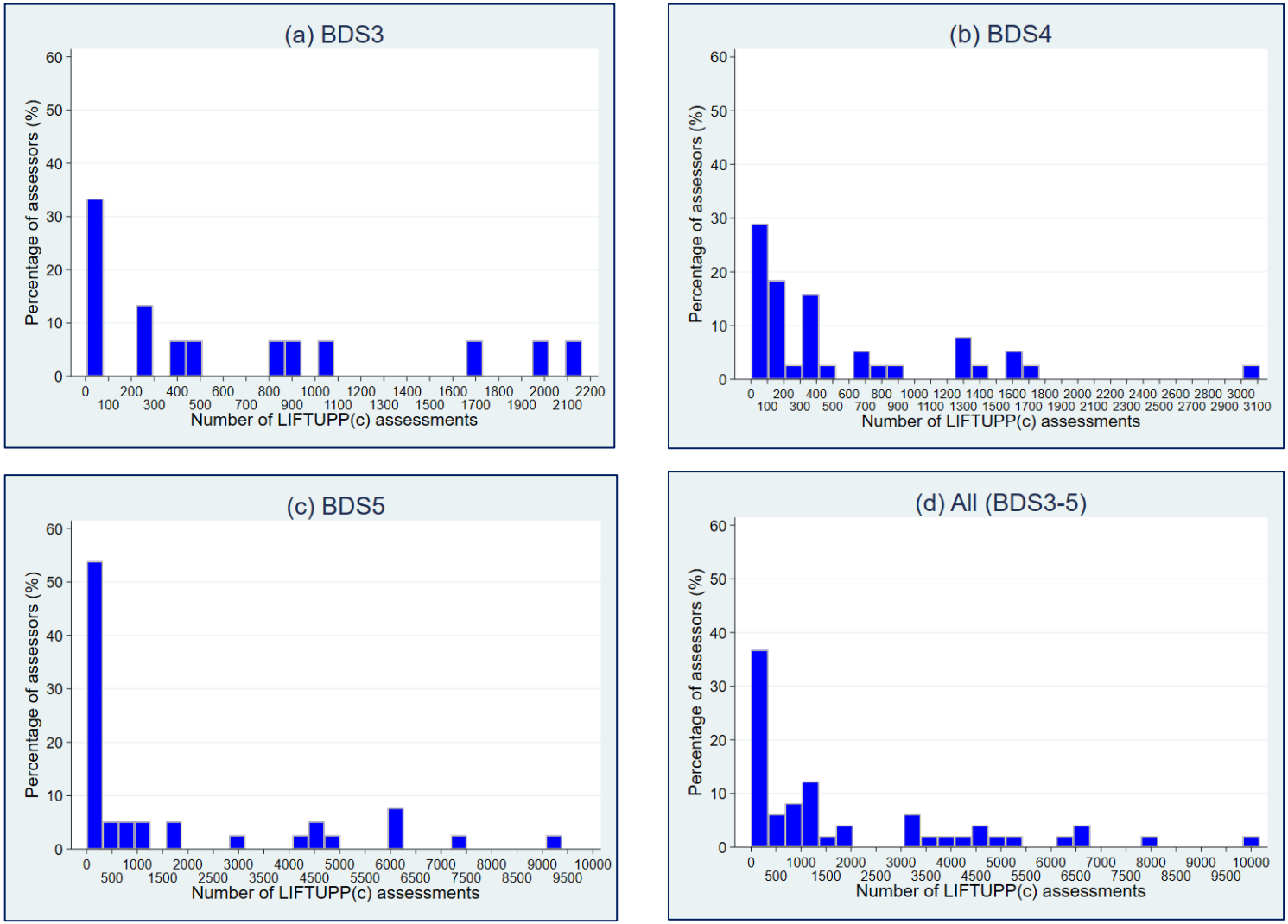
Considerable variability was also observed in the number of individual students assessed by assessors. Some assessors had only assessed a single student in some BDS years whereas others had assessed each student in the cohort. Examples of the former are seen in BDS3/4/5 for cohort 1 (Table 5.7), and BDS4/5 for both cohorts 2 (Table 5.8) and 3 (Table 5.9). Examples of latter can be observed in BDS5 for cohort 1 (Table 5.7), BDS4 for cohort 2 (Table 5.8), and BDS4/5 for cohort 3 (Table 5.9).



Figures 5.6 (a-d) - Cohort 1: Number of clinical LIFTUPP© procedural stage assessments completed per assessor.



Figures 5.7 (a-d) - Cohort 2: Number of clinical LIFTUPP© procedural stage assessments completed per assessor.



Figures 5.8 (a-d) - Cohort 3: Number of clinical LIFTUPP© procedural stage assessments completed per assessor.

Table 5.4 - Cohort 1: Summary statistics for the number of clinical LIFTUPP© procedural stage assessments completed per assessor within and across BDS academic years.

BDS year	*Number of assessors providing assessment*	*Mean*	*Standard deviation*	*Minimum*	*Q1*	*Median*	*Q3*	*Maximum*
3	14	677.79	573.47	4	51	613.50	1164	1536
4	36	486.39	606.54	2	38	188	676.50	2004
5	31	2263.45	2911.01	1	65	407	4803	8596
All	42	2313.48	2703.22	1	193	1088	3524	9235

Table 5.5 – Cohort 2: Summary statistics for the number of clinical LIFTUPP© procedural stage assessments completed per assessor within and across BDS academic years.

BDS year	*Number of assessors providing assessment*	*Mean*	*Standard deviation*	*Minimum*	*Q1*	*Median*	*Q3*	*Maximum*
3	13	813.46	783.64	14	260	515	962	2221
4	39	597.33	756.18	3	30	226	1086	2345
5	39	1721.05	2578.21	1	40	297	3170	8414
All	46	2208.76	2550.27	1	212	1125.50	3535	9032

Table 5.6 – Cohort 3: Summary statistics for the number of clinical LIFTUPP© procedural stage assessments completed per assessor within and across BDS academic years.

BDS year	*Number of assessors providing assessment*	*Mean*	*Standard deviation*	*Minimum*	*Q1*	*Median*	*Q3*	*Maximum*
3	15	681.87	746.16	6	28	427	1028	2162
4	38	540.42	678.50	1	99	282.50	788	3113
5	39	1687.13	2524.19	5	29	294	3024	9385
All	49	1986.20	2460.06	4	154	965	3311	10173

Table 5.7 – Cohort 1: Summary statistics for the number of individual students assessed per assessor within and across BDS academic years.

BDS year	*Number of assessors providing assessment*	*Mean*	*Standard deviation*	*Minimum*	*Q1*	*Median*	*Q3*	*Maximum*
3	14	27	16.41	1	10	30.50	37	52
4	36	26.78	25.35	1	6	16.50	50.50	77
5	31	28.77	27.73	1	9	19	46	80
All	42	37.81	28.31	1	13	28	72	80

Table 5.8 – Cohort 2: Summary statistics for the number of individual students assessed per assessor within and across BDS academic years.

BDS year	*Number of assessors providing assessment*	*Mean*	*Standard deviation*	*Minimum*	*Q1*	*Median*	*Q3*	*Maximum*
3	13	36.77	20.10	2	23	39	55	64
4	39	30.69	31.05	1	7	14	62	86
5	39	30.87	29.95	1	8	18	49	85
All	46	44.78	32.65	1	11	37	82	86

Table 5.9 – Cohort 3: Summary statistics for the number of individual students assessed per assessor within and across BDS academic years.

BDS year	*Number of assessors providing assessment*	*Mean*	*Standard deviation*	*Minimum*	*Q1*	*Median*	*Q3*	*Maximum*
3	15	29.60	21.40	2	9	26	50	61
4	38	26.42	25.28	1	6	13.50	54	68
5	39	24.13	23.29	1	7	15	37	68
All	49	34.37	25.05	2	13	25	65	68

5.4.1.4 LIFTUPP© performance indicators – Assessor perspective

Frequencies of LIFTUPP© performance indicators awarded by assessors are displayed for each cohort in Tables 5.10-5.12 and illustrated using bar charts in Figures 5.9 (a-c). NOTE: These are ***all*** LIFTUPP© performance indicators awarded by assessors for each stage of clinical procedure as opposed to minimum LIFTUPP© performance indicators for eligible procedures which were used to analyse the data from the student perspective (see section 5.4.1.2 above).

In all three cohorts, the most common performance indicator awarded by assessors for clinical procedures performed in BDS3 was 4, which made up for 45.2% (cohort 1), 47.8% (cohort 2) and 52.4% (cohort 3) of all BDS3 performance indicators. In cohort 1, a 5 was the most frequent performance indicator in BDS4 (47.2%), however, in cohorts 2 and 3, it was 4 (Cohort 2: 44.8% |Cohort 3: 45.2%). A 5 was the most prevalent performance indicator in BDS5 for all three cohorts, accounting for 64.4%, 58.6% and 56.9% of all BDS5 performance indicators in cohorts 1, 2 and 3, respectively. A 5 was also the most predominant performance indicator awarded over the entire BDS course in all three cohorts (Cohort 1: 58.7%|Cohort 2: 52.9%|Cohort 3: 51.7%).

Over subsequent student cohorts, the proportion of 5s and 6s awarded decreased and the proportion of 4s increased. Very little change in the proportion of 1s, 2s and 3s between all three cohorts was also noted (see Tables 5.10-5.12).

Table 5.10 – Cohort 1: Frequencies of LIFTUPP© performance indicators awarded for procedural stages by assessors within and across BDS academic years.

BDS year	Total LIFTUPP© assessments completed	LIFTUPP© performance indicators [n (%)]					
		1	***2***	***3***	***4***	***5***	***6***
3	9,489	5 (0.05)	96 (1.01)	964 (10.16)	**4,288 (45.19)**	3,605 (37.99)	531 (5.60)
4	17,510		137 (0.78)	1,047 (5.98)	6,948 (39.68)	**8,264 (47.20)**	1,114 (6.36)
5	70,167		133 (0.19)	1,537 (2.19)	16,966 (24.18)	**45,200 (64.42)**	6,329 (9.02)
All (BDS3-5)	97,166	7 (0.01)	366 (0.38)	3,548 (3.65)	28,202 (29.02)	**57,069 (58.73)**	7,974 (8.21)

Mode performance indicators (within and across all BDS year groups) are in **bold**. Cells where n<5 have been greyed out.

Table 5.11 – Cohort 2: Frequencies of LIFTUPP© performance indicators awarded for procedural stages by assessors within and across BDS academic years.

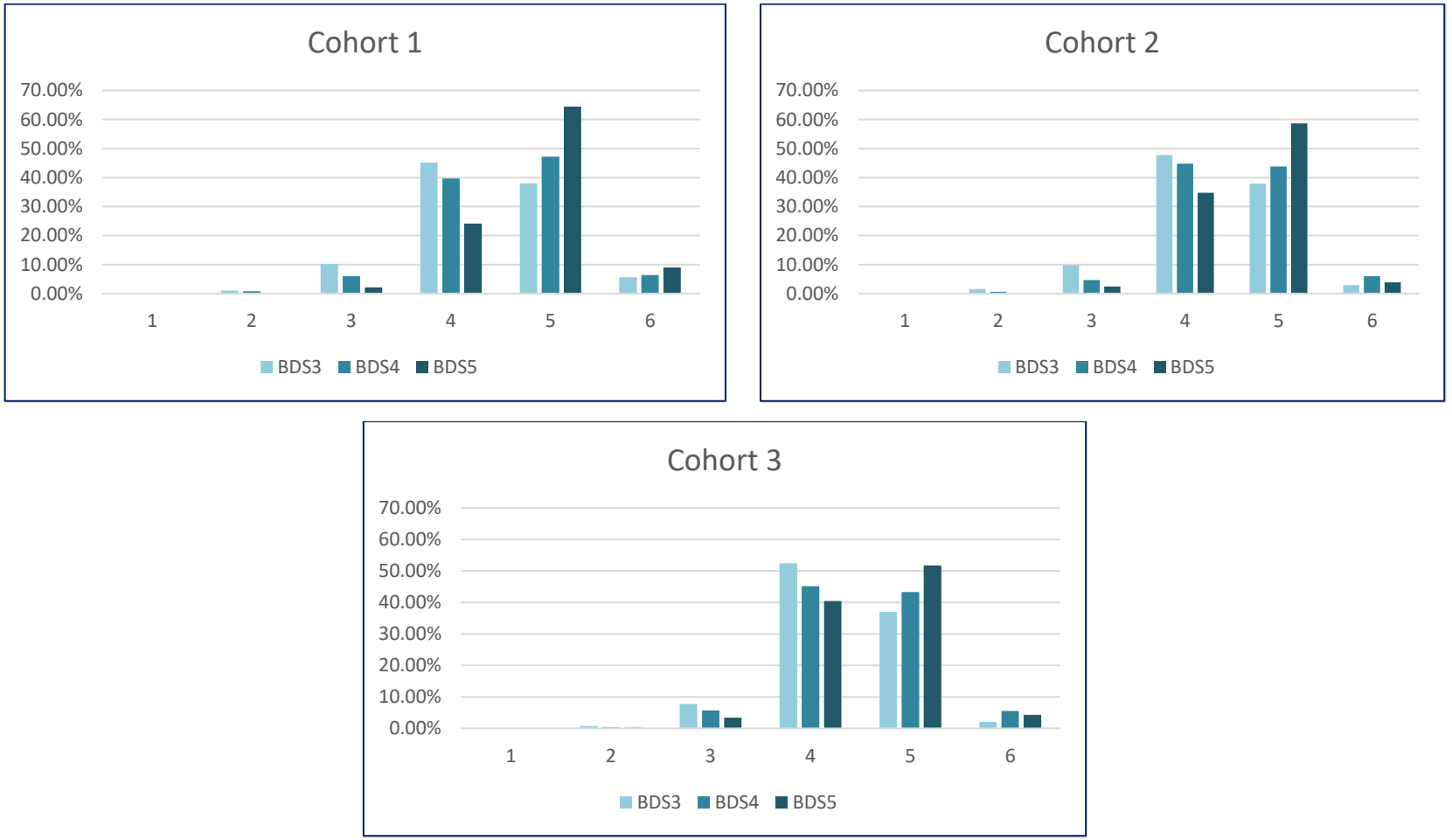
BDS year	Total LIFTUPP© assessments completed	LIFTUPP© performance indicators [n (%)]					
		1	***2***	***3***	***4***	***5***	***6***
3	10,575	5 (0.05)	163 (1.54)	1,030 (9.74)	**5,056 (47.81)**	4,008 (37.90)	313 (2.96)
4	23,296	8 (0.03)	149 (0.64)	1,092 (4.69)	**10,439 (44.81)**	10,212 (43.84)	1,396 (5.99)
5	67,121	11 (0.02)	148 (0.22)	1,654 (2.46)	23,332 (34.76)	**39,361 (58.64)**	2,615 (3.90)
All (BDS3-5)	101,603	31 (0.03)	490 (0.48)	3,873 (3.81)	39,161 (38.54)	**53,694 (52.85)**	4,354 (4.29)

Mode performance indicators (within and across all BDS year groups) are in **bold**.

Table 5.12 – Cohort 3: Frequencies of LIFTUPP© performance indicators awarded for procedural stages by assessors within and across BDS academic years.

BDS year	Total LIFTUPP© assessments completed	LIFTUPP© performance indicators [n (%)]					
		1	***2***	***3***	***4***	***5***	***6***
3	10,228		77 (0.75)	796 (7.78)	**5,360 (52.41)**	3,778 (36.94)	213 (2.08)
4	20,536		58 (0.28)	1,172 (5.71)	**9,272 (45.15)**	8,886 (43.27)	1,145 (5.58)
5	65,798		84 (0.13)	1,223 (1.86)	24,225 (36.82)	**37,461 (56.93)**	2,804 (4.26)
All (BDS3-5)	97,324	8 (0.01)	232 (0.24)	3,291 (3.38)	39,324 (40.41)	**50,303 (51.69)**	4,166 (4.28)

Mode performance indicators (within and across all BDS year groups) are in **bold**. Cells where n<5 have been greyed out.



Figures 5.9 (a-c) – Cohorts 1, 2 and 3: Bar chart representations for proportions of LIFTUPP© performance indicators awarded for procedural stage assessments by assessors within each BDS academic year.

5.4.2 Modelling LIFTUPP© data using group-based trajectory modelling

The following sections describe the GBTMs selected to represent minimum performance indicators from clinical LIFTUPP© data according to the data distribution (censored normal or Bernoulli) chosen and any imposed minimum number of 5, 10, 15 and 20 students per group. Details of the processes used to generate, evaluate, and select the GBTMs have previously been provided in chapter 3 (section 3.5.3.6).

Chapter 3 (section 3.5.3.6) also detailed the numerical codes that depict the shape of the trajectories within each GBTM were as follows:

0 = zero order (straight line)

1 = linear

2 = quadratic (one turning point)

3 = cubic (two turning points)

The combination of numerical codes equates to the number of group trajectories within each model. For example, a 1 1 1 1 model is comprised of four groups following linear trajectories, whereas a 3 3 2 model would show a three-group model with two cubic trajectories and one quadratic trajectory.

As per section 5.4.1.1 (above), the number of LIFTUPP© assessments eligible for inclusion in the study for cohorts 1, 2 and 3 were 19,999, 20,312 and 20,817, respectively. Therefore, GBTMs based on censored normal and Bernoulli data distributions were generated from these number of assessments spread across 80 (cohort 1), 86 (cohort 2) and 68 (cohort 3) students.

For both data distributions, the standard error for the coefficient could not be calculated in two-, three- and four-group models which had an insufficient number of students in at least one group as the variance matrix was non-symmetric or highly singular. Some models were returned following false

convergence. These occurrences were highlighted by the Stata software traj plugin.

Variance matrix errors relate to the variance/covariance matrix of coefficient estimates. Trajectory groups which display no variance (e.g., a group with only one outlier row and/or a constant value across all time periods) will have a near singular covariance matrix and, therefore, cannot be inverted. As a result, the parameter estimates of the group cannot be calculated. This error will eventually occur through increasing the number of groups in GBTM. For example, if there was a cohort of 100 dental students, and the traj plugin was asked to estimate 100 group trajectories, a variance matrix error is guaranteed to be returned.

False convergence errors signify too many groups have been estimated or the specific combination of trajectory shapes just does not fit the data for the given the number of time periods. Models which returned either of these statistical errors were discarded even if a BIC was given as part of the traj statistical output.

The “average” performance indictors and threshold performance indicators used to generate censored normal and Bernoulli GBTMs, respectively, are based on the minimum performance indicators awarded to students per clinical procedure (see chapter 3, section 3.5.3.4).

A “best”/“better” performing trajectory group was sought within each GBTM. Various factors were taken into consideration when judging which trajectory groups were considered “better” performing. These included:

- The duration over which a group achieved higher group average (minimal) LIFTUPP© performance indicators (censored normal models)/probability of being awarded the threshold performance indicator (Bernoulli models).

- The rate of positive change.

- A higher trajectory end point.

If these criteria didn't provide a clear objective decision, a subjective judgement was made to select a "better" performing group for comparisons with other assessment outcomes (undergraduate examinations and LEPs) (see chapter 7).

5.4.2.1 Censored normal data distributions

Model generation

For each cohort, simulations for all potential trajectory shape combinations were run for up to maximum of four trajectory groups per model using the traj Plugin within Stata® 15 statistical software. In total, 340 were generated for each cohort which included four single trajectory group models, 16 two-group models, 64 three-group models, and 256 four-group models.

Model evaluation and selection

Out of the 340 models generated per cohort, 100, 18 and 17 models were returned with no statistical errors for cohort 1, 2 and 3, respectively. Appendix 9 provides full lists of censored normal GBTMs returned with no statistical errors for each cohort and ranks them by BIC from highest (i.e., least negative) to lowest (i.e., most negative).

From the error-free subsets - and following the criteria described in chapter 3 (section 3.5.3.6) - three two-group GBTMs were selected for cohort 1 and one two-group GBTM was each selected for cohorts 2 and 3 (see Table 5.13).

The GBTMs selected for cohorts 2 and 3 (models 1 3 and 3 2, respectively) had the highest BICs out of all the returned models and displayed a BIC difference of at least two compared to the model with the second highest BIC. Both models also satisfied the minimum group number restrictions of at least 5, 10, 15 and 20 students. However, in cohort 1, the GBTM with the highest BIC (model 1 3 2 3) did not contain at least 20 students in all its groups. Model 2 1 0 was the GBTM with the highest BIC to meet this respective criterion and was, therefore, also selected for further analysis and comparison with undergraduate examination data and postgraduate performance (see chapter 8). This model also displayed a

BIC difference of at least two compared to its nearest counterparts (i.e., those GBTMs which contained at least 20 students in its smallest group).

The GBTMs selected - and listed in Table 5.13 - performed well on all tests of model adequacy laid out in Nagin (2005) and Nagin and Odgers (2010). Table 5.14 shows the AvePP was at least 0.88 for each group in all models, i.e., greater than the 0.70 minimum value recommended by Nagin and Odgers (2010). The odds of correct classification for all groups in each GBTM were significantly greater than 5.0, indicating each model's assignment accuracy was good. There was also close correspondence between each trajectory group's estimated probability of group membership and the proportion of students classified to that group according to posterior probability of group membership across all models.

Table 5.13 - Censored normal distribution group-based trajectory models (GBTMs) selected to represent clinical LIFTUPP© data for cohorts 1, 2 and 3.

Cohort	Number of students (n)	Number of assessments	Model	Contains at least X students per group, where X =				BIC (based on number of students)
				5	*10*	*15*	*20*	
1	80	19,199	1 3 2 3	✓	✓	✓	✗	-22537.16
			2 1 0	✓	✓	✓	✓	-22822.41
2	86	20,312	1 3	✓	✓	✓	✓	-22449.02
3	68	20,817	3 2	✓	✓	✓	✓	-22062.14

Table 5.14 - Average posterior probabilities, odds of correct classification, estimated probability and proportion of group membership (as per posterior probability of group membership) for censored normal data distribution group-based trajectory models (GBTMs) selected to represent clinical LIFTUPP© data in cohorts 1, 2 and 3. All values are rounded to two decimal places.

Cohort (n)	Model	BIC	Group	n per group	Average posterior probability	Odds of correct classification	Estimated probability of group membership	Proportion of students classified to group according to posterior probability of group membership
1 (80)	1 3 2 3	-22537.16	1-1	10	.99	1104.38	.13	.13
			1-2	26	.97	68.48	.33	.33
			1-3	28	.95	35.31	.34	.35
			1-4	16	.99	764.45	.21	.20
	2 1 0	-22822.41	1-1	24	1	470.81	.31	.30
			1-2	34	.97	42.87	.43	.43
			1-3	22	.93	39.09	.26	.28
2 (86)	1 3	-22449.02	2-1	16	.99	318.93	.20	.19
			2-2	70	.98	13.48	.80	.81
3 (68)	3 2	-22062.14	3-1	48	1	121.89	.71	.71
			3-2	20	.99	230.41	.29	.29

Selected model trajectories (censored normal)

Cohort 1

The trajectories for cohort 1's 1 3 2 3 model are illustrated in Figure 5.10. This GBTM was the most suitable model when each group was required to have a minimum of 5, 10 and 15 students.

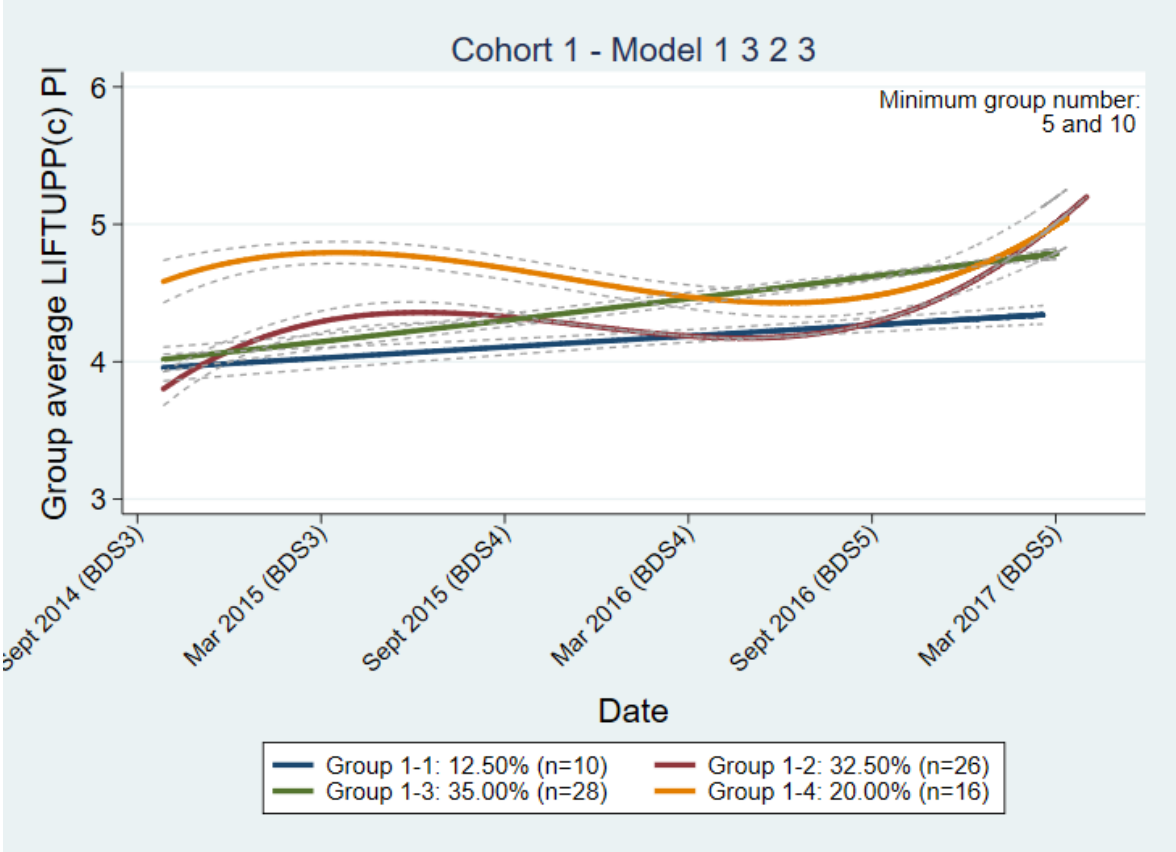


Figure 5.10 – Cohort 1: Trajectory groups for censored normal distribution model 1 3 2 3. NOTES: Individual data points have been removed to ensure trajectory lines are visible. 95% confidence intervals around estimated group probabilities are depicted by the dotted lines around each trajectory.

Students most likely to be members of groups 1-2 (12.5%; n = 10) and 1-3 (35.0%; n = 28) had a similar average starting [minimum] LIFTUPP© performance indicator at the beginning of BDS3 (~4) and both increased over time. However, by the end of BDS5, the minimum performance indicators awarded to those most likely to follow the trajectory of group 1-3 averaged at 4.8, compared to 4.3 (group 1-1).

Trajectories for group 1-2 (32.5%; n = 26) and 1-4 (20.0%; n = 16) were cubic with very similar shapes over the three BDS years. Students more likely to belong

to group 1-2 had, on average, a lower starting [minimum] performance indicator but appeared to "catch up" with group 1-4 students by the end of BDS5 (~5.3). From the trajectory patterns presented, it was difficult to determine which group was the "better performing" using the criteria previously described in section 5.4.2.

Narrow 95% confidence intervals were observed around the estimated group probabilities for each trajectory group. This satisfies another of Nagin (2005) and Nagin and Odgers (2010) criteria for appropriate GBTM selection.

Trajectory groups for cohort 1's 2 1 0 model are presented in Figure 5.11. This was the most suitable model when each trajectory group was restricted to a minimum of 20 students.

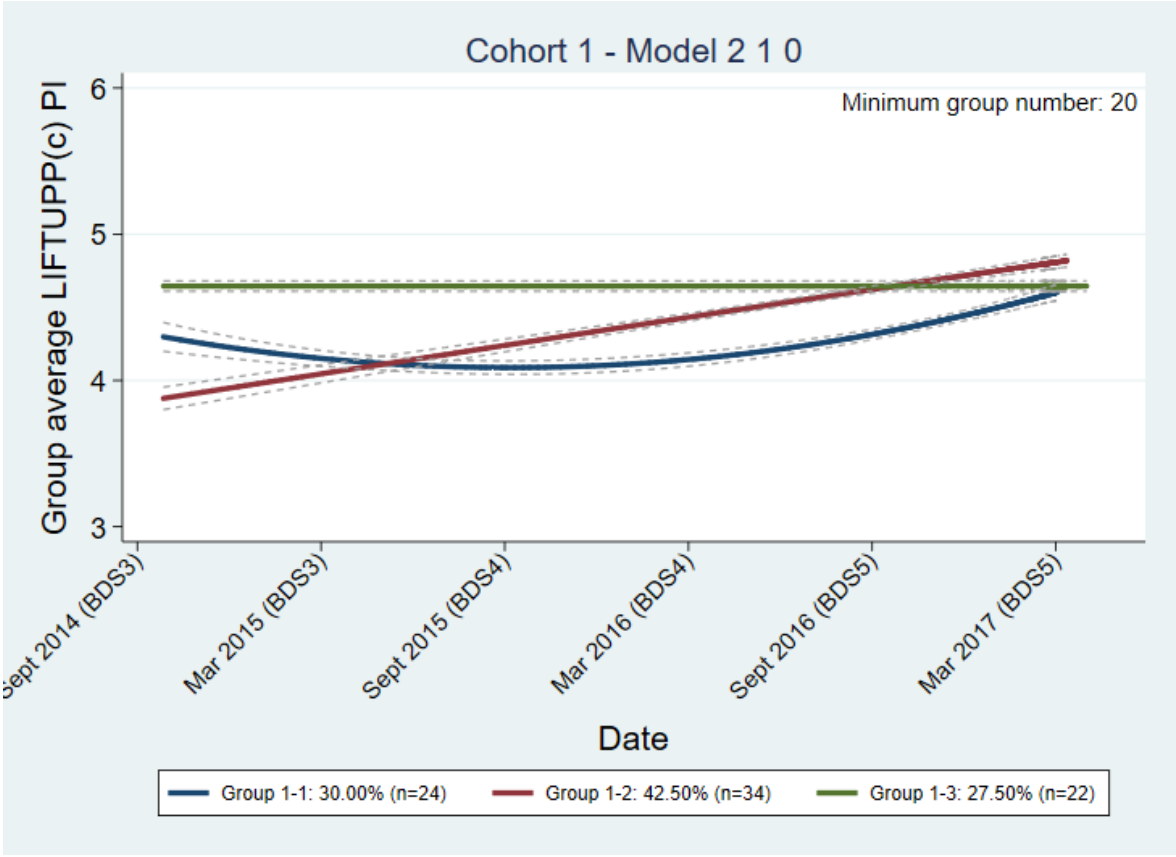


Figure 5.11 - Cohort 1: Trajectory groups for censored normal distribution model 2 1 0. NOTES: Individual data points have been removed to ensure trajectory lines are visible and 95% confidence intervals around estimated group probabilities are depicted by the dotted lines around each trajectory.

Trajectories generated for groups 1-1 (30.0%; n = 24) and 1-3 (27.5%; n = 22) indicated there was very little or no improvement in clinical performance over the BDS course. Group 1-1 followed a quadratic trajectory whereby the average

[minimum] LIFTUPP© performance indicator initially decreased from 4.3 to 4.1 over the course of BDS3 and BDS4 before gradually increasing up to 4.7 by the end of BDS5. Group 1-3 was a straight line through the y-axis where the average [minimum] LIFTUPP© performance indicator remained constant at 4.7. Alternatively, the linear trajectory for group 1-2 (42.5%; n = 34) showed a gradual improvement from 3.9 to 4.8 over the BDS curriculum. Overall, group 1-3 were identified as the "better" performing group over the duration of the BDS course.

The 2 1 0 model also had tight confidence intervals around the estimated group probabilities for each trajectory group.

Cohort 2

The trajectories for the 1 3 model, which was selected as the best fitting model for cohort 2, are shown in Figure 5.12. Unlike cohort 1, the model with the highest BIC was suitable for any group size larger than five students.

Group 2-1 (linear) (18.6%; n = 16) and group 2-2 (cubic) (81.4%; n = 70) both demonstrated an increase in their average [minimum] LIFTUPP© performance indicator by the end of the BDS course. However, group 2-2 displayed a greater overall increase in performance than group 2-1, as their average [minimum] LIFTUPP© performance indicator rose from 3.7 at the beginning of BDS3 to 4.8 by the end of BDS5. Group 2's clinical assessment data followed a cubic trajectory, with an increase in average [minimum] performance indicator over BDS3, a decrease over BDS4, followed by an additional increase in BDS5. Whereas group 2-1's linear trajectory illustrated a gradual increase in average [minimum] LIFTUPP© performance indicator from 3.9 at the beginning of BDS3 to 4.3 at the end of BDS5. Overall, group 2-2 were the "better performing" group.

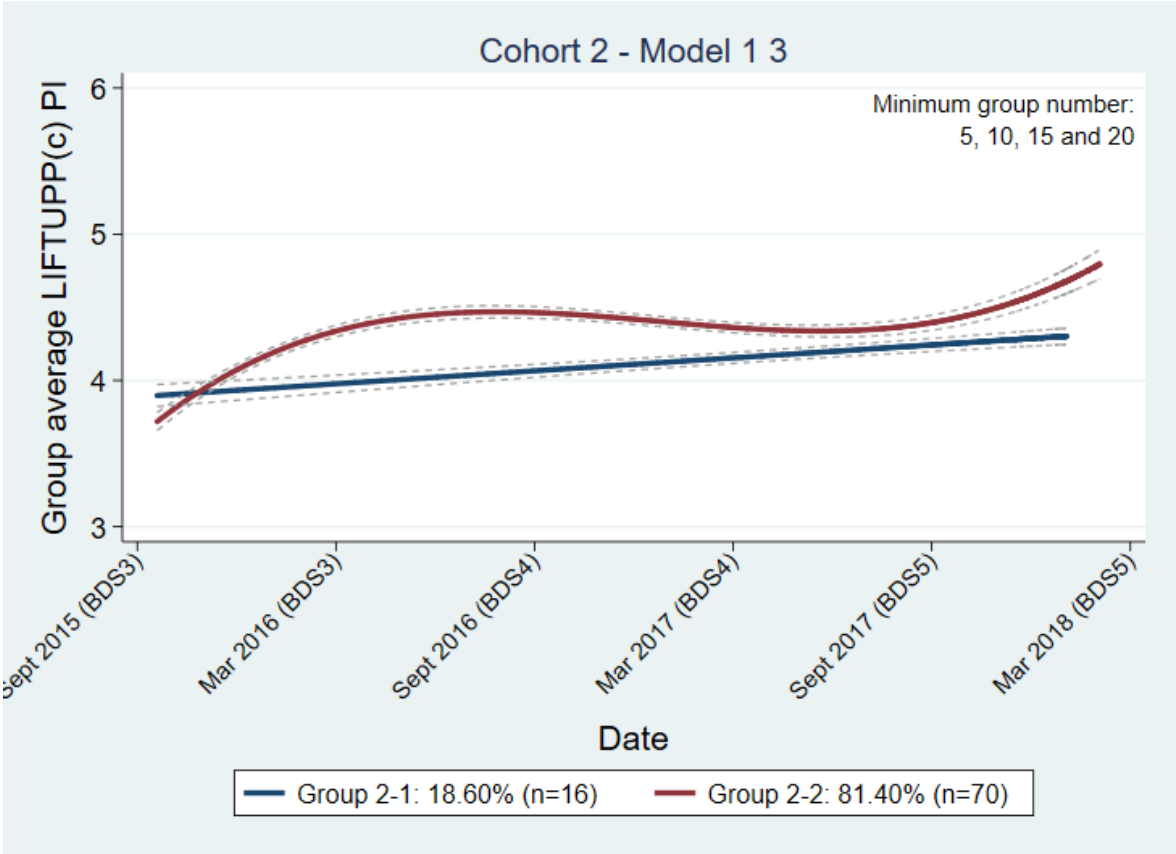


Figure 5.122 - Cohort 2: Trajectory groups for censored normal distribution model 1 3. NOTES: Individual data points have been removed to ensure trajectory lines are visible and 95% confidence intervals around estimated group probabilities are depicted by the dotted lines around each trajectory.

Tight confidence intervals are seen around the estimated group probabilities for each of group 1 and 2's trajectories.

Cohort 3

Figure 5.13 presents the trajectories for cohort 3's 3 2 model. Like cohort 2, the model with the highest BIC was the best model for any group size larger than five students.

Once again, both trajectory groups indicated an increase in clinical performance over the duration of the BDS course. Group 3-1 (cubic) (70.6%; n = 48) produced a cubic shaped trajectory where there was an initial increase in their average [minimum] LIFTUPP© performance indicator from 3.7 to 4.4 in BDS3, followed by decrease to 4.2 in BDS4 and a final increase to 4.5 during BDS5. Group 3-2 (29.4%; n = 20) generated a quadratic shaped trajectory which showed a gradual increase in average [minimum] LIFTUPP© performance indicator from 4.3 to 4.6. Therefore, despite differences in their trajectory shapes and starting points,

there was very little difference between each group's average [minimum] LIFTUPP© performance indicator by the end of BDS5. Group 3-2 were considered to be the "better performing group" overall.

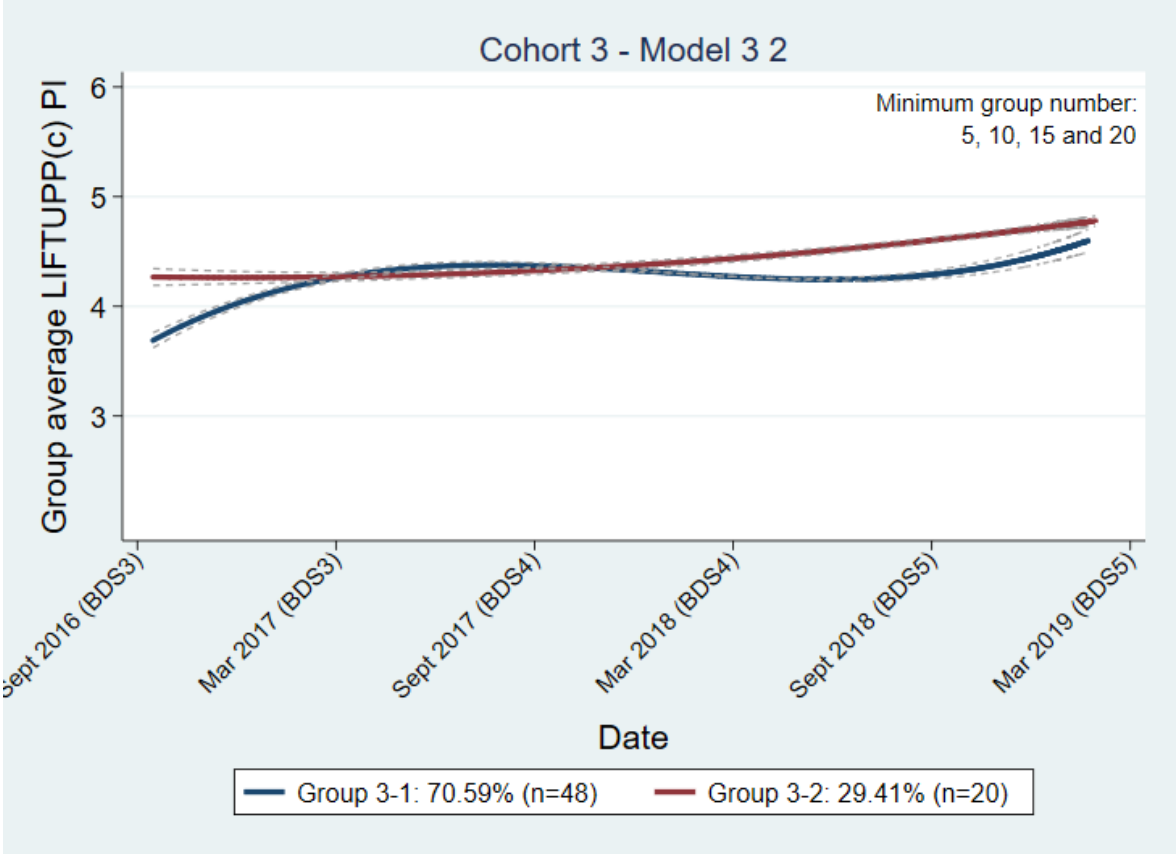


Figure 5.133 - Cohort 3: Trajectory groups for censored normal distribution model 3 2. NOTES: Individual data points have been removed to ensure trajectory lines are visible and 95% confidence intervals around estimated group probabilities are depicted by the dotted lines around each trajectory.

Tight confidence intervals around the estimated group probabilities for each trajectory group were also evident in this model.

5.4.2.2 Bernoulli data distributions (Threshold performance indicator = 4)

The following section describes the simulations for all potential trajectory shape combinations - which were run for up to maximum of four trajectory groups per model for all three cohorts. Therefore, 340 models were generated for each cohort (four single trajectory group models, 16 two-group models, 64 three-group models, and 256 four-group models).

Model evaluation and selection

Out of the 340 models generated per cohort, only four single-group trajectory models were returned with no statistical errors in all three cohorts. These models were subsequently ranked from least negative to most negative BIC.

Single group trajectories cannot have their model adequacy checked statistically (in accordance with recommendations by Nagin (2005) and Nagin and Odgers (2010) – see chapter 3, section 3.5.3.6). However, they could still explain student progress within and across each BDS year, and, therefore, GBTMs with threshold performance indicator of 4 were used to assess content validity.

To investigate criterion validity, there needs to be distinction between different performing student groups, but a threshold performance indicator of 4 did not provide this discrimination. This suggests a higher threshold performance indicator was needed to facilitate investigations on criterion validity (see section 5.4.2.3 for further details).

The single-group trajectory models with the least negative BICs were selected for presentation within this thesis (Figures 5.14-5.16). All trajectories were cubic and showed an increase in probability of achieving LIFTUPP© performance indicator ≥4 between BDS3 and BDS5. By the end of BDS5, all students in each cohort had a 0.95 probability of being awarded a (minimum) performance indicator ≥4.

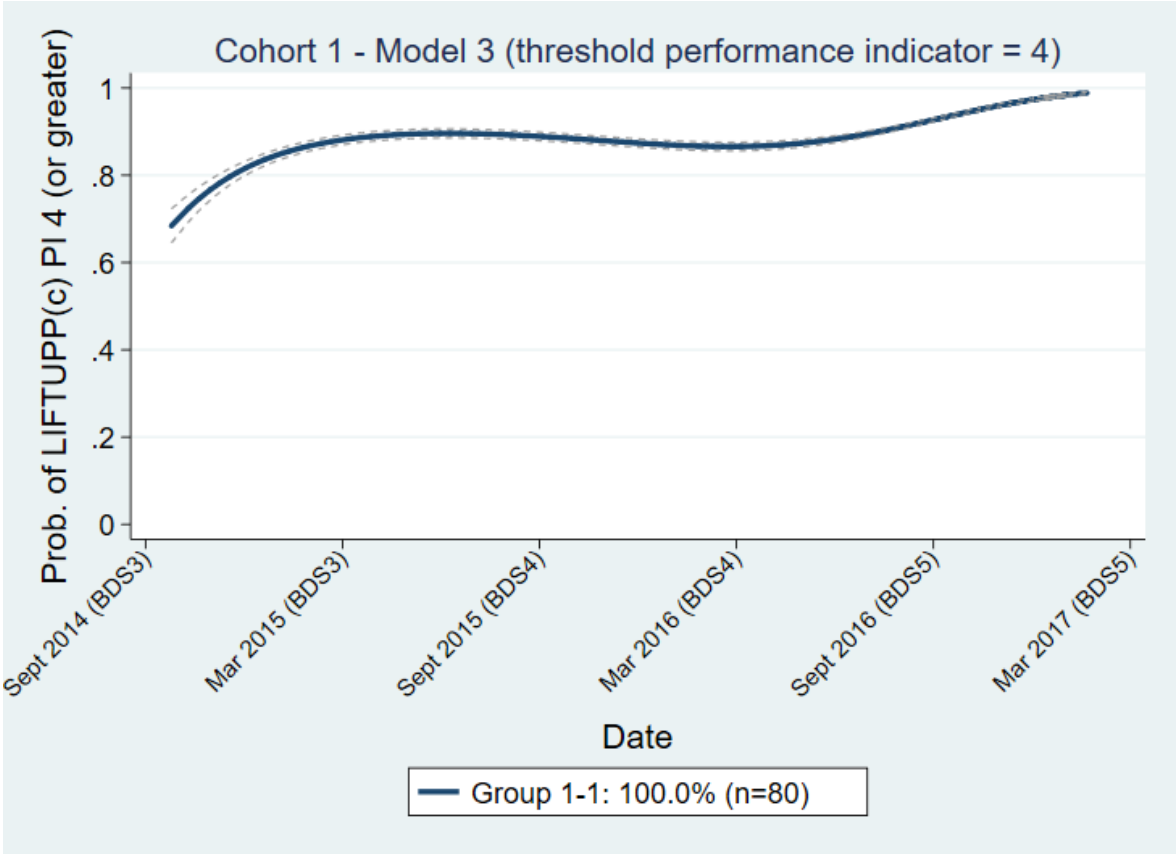


Figure 5.144 – Cohort 1: The trajectory for the clinical data if a threshold LIFTUPP© performance indicator (PI) of 4 was used in a Bernoulli data distribution.

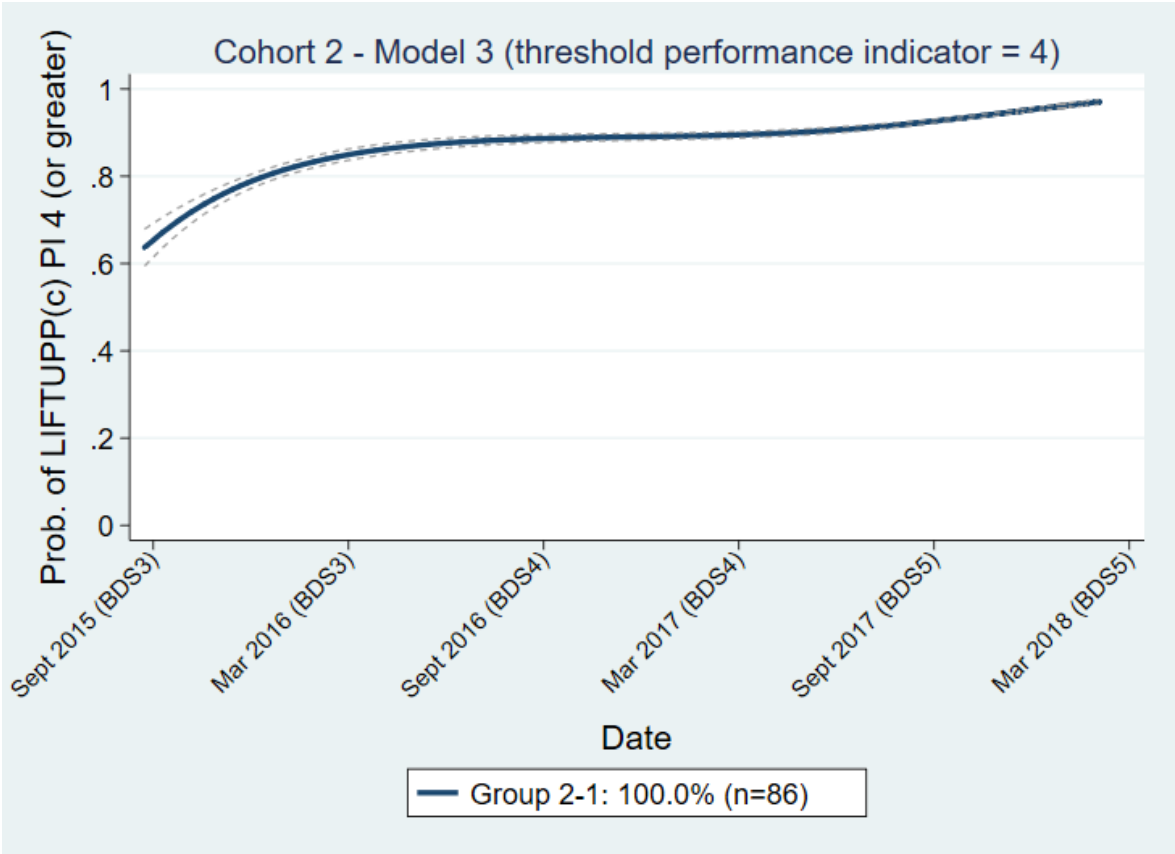


Figure 5.15 – Cohort 2: The trajectory for the clinical data if a threshold LIFTUPP© performance indicator (PI) of 4 was used in a Bernoulli data distribution.

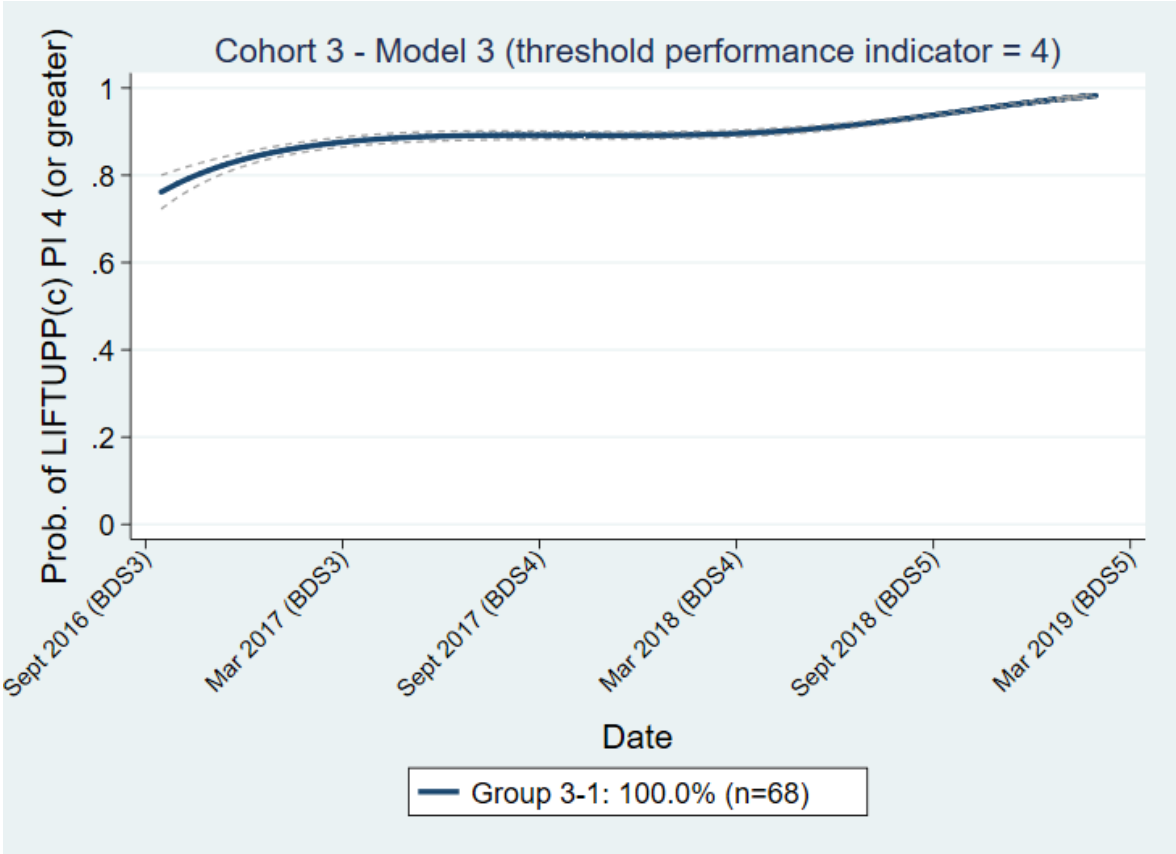


Figure 5.166 – Cohort 3: The trajectory for clinical data if a threshold LIFTUPP© performance indicator (PI) of 4 was used in a Bernoulli data distribution.

Narrow confidence intervals were observed around the estimated group probabilities for the trajectory in each cohort (dashed lines around trajectories in Figures 5.14-5.16).

5.4.2.3 Bernoulli data distributions (Threshold performance indicator = 5)

Model generation

Like GBTMs with a threshold performance indicator of 4, 340 were generated for each cohort (four single trajectory group models, 16 two-group models, 64 three-group models, and 256 four-group models).

Model evaluation and selection

Out of the 340 models generated per cohort, only 17, 19 and 15 models were returned which no statistical errors for cohorts 1, 2 and 3, respectively. These models were subsequently ranked from least negative to most negative BIC.

In contrast to Bernoulli GBTMs which used a threshold performance indicator of 4 (see section 5.4.2.2), a threshold performance indicator of 5 allowed multiple different student clinical development patterns to be distinguished. This will allow comparisons with other assessment outcomes to be made to investigate criterion validity (see chapter 7). However, the maximum number of groups identified within the LIFTUPP© data was only two, as no three- or four-group models were found in all three cohorts.

Using the model selection process described in chapter 3 (section 3.5.3.6), two two-group GBTMs were selected for cohort 1 and one two-group model was each selected for cohorts 2 and 3. Table 5.15 lists each of these models, all of which returned the highest BIC (determined by the total number of student participants) for their respective cohorts in accordance with any minimum group membership restrictions (~ 5, 10, 15 and 20 students). Appendix 9 provides full lists of Bernoulli GBTMs returned with no statistical errors for each cohort and ranks them by BIC (based on the number of participants) from least negative (highest) to most negative (lowest).

Table 5.15 - Bernoulli distribution group-based trajectory models (GBTMs) selected to represent clinical LIFTUPP© data for cohorts 1, 2 and 3.

Cohort	Number of students (n)	Number of assessments	Model	Contains at least X students per group, where X =				BIC (based on number of students)
				5	*10*	*15*	*20*	
1	80	19,199	3 2	✓	✓	✓	✗	-12432.70
			0 1	✓	✓	✓	✓	-12684.83
2	86	20,312	1 3	✓	✓	✓	✓	-13441.60
3	68	20,817	3 2	✓	✓	✓	✓	-13592.67

The first model selected for cohort 1 (model 3 2) had the highest BIC and displayed a BIC difference greater than two compared to the model with the second highest. However, it did not contain at least 20 students in all trajectory groups. Only one other multiple group model was returned with no statistical errors and contained at least 20 students in its smallest group – model 0 1 (appendix 9).

In cohort 2, the 1 3 model displayed the highest BICs out of all models returned without statistical errors and satisfied the minimum group number restrictions of at least 5, 10, 15 and 20 students. The difference in BIC between model 1 3 and the GBTMs with the second and third highest BICs (models 3 1 and 3 2, respectively) was less than two (appendix 9). Therefore, further scrutiny was undertaken to identify the most parsimonious model. Upon comparison, the graphical trajectories of these three models appeared very similar and would not differ significantly when used to describe the development of clinical performance for cohort 2 students (appendix 9). Furthermore, both the model parameters produced by the traj plugin (during generation of the GBTMs) and the results of the additional model adequacy testing (proposed by Nagin (2005) and Nagin and Odgers (2010)) were similar for models 1 3 and 3 1. This was because both models contained the same two trajectory patterns – the only difference was the order in which the traj plugin identified and presented them (appendix 9). Therefore, either one of models 1 3, 3 1 or 3 2 could have been selected to represent cohort 2's clinical data. In this case, model 1 3 was chosen by the researcher since it returned the highest BIC.

In cohort 3, model 3 2 returned the highest BIC out of all the models with no statistical errors and satisfied the minimum group number restrictions of at least 5, 10, 15 and 20 students. However, model 3 2's BIC was not greater than two when compared to model 1 3, which had the second highest BIC. On further investigation, there was negligible difference in the graphical appearance of both GBTMs (appendix 9), so choosing one model over the other would not result in the clinical performance of the student's being described differently. Therefore, model 3 2 was accepted as the most parsimonious model and selected to represent cohort 3's clinical LIFTUPP© data.

All GBTMs selected for each cohort were satisfactory based on the results of additional statistical testing for model adequacy (Nagin, 2005; Nagin and Odgers, 2010). The AvePP was 0.97 or higher for each group, i.e., greater than the 0.70 minimum value recommended (Nagin, 2005; Nagin and Odgers, 2010). The odds of correct classification for both groups were over 5.0, indicating the model's assignment accuracy was good, and there was close correspondence between each trajectory group's estimated probability of group membership and the

proportion of students classified to that group according to the posterior probability of group membership (see Table 5.16).

Table 5.16 - Average posterior probabilities, odds of correct classification, estimated probability, and proportion of group membership (as per posterior probability of group membership) for Bernoulli data distribution group-based trajectory models (GBTMs) selected to represent clinical LIFTUPP© data in cohorts 1, 2 and 3. NOTE: All values are rounded to two decimal places.

Cohort (n)	Model	BIC	Group	n per group	Average posterior probability	Odds of correct classification	Estimated probability of group membership	Proportion of students classified to group according to posterior probability of group membership
1 (80)	3 2	-12432.70	1-1	63	1	111.75	.79	.79
			1-2	17	.98	156.22	.21	.21
	0 1	-12684.83	1-1	21	.97	77.36	.26	.26
			1-2	59	.98	23.28	.74	.74
2 (86)	1 3	-13441.60	2-1	21	.99	211.17	.25	.24
			2-2	65	.99	23.85	.75	.76
3 (68)	3 2	-13592.67	3-1	43	.99	58.05	.63	.63
			3-2	25	.99	162.17	.37	.37

Selected model trajectories (Threshold performance indicator = 5)

Cohort 1

Figure 5.17 presents the group trajectories for model 3 2. This GBTM was the most suitable model when each group was required to have a minimum of 5, 10 and 15 students.

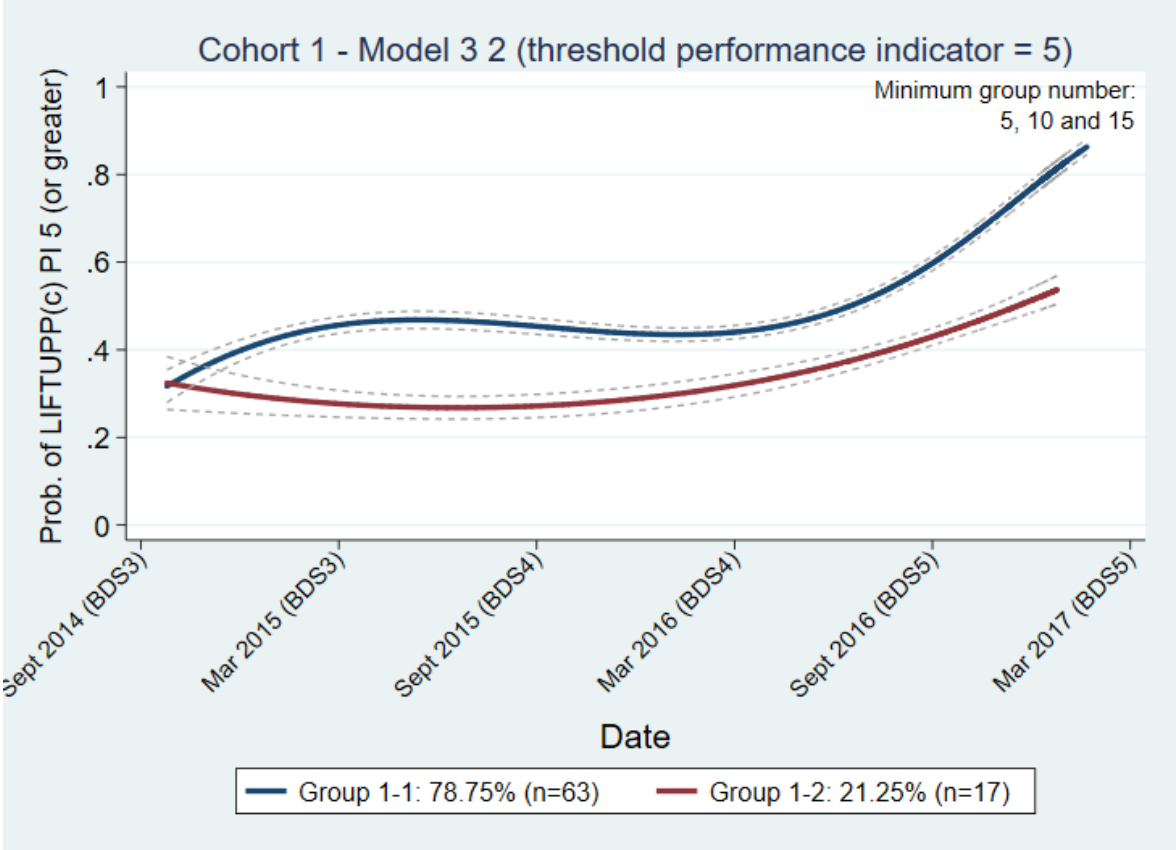


Figure 5.17 - Cohort 1: Trajectory groups for Bernoulli distribution model 3 2. NOTES: Individual data points have been removed to ensure trajectory lines are visible and 95% confidence intervals around estimated group probabilities are depicted by the dotted lines around each trajectory.

Both groups demonstrated an increase in their probability of achieving a LIFTUPP© performance indicator ≥5 over the duration of the BDS course. However, the trajectory pattern for each group differs overall despite both groups having the same probability at the beginning of BDS3 (~0.31). Group 1-1 (78.8% of cohort 1 students; n = 63) followed a cubic trajectory with an initial increase in probability over BDS3, a decrease during BDS4, followed by rapid increase over BDS5. Group 1-2 (21.3%; n = 17) followed a quadratic trajectory with an early decrease in probability over BDS3 and a subsequent steady increase over the duration of both BDS4 and BDS5. By the end of BDS5, students

in group 1-1 were achieving a LIFTUPP© performance indicator ≥5 in 85% of their assessments, whereas group 1-2 were achieving this threshold in 55%. Overall, group 1-1 were the "best" performing group.

Narrow confidence intervals are seen around the estimated group probabilities for each trajectory in both groups.

The trajectories for the 0 1 model (i.e., the most suitable model when a minimum group number restriction of 20 students was applied) are shown in Figure 5.18.

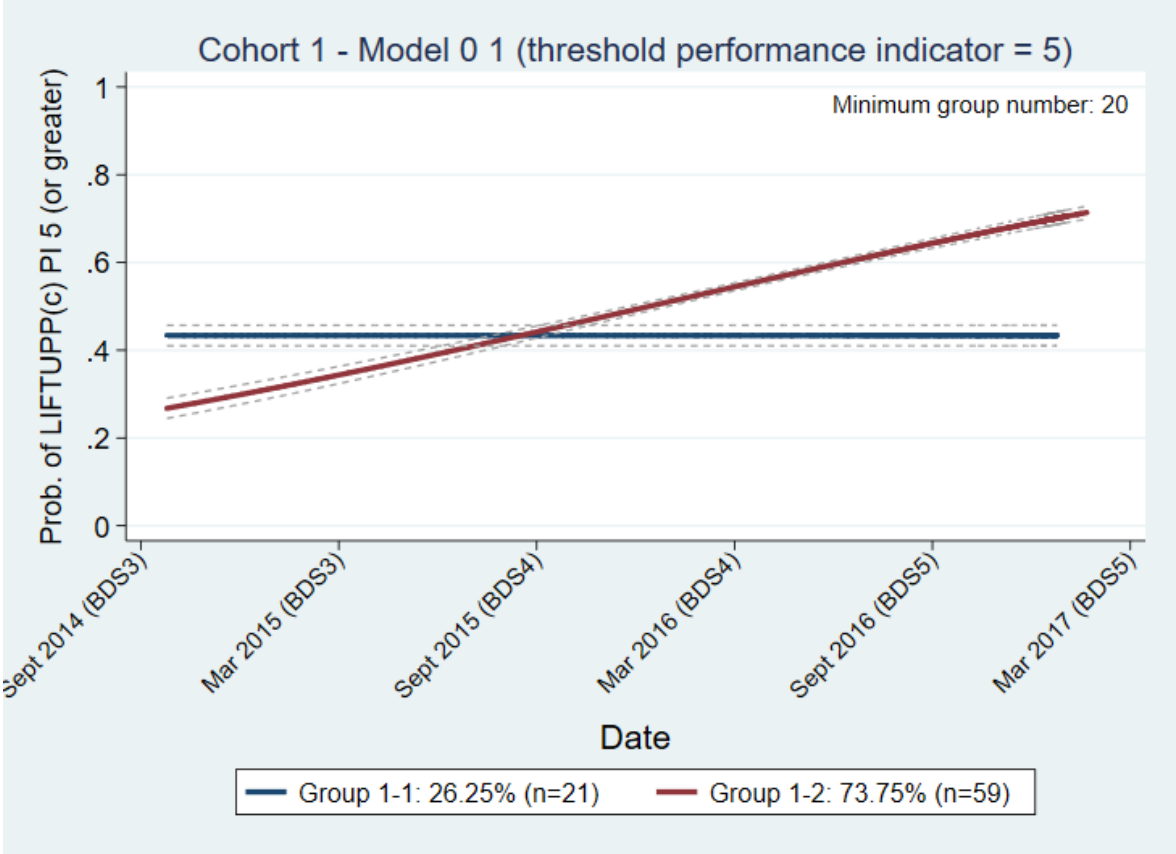


Figure 5.188 - Cohort 1: Trajectory groups for Bernoulli distribution model 0 1. NOTES: Individual data points have been removed to ensure trajectory lines are visible and 95% confidence intervals around estimated group probabilities are depicted by the dotted lines around each trajectory.

In this model, group 1-1 (26.3%; n = 21) produced a zero-order trajectory which indicated no change in a probability of 0.42 over the entire BDS course. Alternatively, group 1-2 (73.8%; n = 59) followed a linear trajectory depicting a gradual increase in the probability of scoring ≥5 from 0.28 at the beginning of BDS to 0.71 by end of BDS5. Overall, group 1-2 were the "best performing".

This model also had tight confidence intervals around the estimated group probabilities for both groups.

Cohort 2

Figure 5.19 presents the trajectories for cohort 2's 1 3 model. Unlike cohort 1, this model returned the highest BIC for any group size larger than five students.

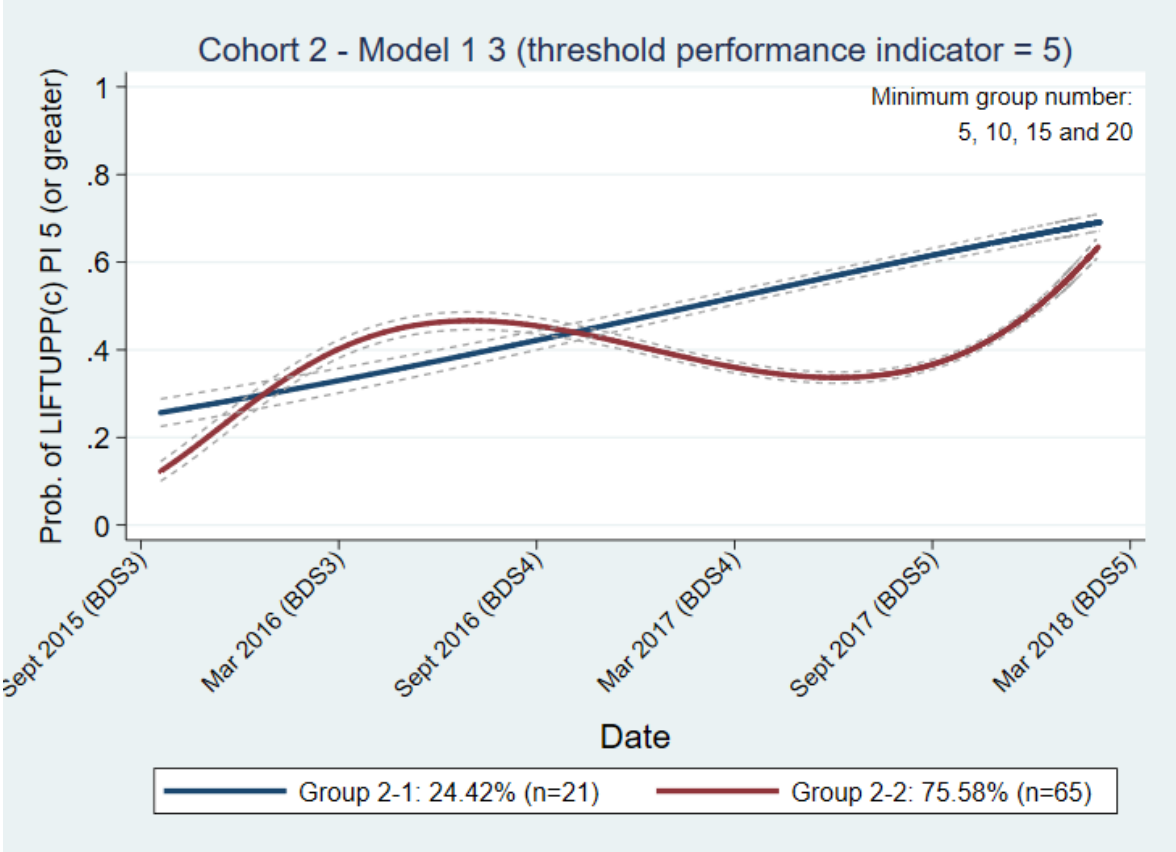


Figure 5.19 - Cohort 2: Trajectory groups for Bernoulli distribution model 1 3. NOTES: Individual data points have been removed to ensure trajectory lines are visible and 95% confidence intervals around estimated group probabilities are depicted by the dotted lines around each trajectory.

An overall increase in the probability of scoring ≥5 was observed in both trajectory groups over the BDS course. The linear trajectory produced by Group 2-1 (24.4%; n = 21) showed an increase from 0.25 at the beginning of BDS3 to 0.70 by the end of the BDS5. Group 2-2 (75.6%; n = 65) displayed followed a cubic shaped trajectory with an increase in probability from 0.11 to 0.45 during BDS3, a decrease to 0.35 during BDS4 and another increase to 0.62 in BDS5. Therefore, at the end of the BDS course, group 2-1 had a 70.0% chance of scoring ≥5, whereas group 2 had a 62.0% chance. Group 2-1 were the "best" performing group overall.

Tight confidence intervals were observed around the estimated group probabilities for both groups.

Cohort 3

Trajectories for the 3 2 model, which was selected to represented cohort 3's clinical LIFTUPP© data, are shown in Figure 5.20. Like cohort 2, the model with the highest BIC was the most suitable for any group size larger than five students.

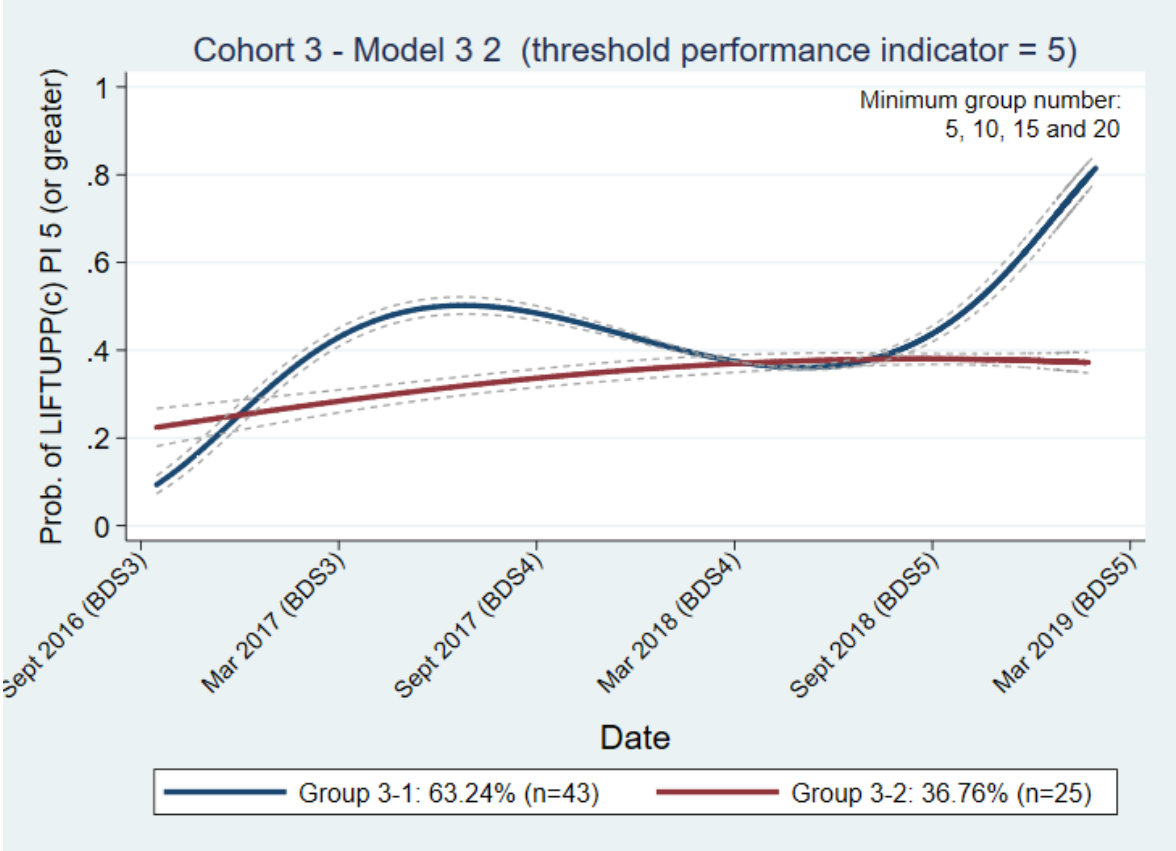


Figure 5.20 - Cohort 3: Trajectory groups for Bernoulli distribution model 3 2. NOTES: Individual data points have been removed to ensure trajectory lines are visible and 95% confidence intervals around estimated group probabilities are depicted by the dotted lines around each trajectory.

Group 3-1 (63.2% of cohort 3 students; n = 43) demonstrated an initial increasing probability (from 0.10 to 0.50) in obtaining a LIFTUPP© performance indicator ≥5 as they progressed through BDS3. The probability then decreased over the BDS4 year (to 0.39) before increasing again over BDS5 (to 0.81). Group 3-2 (36.8%; n = 25), also demonstrated an increase in probability, but not as large as the one demonstrated by group 3-1, as it only increased from 0.23 to 0.39 over the whole BDS course. Therefore, by the end of BDS5, group 3-1 were 42.0% more

likely to be awarded a LIFTUPP© performance indicator ≥5 in a clinical assessment and were, overall, considered as the "best" performing group.

Confidence intervals around the estimated group probabilities for each trajectory group were narrow.

5.4.3 A note on censored normal and Bernoulli models

From the findings presented in sections 5.4.2.1-5.4.2.3, it became apparent that Bernoulli data distributions were preferable for modelling and interpreting LIFTUPP© data patterns compared to those based on censored normal data distributions.

Models based on censored normal data distributions presented student group progression through their average [minimum] LIFTUPP© performance indicator and, for some groups, this displayed little change over time. Alternatively, Bernoulli distribution models displayed more distinctive patterns of student progression through successful achievement of a threshold performance indicator, making it easier to determine which trajectory groups had performed better than others within each cohort, i.e., Bernoulli models appeared to be more parsimonious than those based on censored normal distributions.

Preference for Bernoulli models is further supported by the need for UK dental students to have developed to the level of the "safe beginner". Attainment of this level of development is easier to interpret when an assessment scale which generates discrete data is used (like LIFTUPP©'s performance indicators) as it allows a threshold for the required standard to be clearly defined. Bernoulli models can be generated using LIFTUPP© performance indicators in their original discrete data format, whereas censored normal models require LIFTUPP© performance indicators to be treated as continuous data. Therefore, censored normal GBTMs may not be the most appropriate modelling method for LIFTUPP© data.

Based on these summations, Bernoulli distribution models were chosen for subsequent comparisons between longitudinal clinical assessment data,

undergraduate examination results and performance in postgraduate vocational training to investigate criterion validity (see chapter 7).

5.5 Summary

This chapter has reported on investigations undertaken to explore the main patterns of longitudinal assessment produced by three recently graduated student cohorts from Glasgow Dental School (2017-19).

Summary statistics revealed that student and assessor longitudinal clinical assessment data patterns changed over subsequent cohorts. More assessments were completed by subsequent student cohorts, the proportion of performance indicators 5 and 6 decreased, and the proportion of 4s increased.

LIFTUPP© data from three cohorts of BDS students were successfully modelled using GBTM with either censored normal (raw minimum LIFTUPP© performance indicators) or Bernoulli data distributions (cut-offs =4 and =5). Bernoulli models with threshold cut offs =4 were used to assess content validity, whereas threshold cut offs =5 were used to discriminate students for assessment of criterion validity (see chapter 7).

Models based on either data distribution were able to illustrate clinical development patterns largely showing improvement in all groups of students over time, demonstrating content validity. Data generated by cohort 1 - as LIFTUPP© was being established in the BDS programme - produced several possible models of equally good fit (censored normal and Bernoulli), whereas only one model emerged as the best fitting model for cohorts 2 and 3, suggesting some stabilising of the LIFTUPP© scoring methods over time.

GBTMs based on Bernoulli data distributions emerged as the preferable means for modelling LIFTUPP© data compared to those using censored normal data distributions, and therefore will be used for investigations on criterion validity in chapter 7. However, there was no apparent preference between the two cohort 1 Bernoulli models (3 2 and 0 1), which were returned as the models of best fit for minimum group number restrictions of 5, 10 and 15 (model 3 2) and 20 (model 0 1), respectively. Therefore, both models were taken forward for

comparisons with undergraduate examination and LEP data (chapter 7) – along with cohort 2's 1 3 model and cohort 3's 3 2 model – to investigate if one model was more stable than the other.

The findings of this chapter are discussed further in chapter 9 in conjunction with results reported in other chapters of this thesis.

Chapter 6 - Modelling of postgraduate longitudinal evaluation of performance

6.1 Introduction

The following chapter presents the results of the analyses of LEP data – the longitudinal clinical assessment format used in Scottish postgraduate VDT schemes (see Figure 1.2 in chapter 1 (section 1.1.2) and Table 3.3 in chapter 3 (section 3.5.3.2)). The results presented are based on data obtained from three cohorts of VDPs who completed their one-year VDT in Scotland.

Like the undergraduate examination data, these data will be used when testing a validity argument for undergraduate longitudinal clinical assessment in later chapters, but it is important to describe these data in the first instance.

The summary results provide an overview of the clinical performance data recorded by LEPs. Whilst these are of interest for each cohort, they also may aid in understanding the patterns of VDP clinical performance produced from GBTM. The trajectory models themselves were created to explore how longitudinal clinical performance data could be modelled and used to track clinical development over time. Since longitudinal data can be modelled differently according to various criteria – such as model data distribution, threshold performance scores (where applicable), and restrictions on the minimum number of participants per trajectory group - multiple models are presented to demonstrate how patterns of LEP performance are influenced according to these criteria. These investigations were necessary to determine which GBTM(s) may represent LEP data best and could be selected for comparisons with other forms of assessment in dental education – namely LIFTUPP© (see chapter 7).

6.2 Aim

To establish the main patterns of postgraduate longitudinal clinical assessment for VDPs. The results generated will be used to compare LEPs with LIFTUPP© performance data (see chapter 7) which will address research question 2c: What is the association between undergraduate longitudinal clinical assessment and

postgraduate assessment? (Criterion validity – predictive) – (see chapter 2, section 2.2).

6.3 Method

As detailed in chapter 3 (section 3.5.3.5), the number of clinical LEP assessments completed per VDP per cohort were summarised using histograms, and, based on the data distributions observed, means, standard deviations, medians, minimum, maximum and Q1 and Q3 statistics. LEP scores awarded per clinical assessment were summarised using frequency tables – stratified by cohort and LEP block within each cohort. Bar charts were used to illustrate the frequency of scores awarded per LEP block and over the duration of the VDT year.

Once summarised, GBTMs were produced to model LEP scores across the VDT year for each cohort using the "traj" plugin for Stata statistical software (Jones and Nagin, 2012; 2013). The processes for generation, evaluation, and selection of GBTMs for LEP data have been described in chapter 3 (section 3.5.3.6).

6.4 Results

6.4.1 A description of longitudinal evaluation of performance data

6.4.1.1 Number of longitudinal evaluation of performance assessments

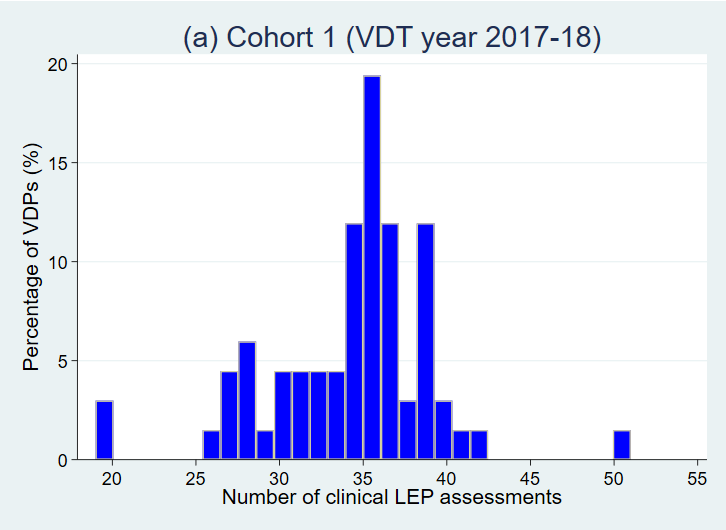
Following completion of data quality assurance, cleaning and management processes outlined in chapter 3 (section 3.5.3.4), the number of LEP assessments eligible for inclusion in the study for cohorts 1, 2 and 3 were 2,294/3,139 (73.1%), 2,839/3,868 (73.4%) and 1,956/2,599 (75.3%), respectively.

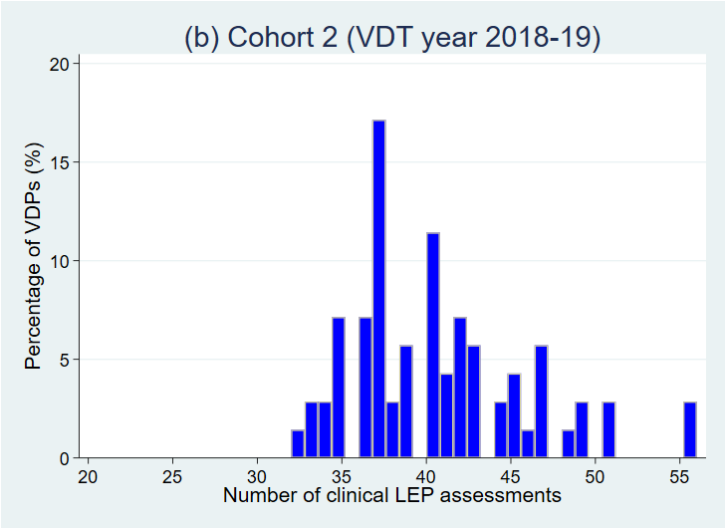
Histograms for the number of clinical LEP assessments performed by VDPs displayed either normal (cohort 1 (n = 67) - Figure 6.1(a); cohort 3 (n = 60) – Figure 6.1(c)) or slight positively skewed distributions (cohort 2 (n = 70): Figure 6.1(b)). Means, standard deviations, medians, minimum, maximum and Q1 and Q3 statistics are summarised for each cohort in appendix 10 and illustrated graphically with boxplots in Figure 6.2.

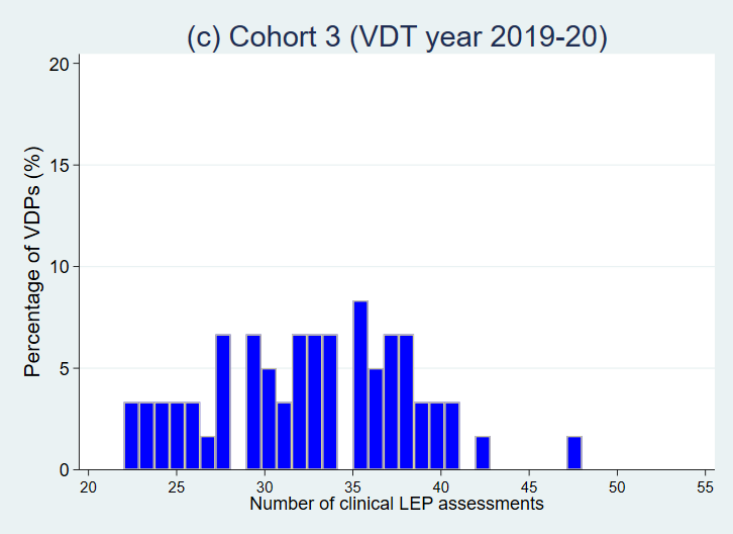
Although the histogram for cohort 2 appeared to show a slight positively skewed distribution, the mean and median were similar in value (40.6 vs 40.0, respectively) – which is indicative of symmetrical distribution.

Positively skewed distributions were initially expected since VDPs are required to complete a minimum of 42 LEPs over the duration of their VDT year (see chapter 3, section 3.5.3.2). However, the means, medians, minimum and Q1 values for both cohorts indicate that, although each VDP may have completed 42 LEP assessments, not all of them incorporated assessment of practical clinical skills. In cohort 1, the mean and median number of LEPs involving assessment of a clinical skill were 34.2 and 35.0, respectively. Only 25% of cohort 1 VDPs completed at least 37 clinical assessments (Q3 = 37.0). As above, cohort 2's mean and median were 40.6 and 40.0 (respectively) but over 75% had completed at least 43 clinical assessments. In cohort 3, the mean and median were 32.6 and 33.0 (respectively) and, like for cohort 1, only 25% of VDPs completed at least 37 clinical assessments (Q3 = 37.0).

It was noted that the number of LEPs completed by cohort 3 would have undoubtedly been impacted by the COVID-19 pandemic. As part of a national lockdown imposed by the Scottish and UK governments, dental clinical activity was greatly reduced, therefore there would have been few or no opportunities for VDPs to complete LEP assessments between 24th March 2020 and the end of the VDT year (31st July 2020). This is discussed further in chapter 9 (section 9.4.1).







Figures 6.1 (a-c) – Data distributions for the number of clinical longitudinal evaluation of performance (LEP) assessments performed per vocational dental practitioner (VDP). (a) Cohort 1 (n = 67); (b) Cohort 2 (n = 70); (c) Cohort 3 (n = 60). VDT = Vocational dental training

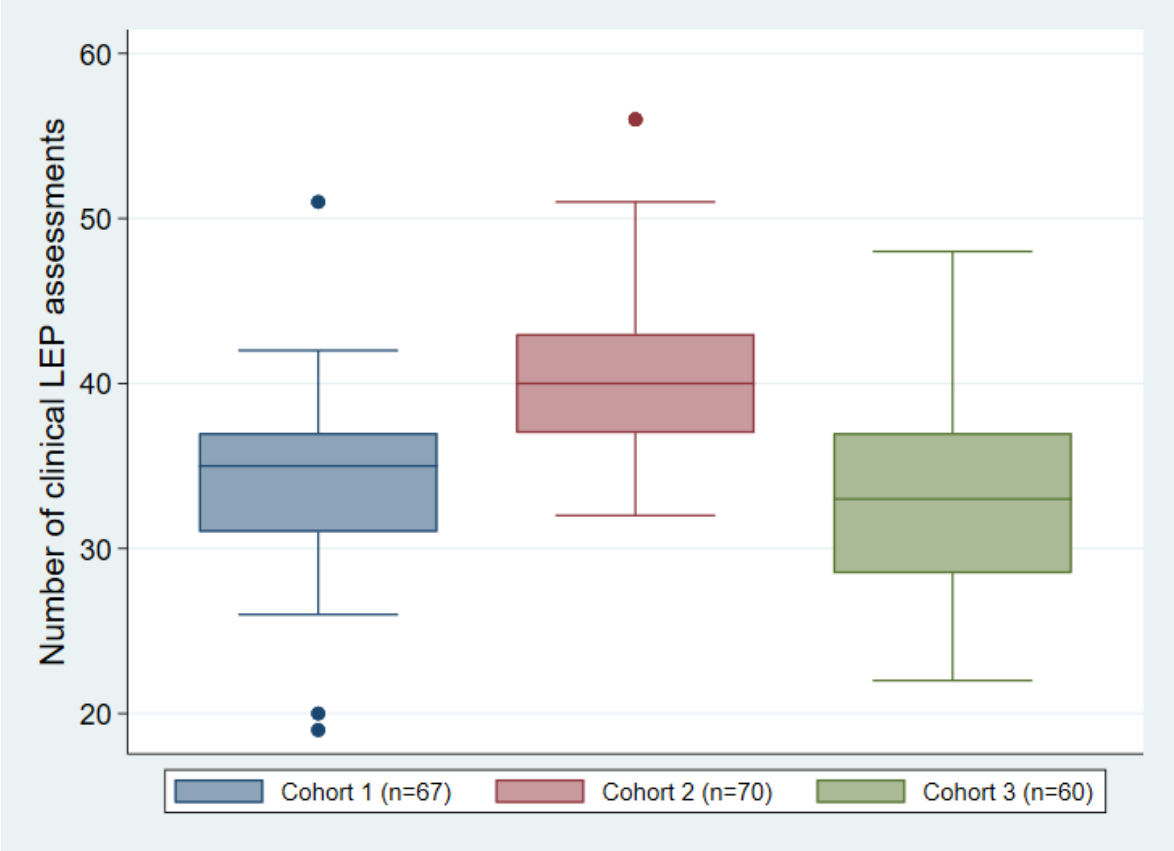


Figure 6.2 – Cohorts 1, 2 and 3: Boxplots for the number of clinical longitudinal evaluation of performance (LEP) assessments completed per vocational dental practitioner (VDP).

Furthermore, there was a wide range in the number of clinical assessments completed by VDPs in each cohort: 19 to 51 (cohort 1), 32 to 56 (cohort 2) and 22 to 48 (cohort 3). The variation between each cohort's LEP data was noted and is discussed further in chapter 9 (section 9.4.1).

6.4.1.2 Longitudinal Evaluation of Performance scores

The frequencies of LEP scores awarded across all eligible clinical procedures in each cohort (per LEP block and across the entire VDT year) are displayed in Table 6.1. Only one score for clinical LEPs (technical skill and manual dexterity) was used for the analyses in this study, as opposed to LIFTUPP© data which required a minimum performance indicator to be identified from multiple performance indicators assigned per clinical assessment completed (see chapter 3, section 3.5.3.4)

In cohorts 1, the most common LEP score for block 1 was 6 (28.3%). In cohorts 2 and 3, it was 5 (cohort 2: 34.3%; cohort 3: 31.7%). For both cohorts 1 and 2, the

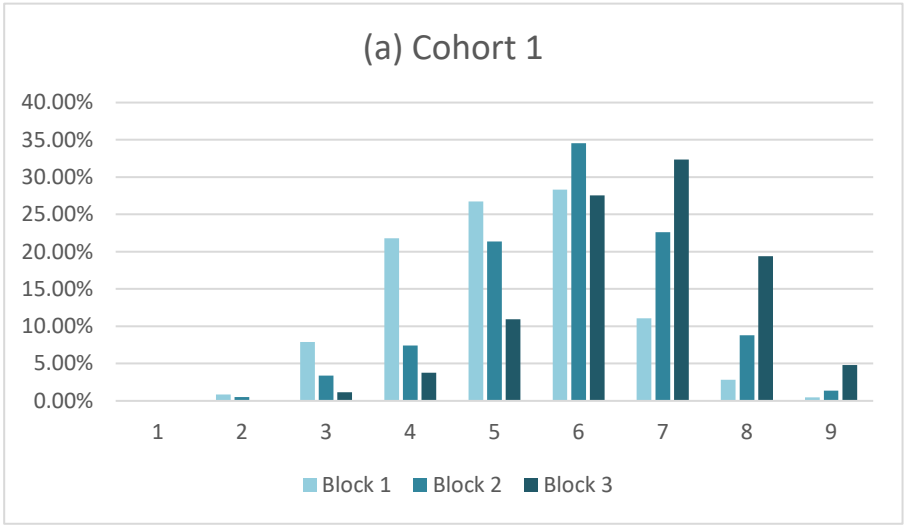
most common scores in blocks 2 and 3 were 6 (cohort 1: 34.6%; cohort 2: 30.7%) and 7 (cohort 1: 34.6%; cohort 2: 30.7%), respectively (Table 6.1 and Figures 6.3 (a-b)). The most common score awarded in blocks 2 and 3 for cohort 3 was 6 (Table 6.1 and Figure 6.3c).

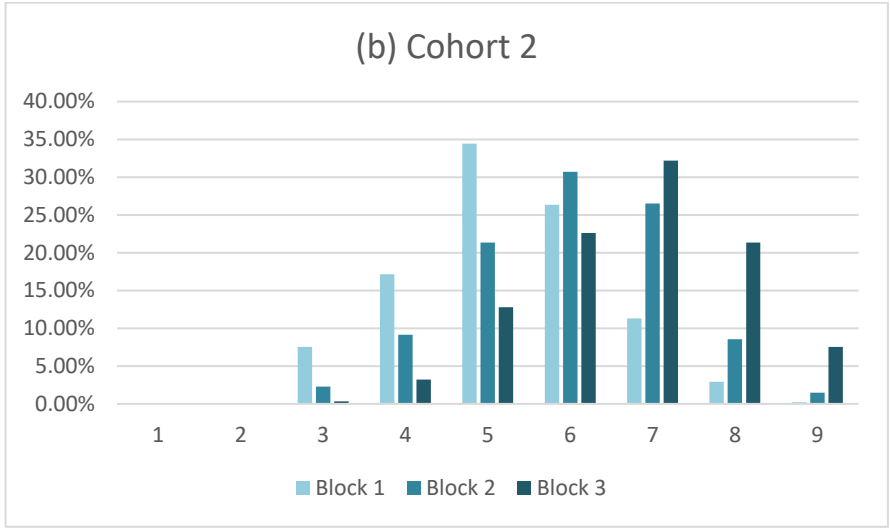
In all three cohorts, the score awarded most over the duration of the VDP year was 6, which accounted for 30.3%, 26.7% and 29.9% of all scores awarded to cohort 1, 2 and 3 VDPs, respectively. “Needs improvement” scores (i.e., 1, 2 or 3) were rarely awarded as, collectively, they comprised less than 5% of all scores in all three cohorts. “Satisfactory” scores (i.e., 4, 5 or 6) made up approximately 60.0% of scores in all three cohorts and “superior” scores accounted for approximately 1/3 of scores in all cohorts (see Table 6.1).

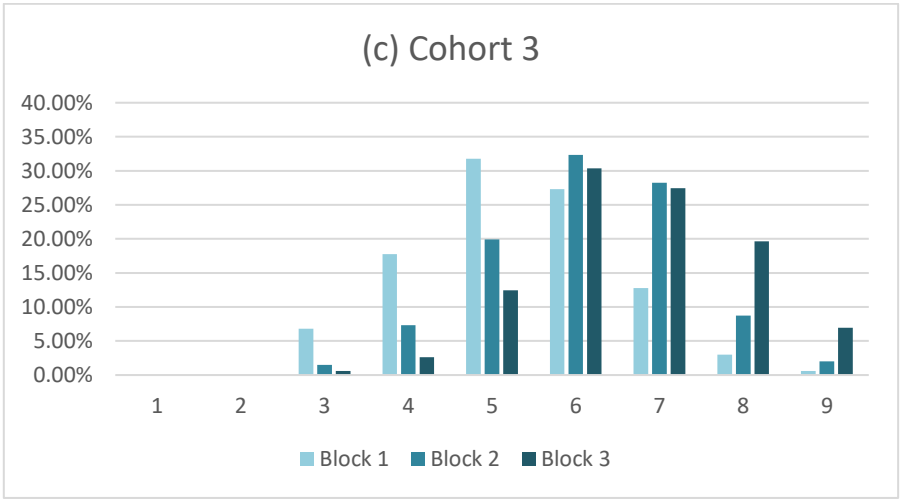
Table 6.1 – Frequencies of longitudinal evaluation of performance (LEP) scores awarded in cohorts 1, 2 and 3.

Cohort	Block	Number of LEP assessments completed	LEP scores [n (%)]								
			1	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***
1	1	812		7 (0.86%)	64 (7.88%)	177 (21.80%)	217 (26.72%)	**230 (28.33%)**	90 (11.08%)	23 (2.83%)	
	2	796			27 (3.39%)	59 (7.41%)	170 (21.36%)	**275 (34.55%)**	180 (22.61%)	70 (8.79%)	11 (1.38%)
	3	686			8 (1.17%)	26 (3.79%)	75 (10.93%)	189 (27.55%)	**222 (32.36%)**	133 (19.39%)	33 (4.81%)
	Total	2,294		11 (0.48%)	99 (4.32%)	262 (11.42%)	462 (20.14%)	**694 (30.25%)**	492 (21.45%)	226 (9.85%)	48 (2.09%)
2	1	956			72 (7.53%)	164 (17.15%)	**329 (34.41%)**	252 (26.36%)	108 (11.30%)	28 (2.93%)	
	2	1007			23 (2.28%)	92 (9.14%)	215 (21.35%)	**309 (30.69%)**	267 (26.51%)	86 (8.54%)	15 (1.49%)
	3	876				28 (3.20%)	112 (12.79%)	198 (22.60%)	**282 (32.19%)**	187 (21.35%)	66 (7.53%)
	Total	2,839			98 (3.45%)	284 (10.00%)	656 (23.11%)	**759 (26.73%)**	657 (23.14%)	301 (10.60%)	83 (2.92%)
3	1	806			55 (6.82%)	143 (17.74%)	**256 (31.76%)**	220 (27.30%)	103 (12.78%)	24 (2.98%)	5 (0.62%)
	2	804			12 (1.49%)	59 (7.34%)	160 (19.90%)	**260** (32.34%)	227 (28.23%)	70 (8.71%)	16 (1.99%)
	3	346				9 (2.60%)	43 (12.43%)	**105 (30.35%)**	95 (27.46%)	68 (19.65%)	24 (6.94%)
	Total	1,956			69 (3.53%)	211 (10.79%)	459 (23.47%)	**585 (29.91%)**	425 (21.73%)	162 (8.28%)	45 (2.30%)

Mode scores highlighted in **bold.** Cells where n<5 have been greyed out.







Figures 6.3 (a-c) – Bar chart representations for proportions of longitudinal evaluation of performance (LEP) scores awarded per vocational dental practitioner (VDP) assessment per LEP block.

The proportions of scores awarded within each cohort across the duration of VDT year appeared similar. There was less than 2% difference in the frequency of most LEP scores between all three cohorts. Exceptions to this finding were observed around scores 5 and 6, whereby a slightly greater proportion of 5s (approximately 3.0%) had been awarded in cohorts 2 and 3 compared to cohort 1, and slightly less 6s (approximately 3.5%) had been awarded in cohort 2 compared to cohorts 1 and 3.

6.4.2 Modelling longitudinal evaluation of performance data using group-based trajectory modelling

The following section describes the GBTMs selected to represent the clinical LEP data according to any imposed minimum number of 5, 10, 15 and 20 VDPs per group. Details of the processes used to generate, evaluate, and select the GBTMs have previously been provided in chapter 3 (section 3.5.3.6). Like for LIFTUPP© data (see chapter 5, section 5.4.2), models for LEPs were generated using zero-order, linear, quadratic, and cubic trajectories.

As per section 6.4.1.1 (above), the number of LEP assessments eligible for inclusion in the study for cohorts 1, 2 and 3 were 2,294, 2,839 and 1,956, respectively. Therefore, GBTMs were generated from these number of assessments spread across 67 (cohort 1), 70 (cohort 2) and 60 (cohort 3) VDPs.

Since Bernoulli models were deemed preferable to censored normal data models for LIFTUPP© data (see chapter 5, section 5.4.3), only Bernoulli models are presented.

6.4.2.1 Threshold score = 4

A score of 4 is the lowest of the three "satisfactory" scores which can be subjectively awarded by assessors using the LEP rating system (see chapter 3, section 3.5.3.2). Adopting a LEP score of 4 as the threshold for competent performance resulted in only single trajectory models being returned (with no statistical errors) in all three cohorts. Figures 6.4-6.6 display the "best fitting" single-group models based on the BIC for each cohort – see chapter 3, section 3.5.3.6).

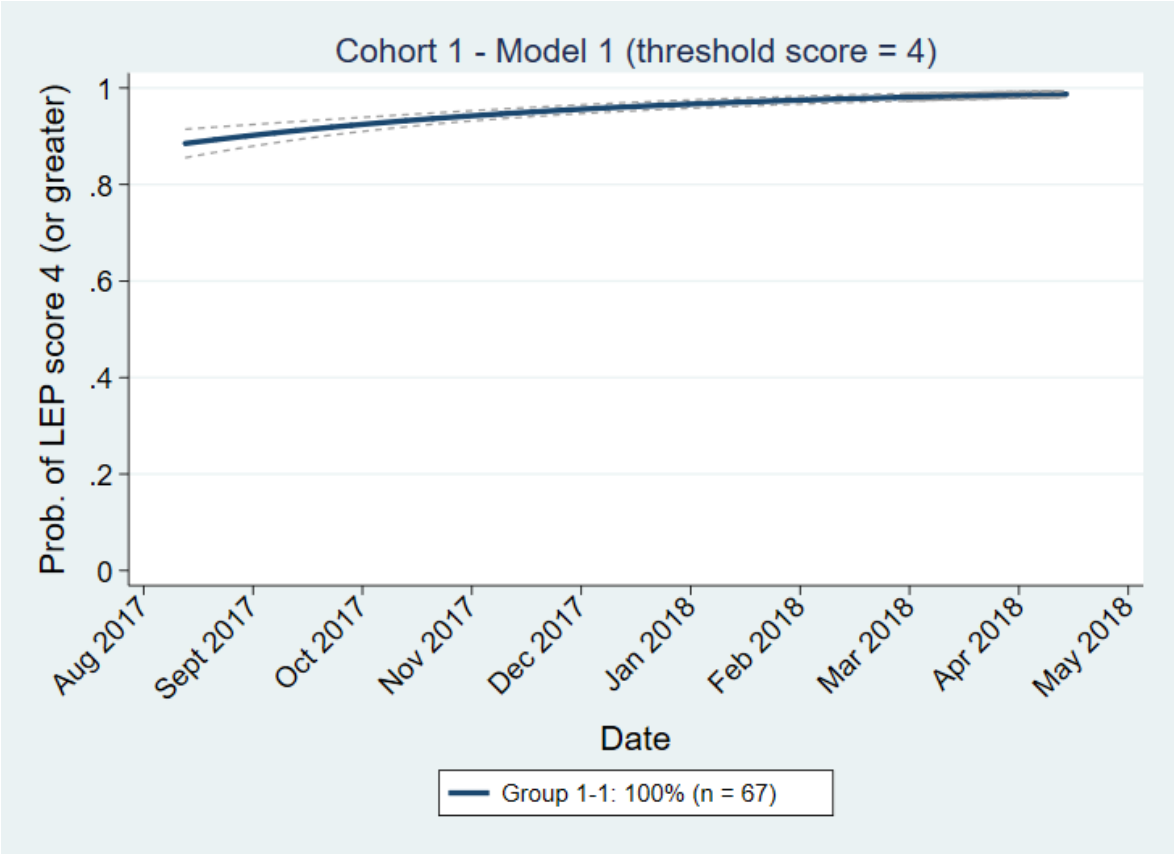


Figure 6.4 – Cohort 1: Single trajectory for clinical longitudinal evaluation of performance (LEP) data if 4 was used as the threshold score for competent performance.

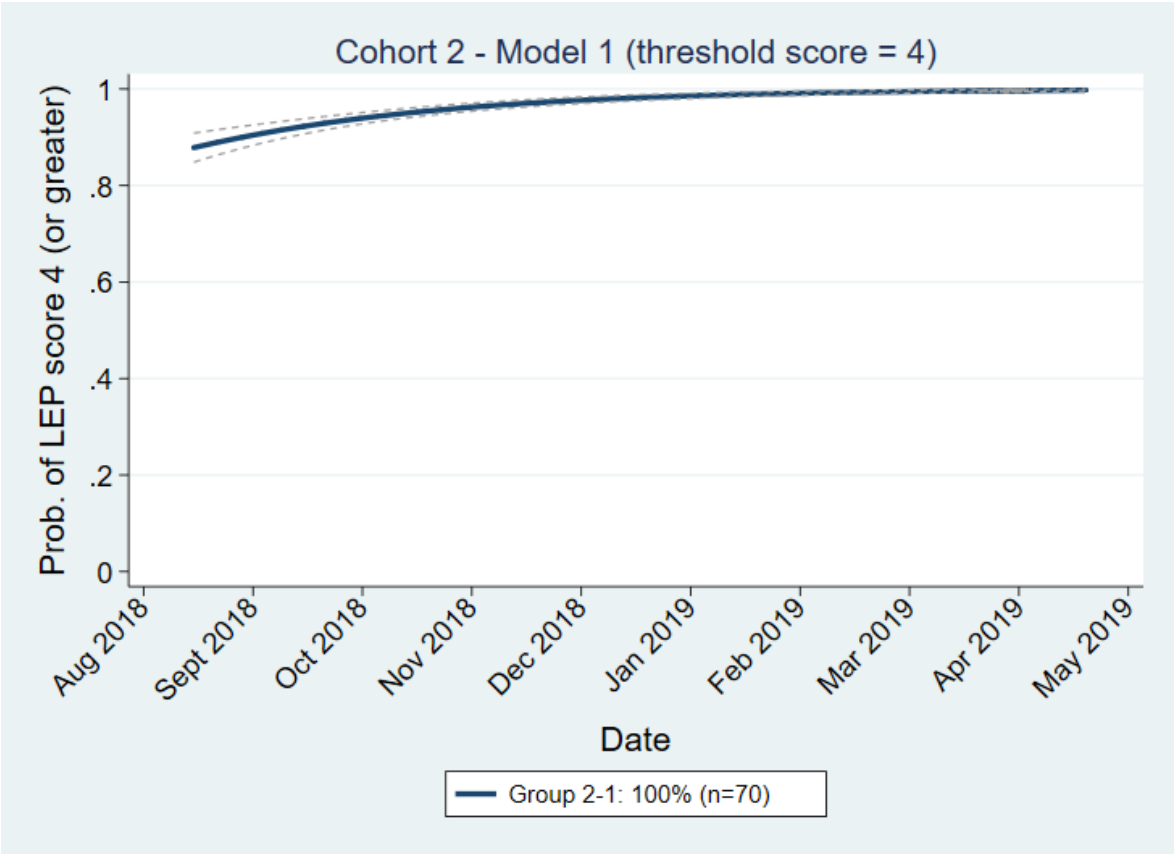


Figure 6.5 – Cohort 2: Single trajectory for clinical longitudinal evaluation of performance (LEP) data if 4 was used as the threshold score for competent performance.

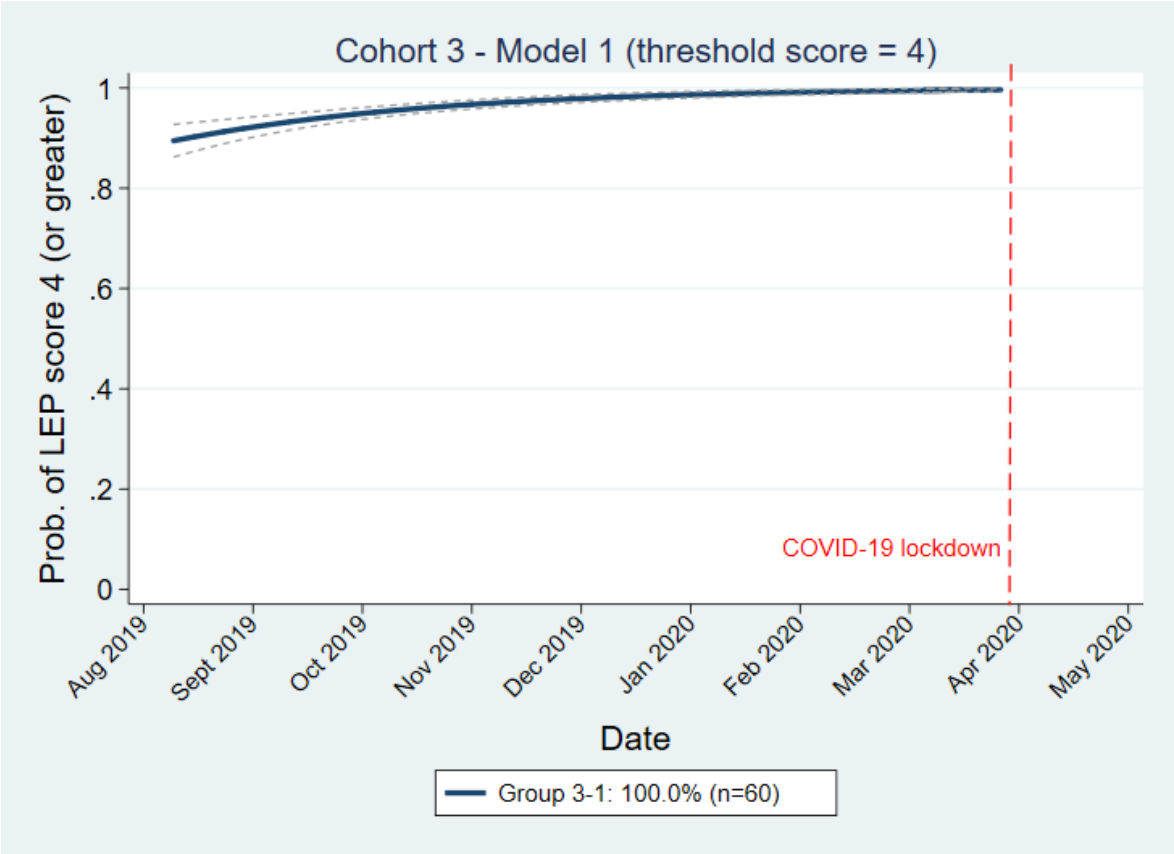


Figure 6.6 – Cohort 3: Single trajectory for clinical longitudinal evaluation of performance (LEP) data if 4 was used as the threshold score for competent performance.

Each cohort's trajectory demonstrated all VDPs increased their probability in scoring ≥4 over the duration of their VDT year. However, single group trajectories do not distinguish different student development patterns, and the lack of discrimination means they were unsuitable for comparisons against other assessment methods. Therefore, models adopting threshold scores of 4 were discounted for investigating the criterion validity of longitudinal assessment (see chapter 7). However, it is worth noting that the threshold 4 models in all three cohorts indicate that VDPs' have a 100% probability of scoring ≥4 by the end of VDT.

6.4.2.2 Threshold score = 5

A score of 5 is the middle of the three "satisfactory" scores which can be subjectively awarded by assessors using the LEP rating system (see chapter 3, section 3.5.3.2).

In cohort 1, a two-group model returned the least negative BIC (model 0 2). This was the only model returned by traj which consisted of more than one trajectory and was the most suitable for any group size larger than five VDPs (Figure 6.7).

Group 1-1 (zero-order; 35.8%; n =24) started their year with a high probability (~0.95) of scoring ≥5 and this remained consistent across the year. Group 1-2 (quadratic; 64.1%; n =43) began VDT with a 0.45 probability for scoring ≥5 which then increased to 0.96 by the end of VDT.

In cohorts 2 and 3, only single trajectory models were returned (with no statistical errors). Figures 6.8 and 6.9 display the "best fitting" single-group models based on the BIC for cohorts 2 and 3, respectively. Both cohort's single trajectories illustrated that VDPs increased their probability in scoring ≥5 (from 0.65 to 0.99) over the course of VDT.

Although different student development patterns could be distinguished in cohort 1, all models adopting threshold scores of 5 were discounted since only single group trajectories were found for cohorts 2 and 3 and, therefore, were unsuitable for comparison against other assessment methods.

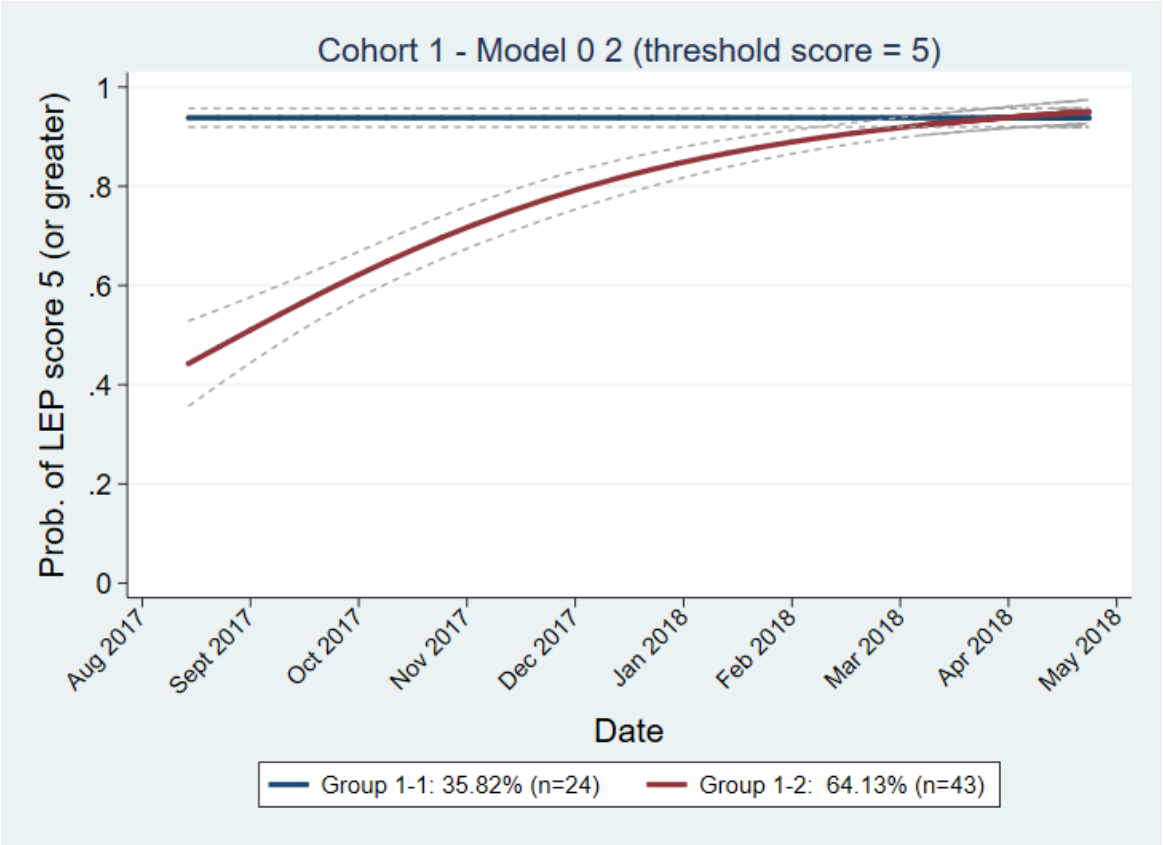


Figure 6.7 - Cohort 1: Single trajectory for clinical longitudinal evaluation of performance data (LEP) if 5 was used as the threshold score for competent performance.

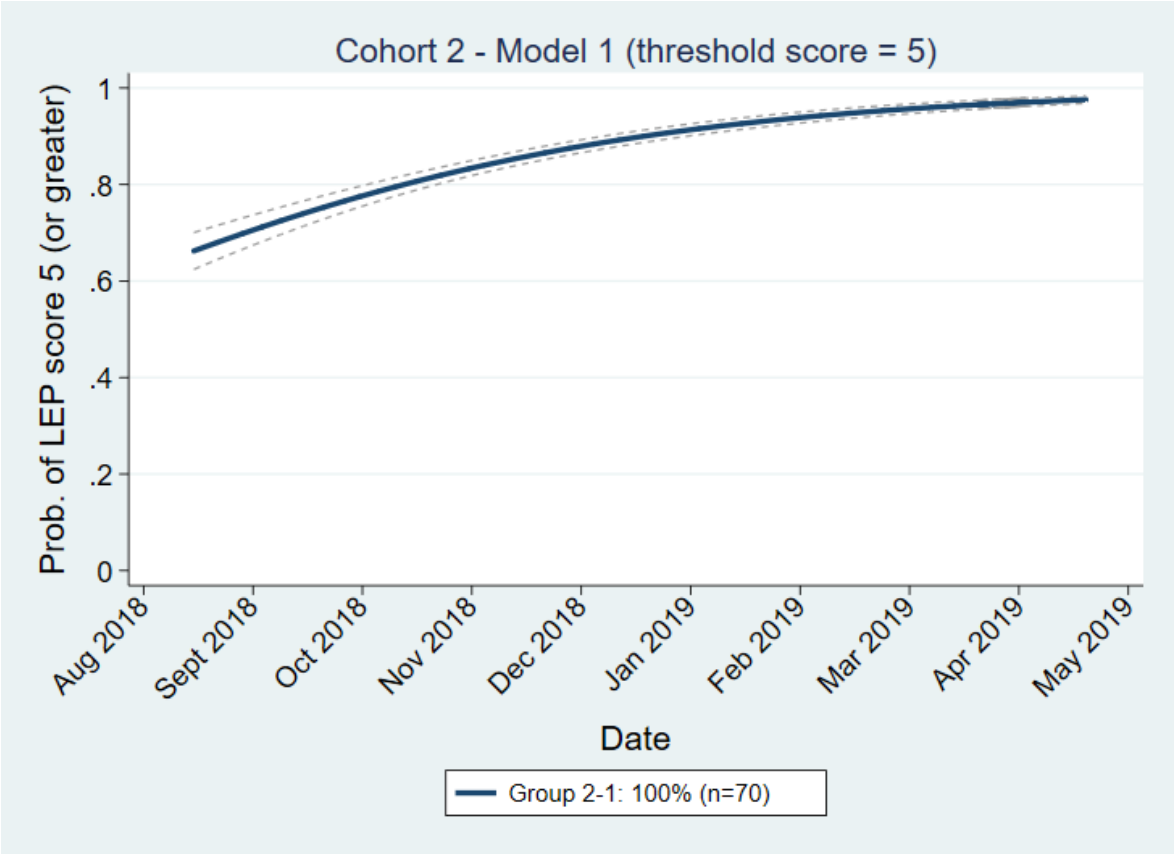


Figure 6.8 - Cohort 2: Single trajectory for clinical longitudinal evaluation of performance (LEP) data if 5 was used as the threshold score for competent performance.

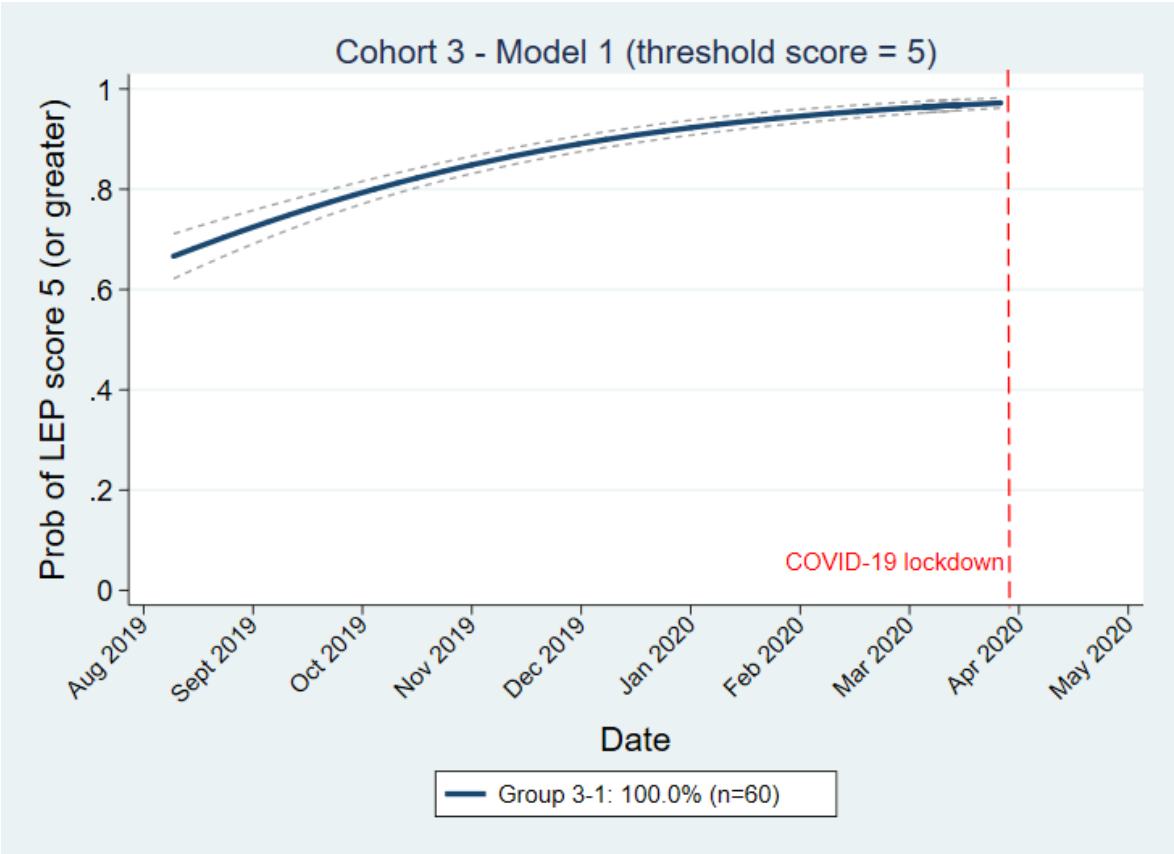


Figure 6.9 - Cohort 3: Single trajectory for clinical longitudinal evaluation of performance (LEP) data if 5 was used as the threshold score for competent performance.

6.4.2.3 Threshold score = 6

A score of 6 is the highest of the three "satisfactory" scores which can be subjectively awarded by assessors using the LEP rating system (see chapter 3, section 3.5.3.2).

Following the selection criteria described in chapter 3 (section 3.5.3.6), there were multiple incidences where the difference in BIC between the model with the highest (least negative) BIC and other models was found to be less than two. These models were subjected to further scrutiny, whereby the trajectory graphs produced by each model were compared to determine the most parsimonious models.

Cohort 1

In cohort 2, the 1 3 model displayed the highest BICs out of all models returned without statistical errors and satisfied the minimum group number restrictions of at least 5, 10, 15 and 20 students.

In cohort 1, model 3 3 3 2 had the highest BIC of all the returned models when a minimum group number restriction of 5 was applied. Its BIC was at least two greater than the model with the second highest. Model 1 1 had the highest BIC of all models when minimum group number restrictions of 10, 15 and 20 were applied, but its BIC was not two greater than the BIC of model 2 1. When the graphical trajectories and parameter estimates of models 1 1 and 2 1 were compared, there was very little difference between them and, therefore, description of VDP clinical performance would not differ significantly. As a result, either of the two models could have been selected to represent cohort 1's LEP data - provided a threshold score of 6, and minimum group memberships of 10, 15 and 20 VDPs were set. In the end, model 1 1 was selected alongside model 3 3 3 2 to represent cohort 1's LEP data (if a threshold score of 6 was implemented) as it returned the highest BIC overall. Model 3 3 3 2 was the only one four-group model with at least five students in the smallest trajectory group returned if a threshold score of 6 was adopted.

Cohort 2

In cohort 2, the highest BIC was returned by model 2 3 1 3. However, this model's BIC was not greater than two when compared to that of models 1 3 1 2, 1 1 1 2, 1 1 3 3, 2 2 1 3 and 2 3 2 3. All these models did not satisfy the criteria of having at least 15 and 20 students in all their trajectory groups. The graphical trajectories and parameter estimates of all six models were similar and, therefore, would have little impact on how VDP clinical performance would be described. However, model 2 3 1 3 was ultimately selected to represent cohort 2's LEP data (if a threshold score of 6, and minimum group number restrictions of 5 and 10 were in effect) as it had the highest BIC overall.

Model 2 2 3 returned the highest BIC of all models implementing a threshold score of 6 and a minimum group number restriction of 15. It's BIC was at least two greater than the model with the second highest and it was therefore selected initially. However, on further investigation, the confidence intervals around the estimated group probabilities of each trajectory group were very wide which was contrary to the criteria for adequate model selection (Nagin, 2005; Nagin and Odgers, 2010).

The model with the second highest BIC which satisfied the criteria of having at least 15 VDPs per trajectory group was model 3 2 – which was also the model with the highest BIC if a restriction of at least 20 VDPs per group was applied. The BIC of model 3 2 was not at least two greater than that of the models 1 3, 2 2 and 3 3. However, upon comparison, the graphical trajectories and parameter estimates were similar and therefore there were no significant advantages to selecting any of these models over mode 3 2 to represent cohort 2's LEP data (where a threshold score of 6 and minimum group number restrictions of 15 and 20 were applied). Model 3 2 was ultimately chosen by the researcher since it had the highest BIC overall.

Cohort 3

The highest BIC was returned by model 3 0 3, this model contained at least five VDPs in each trajectory group and had a BIC difference greater than two when compared against the model with the second highest BIC.

If the criteria for having at least 10, 15 and 20 VDPs per trajectory group were applied, the model with the highest BIC was model 1 3. This model also had a BIC difference greater than two when compared against the model with the second highest BIC for minimum group number restrictions of 10, 15 and 20 VDPs.

A summary of the models with the least negative BICs within each cohort is presented in Table 6.2.

Table 6.2 - Bernoulli distribution group-based trajectory models (GBTMs) (threshold score = 6) selected to represent clinical longitudinal evaluation of performance (LEP) data for cohorts 1, 2 and 3. VDPs = Vocational dental practitioners. BIC = Bayesian information criterion.

Threshold score ≥6								
Cohort	**Number of VDPs (n)**	**Number of assessments**	**Model**	**Contains at least X VDPs per group, where X =**				**BIC (based on number of VDPs)**
				5	***10***	***15***	***20***	
1	67	2,294	3 3 3 2	✓	✗	✗	✗	-1274.38
			1 1	✓	✓	✓	✓	-1280.60
2	70	2,839	2 3 1 3	✓	✓	✗	✗	-1505.93
			3 2	✓	✓	✓	✓	-1538.91
3	60	1,956	3 0 3	✓	✗	✗	✗	-1126.03
			1 3	✓	✓	✓	✓	-1128.86

Each of the selected models performed well on all tests of model adequacy laid out in Nagin (2005) and Nagin and Odgers (2010) (Tables 6.3-6.5). Across all three cohorts, the AvePP was at least 0.86 for each group in all models, i.e., greater than the 0.70 minimum value recommended by Nagin and Odgers (2010). The odds of correct classification for all groups in each GBTM were greater than 5.0, indicating each model's assignment accuracy was good. There was also close correspondence between each trajectory group's estimated probability of group membership and the proportion of students classified to that group according to posterior probability of group membership across all models.

Table 6.3 – Cohort 1: Average posterior probabilities, odds of correct classification, estimated probability, and proportion of group membership (according to posterior probability of group membership) for clinical longitudinal evaluation of performance (LEP) trajectory group models (based on Bernoulli data distribution). NOTE: All values are rounded to two decimal places. BIC = Bayesian information criterion.

Cohort (n)	**Threshold score**	**Model**	**BIC**	**Group**	**n per group**	**Average posterior probability**	**Odds of correct classification**	**Estimated probability of group membership**	**Proportion of students classified to group according to posterior probability of group membership**
1 (67)	6	3 3 3 2	-1274.38	1-1	10	.94	84.84	.17	.15
				1-2	28	.86	8.96	041	.42
				1-3	22	.87	14.71	.32	.33
				1-4	7	.97	256.36	.10	.10
		1 1	-1280.60	1-1	24	.95	31.28	.38	.36
				1-2	43	.93	8.60	.62	.64

Table 6.4 – Cohort 2: Average posterior probabilities, odds of correct classification, estimated probability, and proportion of group membership (according to posterior probability of group membership) for clinical longitudinal evaluation of performance (LEP) trajectory group models (based on Bernoulli data distribution). NOTE: All values are rounded to two decimal places. BIC = Bayesian information criterion.

Cohort (n)	Threshold score	Model	BIC	Group	n per group	Average posterior probability	Odds of correct classification	Estimated probability of group membership	Proportion of students classified to group according to posterior probability of group membership
2 (70)	6	2 3 1 3	-1505.93	2-1	11	.93	77.17	.15	.16
				2-2	29	.91	14.47	.41	.41
				2-3	18	.91	29.90	.26	.26
				2-4	12	.92	54.51	.18	.17
		3 2	-1538.91	2-1	29	.95	24.79	.42	.41
				2-2	41	.96	16.06	.58	.59

Table 6.5 – Cohort 3: Average posterior probabilities, odds of correct classification, estimated probability, and proportion of group membership (according to posterior probability of group membership) for clinical longitudinal evaluation of performance (LEP) trajectory group models (based on Bernoulli data distribution). NOTE: All values are rounded to two decimal places. BIC = Bayesian information criterion.

Cohort (n)	**Threshold score**	**Model**	**BIC**	**Group**	**n per group**	**Average posterior probability**	**Odds of correct classification**	**Estimated probability of group membership**	**Proportion of students classified to group according to posterior probability of group membership**
3 (60)	6	3 0 3	-1126.03	3-1	27	.90	12.10	.43	.45
				3-2	6	.90	74.22	.11	.10
				3-3	27	.93	15.17	.45	.45
		1 3	-1128.86	3-1	31	.92	10.53	.51	.52
				3-2	29	.92	12.46	.49	.48

Cohort 1 - Threshold = 6; Minimum number VDPs per group = 5

Figure 6.10 presents the group trajectories for model 3 3 3 2. This GBTM was the most suitable model if each trajectory group was required to have a minimum of 5 VDPs and a threshold LEP score of 6 was adopted.

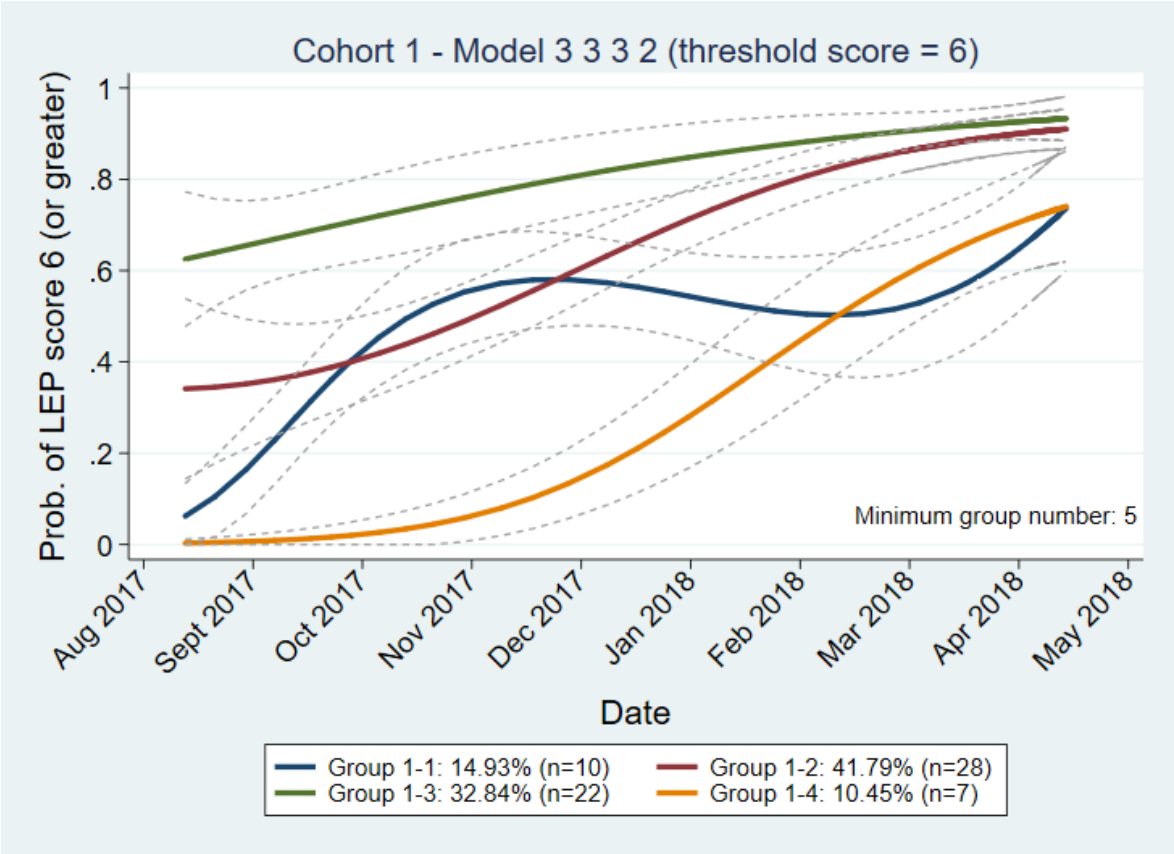


Figure 6.10 - Cohort 1: Trajectory groups for Bernoulli distribution model 3 3 3 2. NOTE: 95% confidence intervals around estimated group probabilities are depicted by the dotted lines around each trajectory.

All four groups demonstrated an increase in their probability of achieving a LEP score ≥6 over the duration of VDT year. Both groups 1-2 (41.8%; (n = 10) and 1-3 (32.8%; n = 22) displayed cubic trajectories and high probabilities (0.88 and 0.90, respectively) of achieving a score ≥6 by the end of VDT. Group 1-2 displayed a greater degree of improvement over the course of the year having started VDT with a probability of 0.35, whereas group 1-3 started with a 0.62 probability. Groups 1-1 (14.9%; n = 10) and 1-4 (10.5%; n = 7) both completed VDT a probability of 0.70. However, group 1-1 produced a cubic shaped trajectory which showed an initial improvement from 0.05 to 0.57 between August and November (2017), followed by a decrease to 0.50 between December (2017) and March (2018), followed by another increase to 0.70 over March and April (2018). In comparison, group 1-4 produced a quadratic curve demonstrated

a gradual increase in probability from 0.00 to 0.70 over the progression of the VDT year. Overall, group 1-3 were regarded as the “best” performing group.

Wider confidence intervals around the estimated group probabilities of each trajectory group were noted in this model when compared to other GBTMs produced for LIFTUPP© and LEP data. However, this was consistent with other “well fitting” LEP models (see below and appendix 10).

Cohort 1 - Threshold = 6; Minimum number VDPs per group = 10, 15 and 20

The group trajectories for model 1 1 are shown in Figure 6.11. This GBTM was the most suitable model if each trajectory group was required to have a minimum of 10, 15 and 20 VDPs and a threshold LEP score of 6 was adopted.

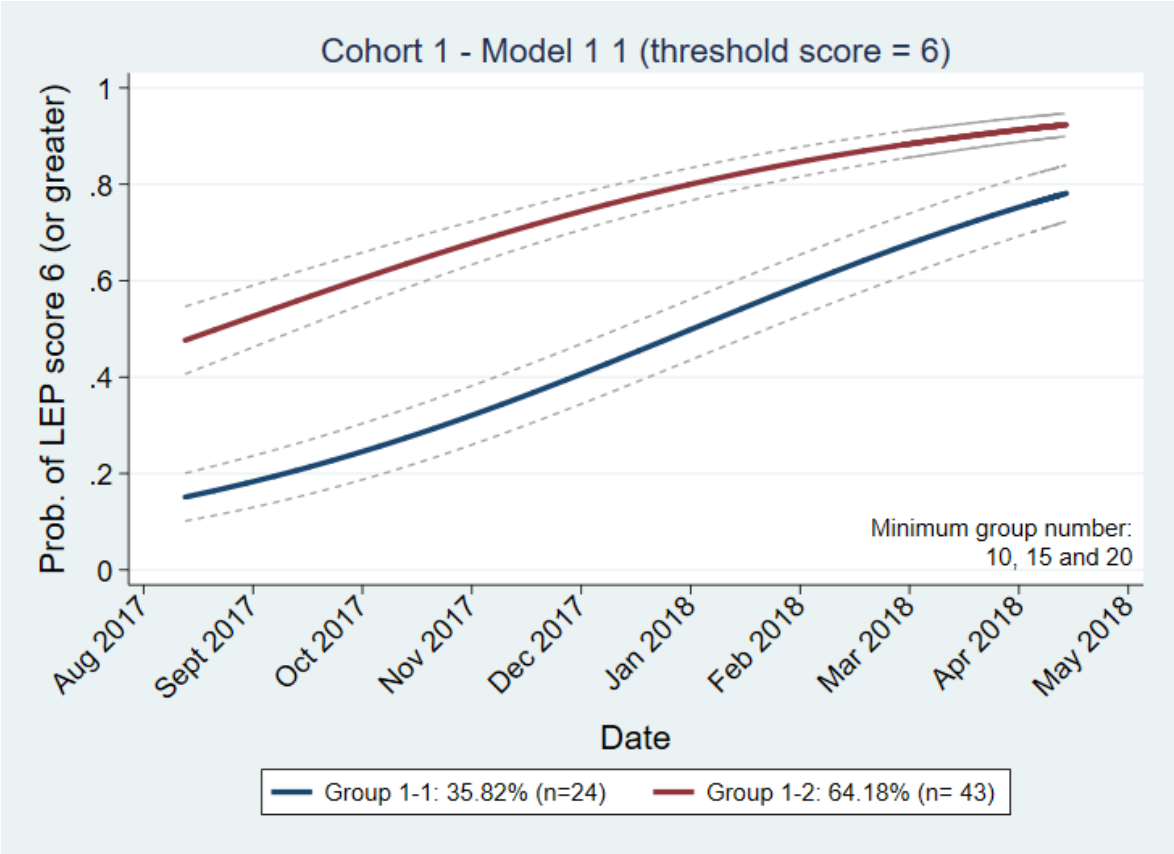


Figure 6.11 - Cohort 1: Trajectory groups for Bernoulli distribution model 1 1 (threshold score = 6). NOTE: 95% confidence intervals around estimated group probabilities are depicted by the dotted lines around each trajectory.

Both groups 1-1 and 1-2 displayed increasing linear trajectories. Group 1-1 (35.8%; n = 24) demonstrated a gradual increase in their probability of scoring a 6 (or greater) from 0.17 to 0.78 over the VDT year. Group 1-2 (64.2%; n = 43) started VDT at a higher point and increased their probability from 0.48 to 0.90.

However, the difference in probability between the two groups had reduced by the end of VDT. Overall, group 1-2 produced the “best” performance.

Narrow confidence intervals are seen around the estimated group probabilities for each trajectory.

Cohort 2 - Threshold = 6; Minimum number VDPs per group = 5 and 10

Figure 6.12 presents the group trajectories for model 2 3 1 3. This model was selected to represent cohort 2’s LEP data if there was to be a minimum of 5 and 10 VDPs per group and a threshold score of 6 was adopted.

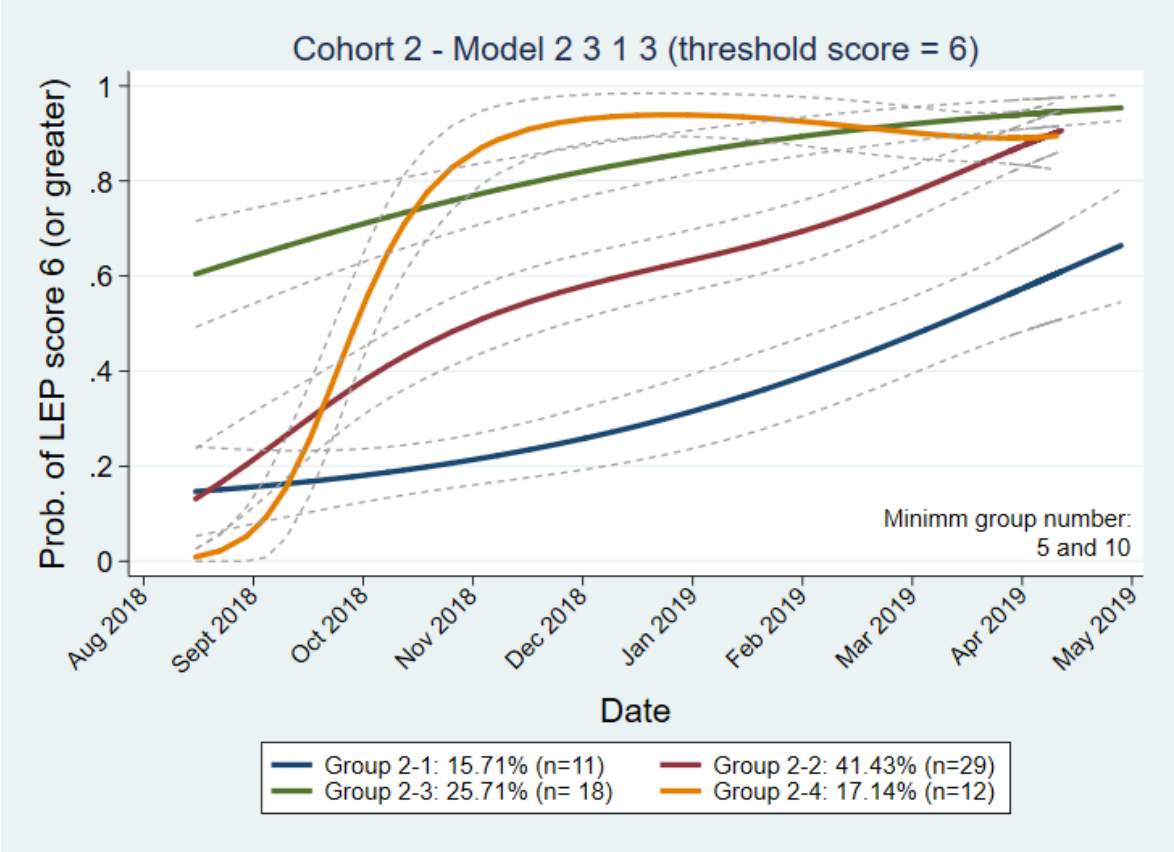


Figure 6.12 - Cohort 2: Trajectory groups for Bernoulli distribution model 2 3 1 3. NOTE: 95% confidence intervals around estimated group probabilities are depicted by the dotted lines around each trajectory.

An overall increase in the probability of scoring ≥6 was observed in all four trajectory groups over the VDT year. The largest increase was observed in group 2-4 (17.1%; n = 12), where probability increased from 0.00 at the beginning of VDT to 0.90 at the end. A similar probability at the end of VDT was seen in group 2-2 (41.4%; n = 29), however this group produced a slight cubic shaped trajectory showing a more gradual increase from a starting probability of 0.15. Group 2-1

(quadratic) (15.7%; n = 11) also had a starting probability of 0.15 but did not show as much of an increase as group 2-2 – ending the VDT year with a 0.65 chance of scoring ≥6. Group 2-3 (25.7%; n = 18) followed a linear trajectory and saw an increase in their probability of scoring ≥6 from 0.60 to 0.95. Based on the trajectory patterns presented and the criteria outlined previously in chapter 5 (section 5.4.2), group 2-3 or group 2-4 were jointly considered as the "best performing".

Confidence intervals around the estimated group probabilities of each trajectory were notably wider in this model compared to other GBTMs selected in this chapter. However, they were still reasonably narrow and none of the other models in cohort 1 which were considered for a threshold score of 6 with a minimum of 5 VDPs per trajectory group had narrower confidence intervals.

Cohort 2 - Threshold = 6; Minimum number VDPs per group = 15 and 20

Figure 6.13 presents the group trajectories for model 3 2. This model was selected to represent cohort 2's LEP data if minimum group number restrictions of 15 and 20 VDPs per group were applied and the threshold score was set at 6.

Both groups increased their probability of scoring ≥6 over the duration of VDT. Group 2-1 (41.4%; n = 29) followed a slight cubic trajectory where their probability increased from 0.10 to 0.85. Group 2-2 (58.6%; n = 41) produced a quadratic trajectory where their probability increased from 0.35 to 0.90. This model appeared similar to the 1 1 model chosen for cohort 1 (Figure 6.11). However, although cohort 2's model demonstrated a greater divergence in probabilities between the two groups in the middle of the VDT, there was a greater convergence towards the end of year. Group 2-2 were the "best performing" overall.

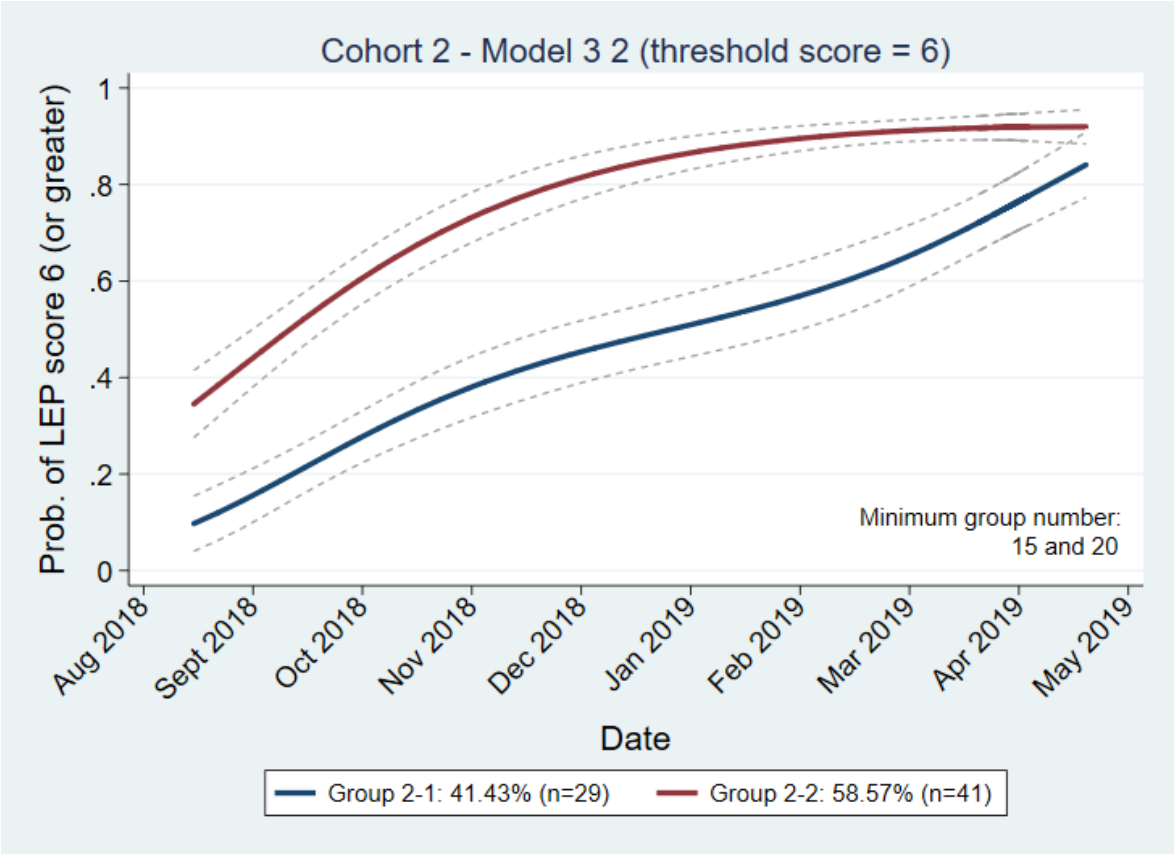


Figure 6.13 - Cohort 2: Trajectory groups for Bernoulli distribution model 3 2. NOTE: 95% confidence intervals around estimated group probabilities are depicted by the dotted lines around each trajectory.

Each trajectory in this model displayed tight confidence intervals around their estimated group probabilities.

Cohort 3 - Threshold = 6; Minimum number VDPs per group = 5

Figure 6.14 presents the group trajectories for model 3 0 3. This GBTM was the most suitable model if each trajectory group was required to have a minimum of 5 VDPs and a threshold LEP score of 6 was adopted.

Groups 3-1 and 3-3 increased their probability of scoring ≥6 over the duration of VDT. Group 1 (45.0%; n = 27) followed a slight cubic trajectory where their probability increased from 0.12 to 0.85. Group 3-3 (45.0%; n = 27) also produced a cubic trajectory where their probability increased from 0.36 to 0.92. A zero-order trajectory was displayed by group 3-2 (10.0%; n = 6), illustrating no change in their probability of scoring >6 during VDT. Out of all three groups, group 3-3 were regarded as the “best performing”.

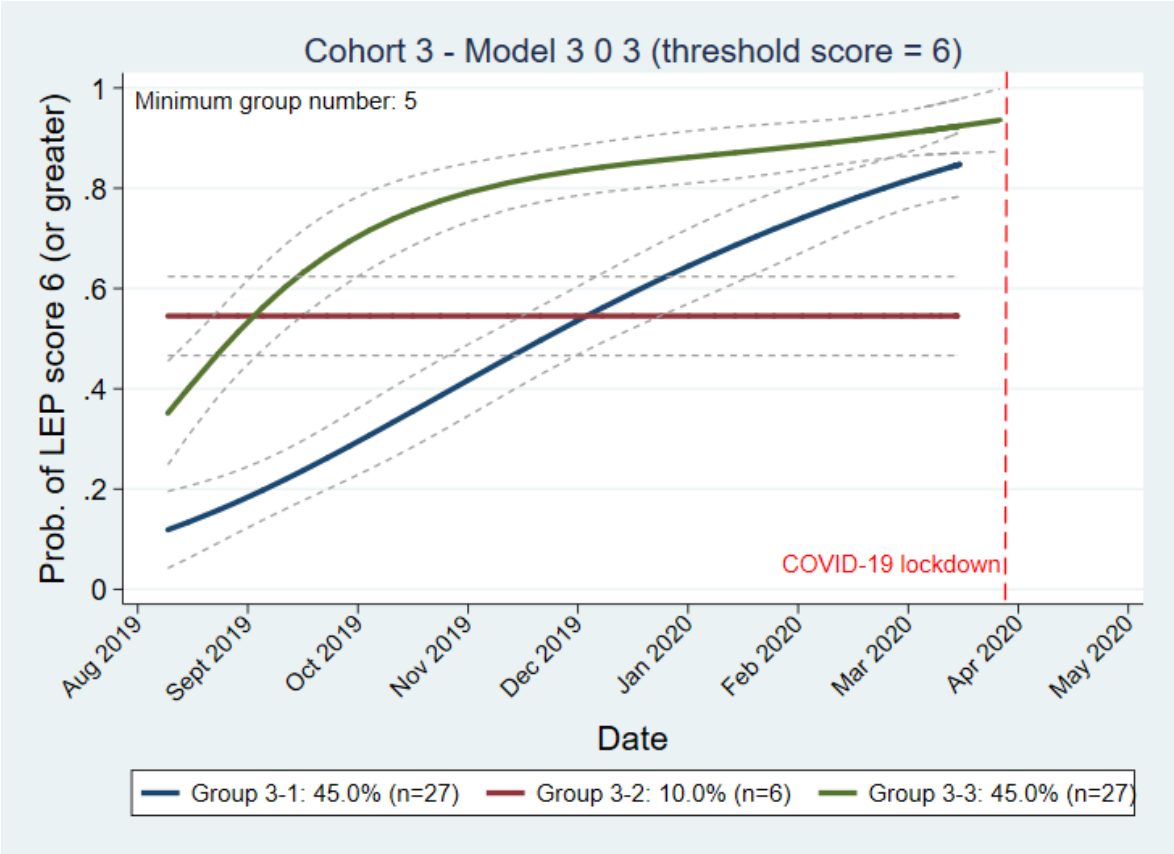


Figure 6.14 - Cohort 3: Trajectory groups for Bernoulli distribution model 3 0 3. NOTE: 95% confidence intervals around estimated group probabilities are depicted by the dotted lines around each trajectory.

Tight confidence intervals around the estimated group probabilities for groups 3-1 and 3-3 were observed. Group 3-2 displayed a slightly wider confidence interval by comparison.

Cohort 3 - Threshold = 6; Minimum number VDPs per group = 10, 15 and 20

The group trajectories for model 1 3 are shown in Figure 6.15. This GBTM was the most suitable model if each trajectory group was required to have a minimum of 10, 15 and 20 VDPs and a threshold LEP score of 6 was adopted.

Both groups 3-1 and 3-2 displayed increasing linear trajectories. Group 3-1 (51.7%; n = 31) demonstrated an increase in probability of scoring $\geq$6 from 0.20 to 0.81 over the VDT year. Group 3-2 (48.33%; n = 29) started VDT at a higher point and increased their probability from 0.38 to 0.90. The difference in probability between the two groups had reduced by the end of VDT. Overall, group 3-2 was identified as the “best performing”.

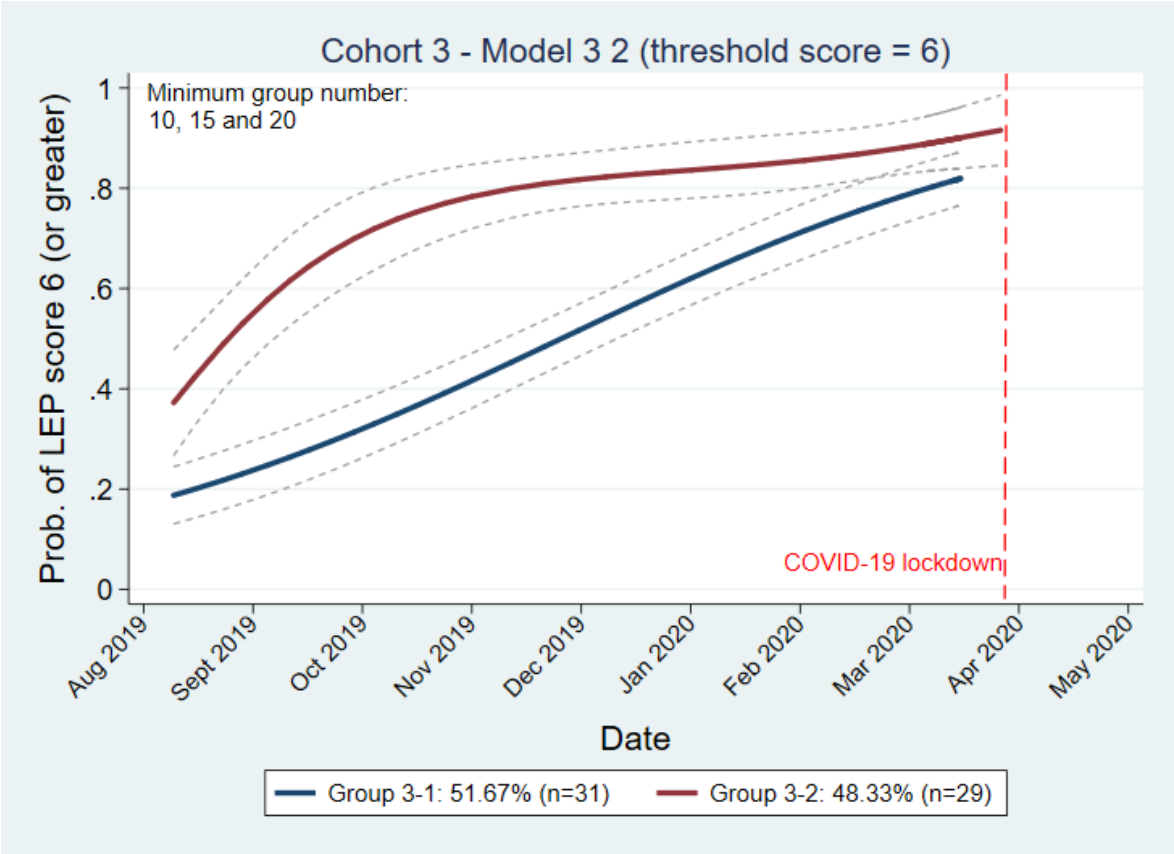


Figure 6.15 - Cohort 3: Trajectory groups for Bernoulli distribution model 1 3. NOTE: 95% confidence intervals around estimated group probabilities are depicted by the dotted lines around each trajectory.

Narrow confidence intervals are seen around the estimated group probabilities for each trajectory.

6.4.2.4 Threshold score = 7

A score of 7 is the lowest of the three "superior" scores which can be subjectively awarded by assessors using the LEP rating system (see chapter 3, section 3.5.3.2).

Like for threshold 6 investigations, there were multiple incidences where the difference in BIC between the model with the highest BIC and other models was found to be less than two. Therefore, these models were subjected to further scrutiny, in which the trajectory graphs produced by each model were compared to determine the most parsimonious models.

Cohort 1

The highest BIC was returned by model 1 1 3, however this was not greater than two when compared to the BIC of six other models (1 3 1 2, 3 1 2, 2 1 2, 3 1 3, 1 2 2 and 1 2 1 2). All these models did not satisfy the criteria of having at least 15 and 20 students in all its trajectory groups if these respective restrictions were implemented.

The parameter estimates and graphical representation of the trajectories appeared similar for all the three-group models (i.e., 1 1 3, 3 1 2, 2 1 2, 3 1 3 and 1 2 2). The four-group models (i.e., 1 3 1 2 and 1 2 1 2) were also alike in terms of their parameter estimates and graphical appearance. When comparing three-group against four-group models, the three-group models exhibited tighter confidence intervals around the estimated group trajectories, indicating greater model adequacy (Nagin, 2005; Nagin and Odgers, 2010). As a result of these considerations, model 1 1 3 was selected to represent cohort 1's LEP data if a threshold score of 7, and minimum group number restrictions of 5 and 10 were in effect.

Model 1 1 had the highest BIC of all models which satisfied minimum group number restrictions of 15 and 20. However, four other models (3 1, 2 1, 1 2, 3 2) had a BIC similar (i.e., a difference no greater than two) when compared to model 1 1. There was very little difference in the appearance if the graphical trajectories and the parameter estimates between these five models and, therefore, any of them could have been selected as the model to represent cohort 1's LEP data (if a threshold score of 7 was implemented). Ultimately, model 1 1 was chosen since it returned the highest BIC.

Cohort 2

There was a clearer choice of model if cohort 2's data were to have a threshold score of 7 and minimum group number restrictions of 5 and 10 implemented. Model 1 1 3 2 returned the highest BIC and it was at least two greater than the model with the second highest. If group restrictions of 15 and 20 were to be adopted, there were two possible models to choose from – models 1 3 and 1 1. However, once again, the trajectories and parameter estimates of both these

models were very similar in terms of graphical appearance and, therefore, no model was preferable to the other. Model 1 3 was chosen since it returned the highest BIC.

Cohort 3

Model 3 1 3 1 returned the highest BIC if minimum group number restrictions of 5 and 10 were applied. However, it did not have a BIC greater than two when compared to model 1 3 0 1, which contained at least 5 VDPs per trajectory group but did not satisfy a minimum group number restriction of 10. Therefore, both models were subjected to further scrutiny to determine if one was more parsimonious than the other (see below).

If minimum group number restrictions of 15 and 20 were applied, then the models with the highest BIC were 1 3 1 and 1 2, respectively. Both these models had a BIC greater than two when compared to those with the second highest BIC (under the same minimum group number restrictions).

A summary of the LEP data GBTMs selected for all three cohorts is shown in Table 6.6. The models listed in Table 6.6 satisfied all tests of model adequacy proposed by Nagin (2005) and Nagin and Odgers (2010). Tables 6.7-6.9 shows the results of these tests for each GBTM selected (according to any implemented threshold scores and minimum group number restrictions). The AvePP was at least 0.88 for each group in all models, i.e., greater than the 0.70 minimum value recommended by Nagin and Odgers (2010). The odds of correct classification for all groups were over 5.0, indicating the model's assignment accuracy was good and there was a difference no greater than 0.04 between each trajectory group's estimated probability of group membership and the proportion of students classified to that group according to posterior probability of group membership across all models.

Table 6.6 – Bernoulli distribution group-based trajectory models (threshold score = 7) selected to represent clinical longitudinal evaluation of performance (LEP) data for cohorts 1, 2 and 3. VDPs = Vocational dental practitioners. BIC = Bayesian information criterion.

Threshold score = 7								
Cohort	**Number of VDPs (n)**	**Number of assessments**	**Model**	**Contains at least X VDPs per group, where X =**				**BIC (based on number of VDPs)**
				5	***10***	***15***	***20***	
1	67	2,294	1 1 3	✓	✓	✗	✗	-1185.98
			1 1	✓	✓	✓	✓	-1204.23
2	70	2,839	1 1 3 2	✓	✓	✗	✗	-1442.14
			1 3	✓	✓	✓	✓	-1484.38
3	60	1,956	3 1 3 1	✓	✓	✗	✗	-1030.88
			1 3 0 1	✓	✗	✗	✗	-1031.20
			1 3 1	✓	✓	✓	✗	-1035.31
			1 2	✓	✓	✓	✓	-1049.64

When comparing models 3 1 3 1 and 1 3 0 1, the latter displayed higher odds of correct classification for each trajectory group compared to the former. As a result, model 1 3 0 1 was selected to represent the data if a minimum group restriction of 5 VDPs was applied, and model 3 1 3 1 was selected to represent the data for a minimum group restriction of 10 VDPs.

Table 6.7 – Cohort 1: Average posterior probabilities, odds of correct classification, estimated probability, and proportion of group membership (according to posterior probability of group membership) for clinical longitudinal evaluation of performance (LEP) trajectory group models (based on Bernoulli data distribution). NOTE: All values are rounded to two decimal places. BIC = Bayesian information criterion.

Cohort (n)	Threshold score	Model	BIC	Group	n per group	Average posterior probability	Odds of correct classification	Estimated probability of group membership	Proportion of students classified to group according to posterior probability of group membership
1 (67)	7	1 1 3	-1185.98	1-1	28	.94	22.84	.41	.42
				1-2	32	.96	23.34	.49	.48
				1-3	7	.96	205.94	.10	.10
		1 1	-1204.23	1-1	30	.97	41.06	.47	.45
				1-2	37	.94	14.07	.53	.55

Table 6.8– Cohort 2: Average posterior probabilities, odds of correct classification, estimated probability, and proportion of group membership (according to posterior probability of group membership) for clinical longitudinal evaluation of performance (LEP) trajectory group models (based on Bernoulli data distribution). NOTE: All values are rounded to two decimal places. BIC = Bayesian information criterion.

Cohort (n)	Threshold score	Model	BIC	Group	n per group	Average posterior probability	Odds of correct classification	Estimated probability of group membership	Proportion of students classified to group according to posterior probability of group membership
2 (70)	7	1 1 3 2	-1442.14	2-1	10	.99	772.18	.15	.14
				2-2	26	.91	19.38	.35	.37
				2-3	10	.96	126.43	.16	.14
				2-4	24	.91	18.89	.34	.34
		1 3	-1484.38	2-1	34	.97	37.79	.47	.49
				2-2	36	1	206.89	.53	.51

Table 6.9 – Cohort 3: Average posterior probabilities, odds of correct classification, estimated probability, and proportion of group membership (according to posterior probability of group membership) for clinical longitudinal evaluation of performance (LEP) trajectory group models (based on Bernoulli data distribution). NOTE: All values are rounded to two decimal places. BIC = Bayesian information criterion.

Cohort (n)	**Threshold score**	**Model**	**BIC**	**Group**	**n per group**	**Average posterior probability**	**Odds of correct classification**	**Estimated probability of group membership**	**Proportion of students classified to group according to posterior probability of group membership**
3 (60)	7	3 1 3 1	-1030.88	3-1	18	.91	24.20	.28	.30
				3-2	20	.93	29.50	.33	.33
				3-3	12	.88	34.82	.18	.20
				3-4	10	.97	118.43	.21	.17
		1 3 0 1	-1031.20	3-1	19	.94	31.58	.32	.32
				3-2	18	.93	30.72	.29	.30
				3-3	7	.88	56.63	.11	.12
				3-4	16	.98	173.59	.27	.27
		1 3 1	-1035.31	3-1	20	.97	52.12	.36	.33
				3-2	18	.90	23.05	.28	.30
				3-3	22	.95	34.07	.36	.37
		1 2	-1049.64	3-1	32	.96	20.31	.53	.53
				3-2	28	.95	22.25	.47	.47

Cohort 1 - Threshold = 7; Minimum number VDPs per group = 5 and 10

The group trajectories for model 1 1 3 are presented in Figure 6.16. This GBTM was selected to represent cohort 1's LEP data if there were to be at least 5 and 10 students per trajectory group and a threshold LEP score of 7.

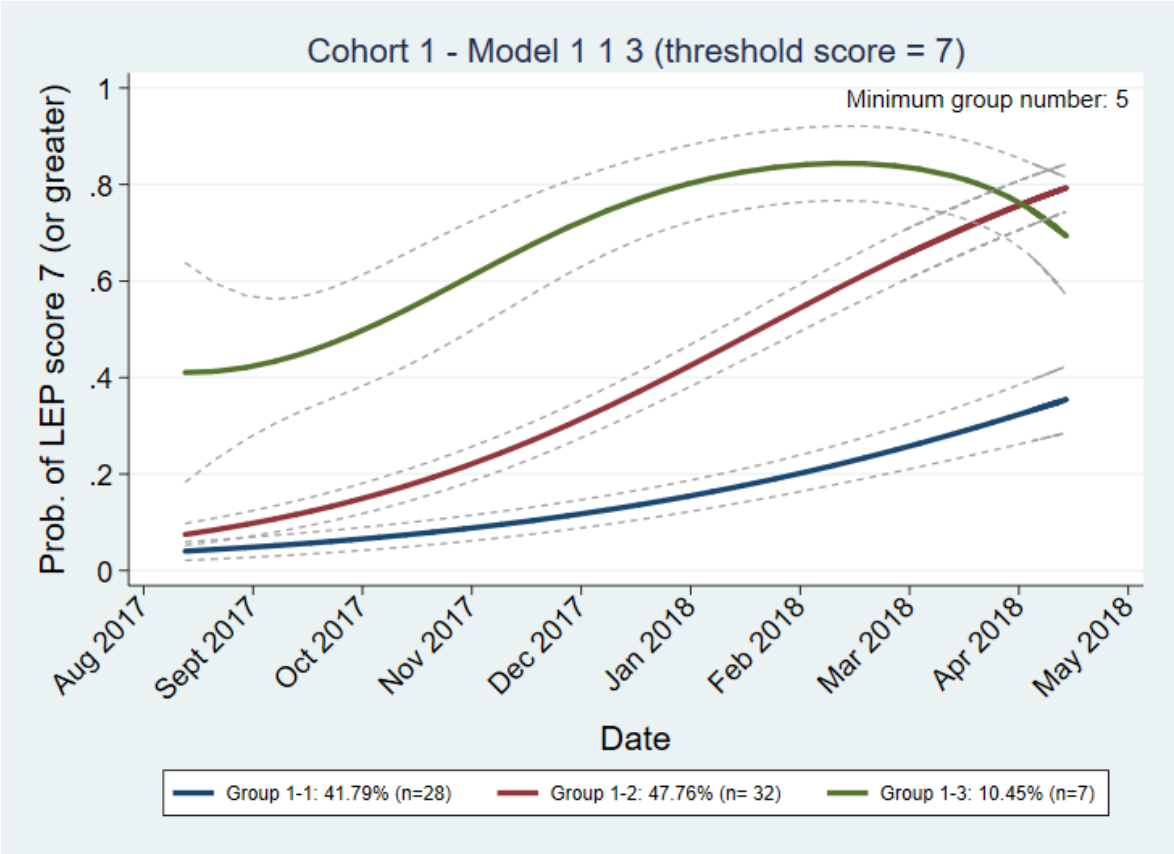


Figure 6.16 - Cohort 1: Trajectory groups for Bernoulli distribution model 1 1 3. NOTE: 95% confidence intervals around estimated group probabilities are depicted by the dotted lines around each trajectory.

Groups 1-1 (41.8%; n = 28) and 1-2 (47.8%; n = 32) demonstrated linear trajectories indicative of increasing probability in achieving a LEP score ≥7 during VDT. However, although both groups began VDT with a similar probability, group 1-2 had a significantly higher chance (~0.79) of achieving a score ≥7 by the end of VDT compared to group 1-1 (~0.30). Group 1-3 (cubic) (10.5%; n = 7) started VDT with the highest probability which increased further (to 0.85) between August (2017) and February (2018). Their probability then decreased to 0.70 over March and April (2018). Despite this dip, group 1-3 were regarded as the "best performing" overall.

Consistently tight confidence intervals were observed around the estimated group probabilities for the trajectories of groups 1-1 and 1-2. The confidence

intervals around group 1-3's trajectory was wider compared to those from groups 1-1 and 1-2 and was especially wide at the beginning of the VDT year. This finding may be due to the smaller percentage of VDPs in group 1-3 (10.5%; n = 7).

Cohort 1 - Threshold = 7; Minimum number VDPs per group = 15 and 20

Group trajectories for the 1 1 model (threshold 7) are shown in Figure 6.17. This GBTM was the most suitable model if each trajectory group was required to have a minimum of 15 and 20 VDPs.

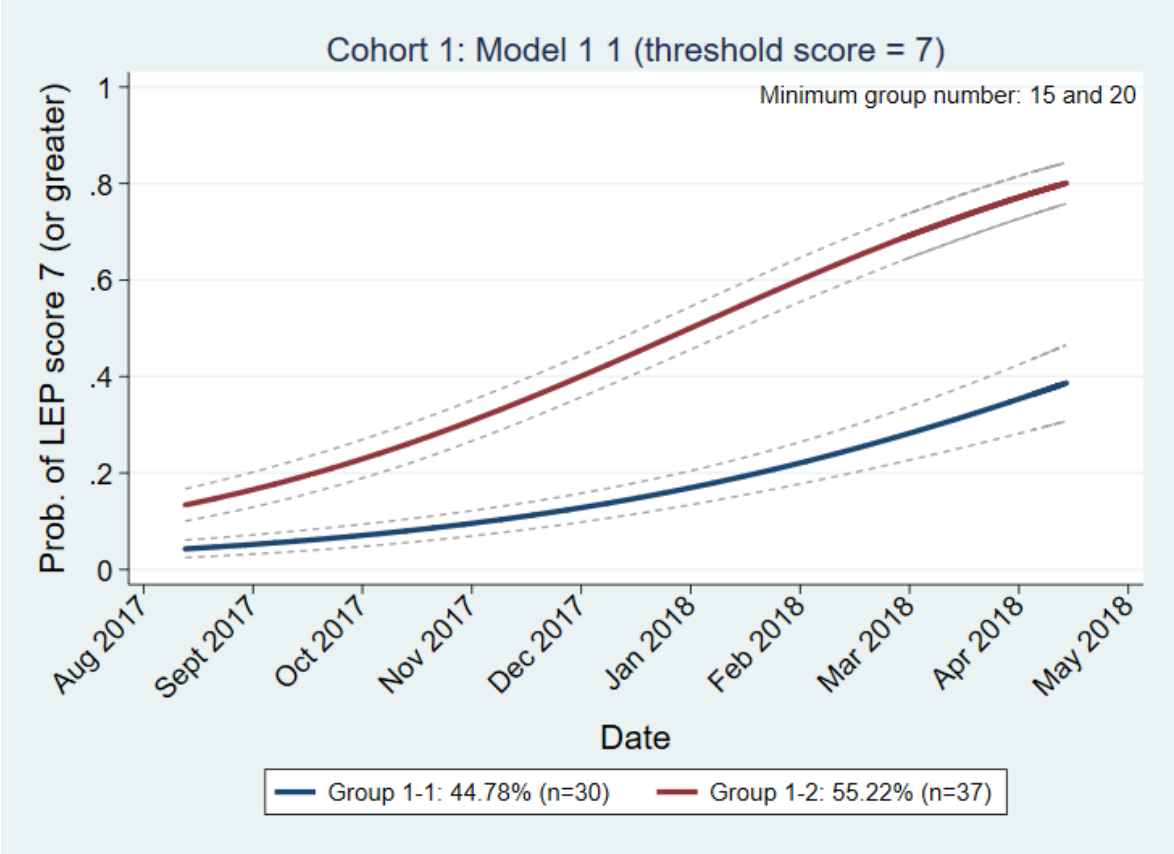


Figure 6.17 - Cohort 1: Trajectory groups for Bernoulli distribution model 1 1 (threshold score = 7). NOTE: 95% confidence intervals around estimated group probabilities are depicted by the dotted lines around each trajectory.

Both groups 1-1 and 1-2 displayed linear trajectories. Group 1-1 (44.8%; n = 30) exhibited an increase in their probability of scoring ≥7 from <0.1 to 0.4 over the VDT year. Group 1-2 (55.2%; n = 37) displayed a more accelerated increase in probability from 0.15 to 0.80. Overall, group 1-2 were the "best performing".

Tight confidence intervals were observed around the estimated group probabilities for the trajectories of groups 1-1 and 1-2.

Cohort 2 - Threshold = 7; Minimum number VDPs per group = 5 and 10

Figure 6.18 presents the trajectories for model 1 1 3 2, which was selected to represented cohort 2's LEP data if minimum group member restrictions of 5 and 10 were imposed along with a threshold score of 7.

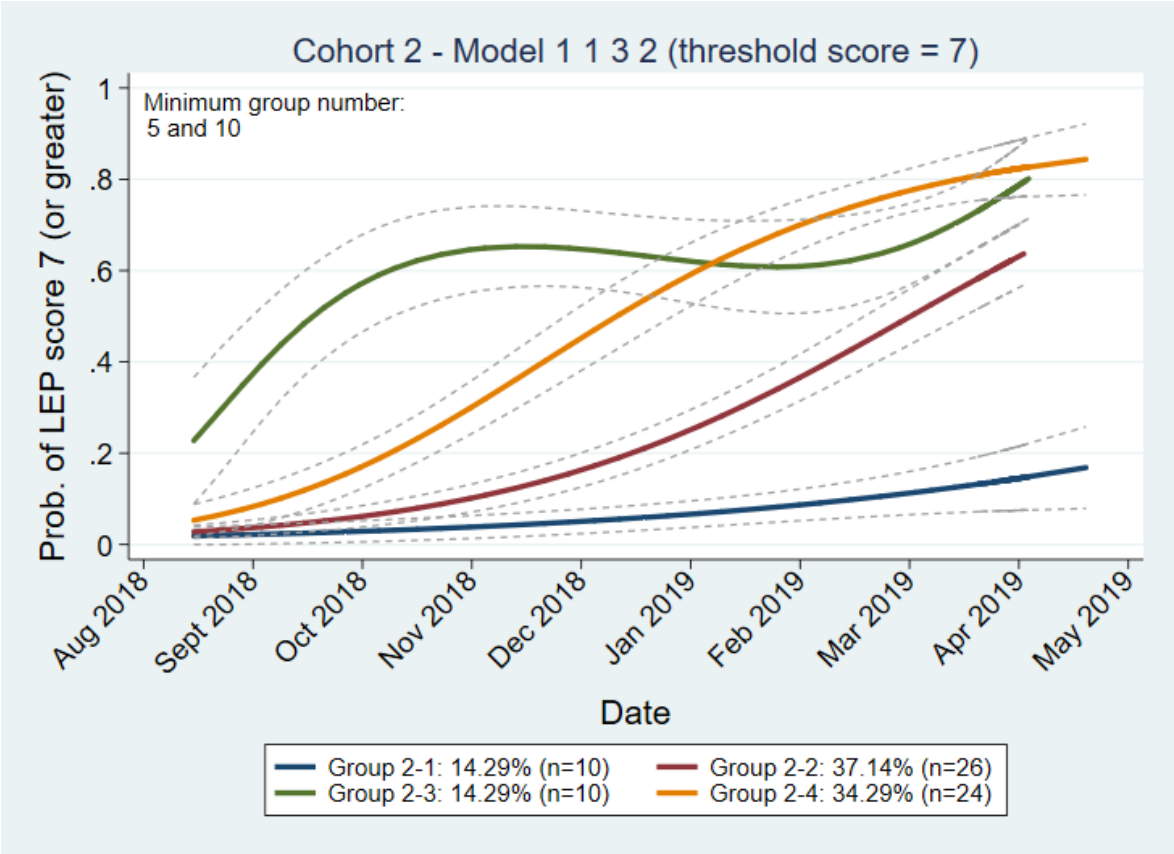


Figure 6.18 - Cohort 2: Trajectory groups for Bernoulli distribution model 1 1 3 2. NOTE: 95% confidence intervals around estimated group probabilities are depicted by the dotted lines around each trajectory.

All four trajectory groups demonstrated an overall increase in the probability of scoring ≥7 over the VDT year. The linear trajectories produced by Groups 2-1 (14.3%; n = 10) and 2-2 (37.1%; n = 26) both originated from a probability of 0.10 at the beginning of VDT However the group 2-1's probability showed very little increase (to 0.18) by the end of the VDT year. In contrast, group 2-2's probability increased to 0.63. Group 2-4 (34.3%; n = 24) also had a low probability at the beginning of VDT (0.03) but eventually displayed the highest probability by the end of the year (~0.85). Group 2-3 (14.3%; n = 10) also had a high probability by the end of VDT (~0.81), although they initially started VDT at 0.21 and followed a cubic trajectory. Overall, group 2-3 were regarded as the "best performing".

Narrow confidence intervals around the estimated group probabilities for the trajectories were observed in groups 2-1, 2-2 and 2-4. However, it was noted that these confidence intervals widened slightly towards the end of the VDT year. Wider confidence intervals were found around the trajectory of group 2-3. This finding may be due to the smaller percentage of VDPs in group 2-3 (14.3%; n = 10).

Cohort 2 - Threshold = 7; Minimum number VDPs per group = 15 and 20

The final GBTM selected in cohort 2 (model 1 3) is illustrated in Figure 6.19. This model was selected to represent cohort 2's LEP data if there was to be a minimum of 15 and 20 VDPs per group and a threshold LEP score of 7 was adopted.

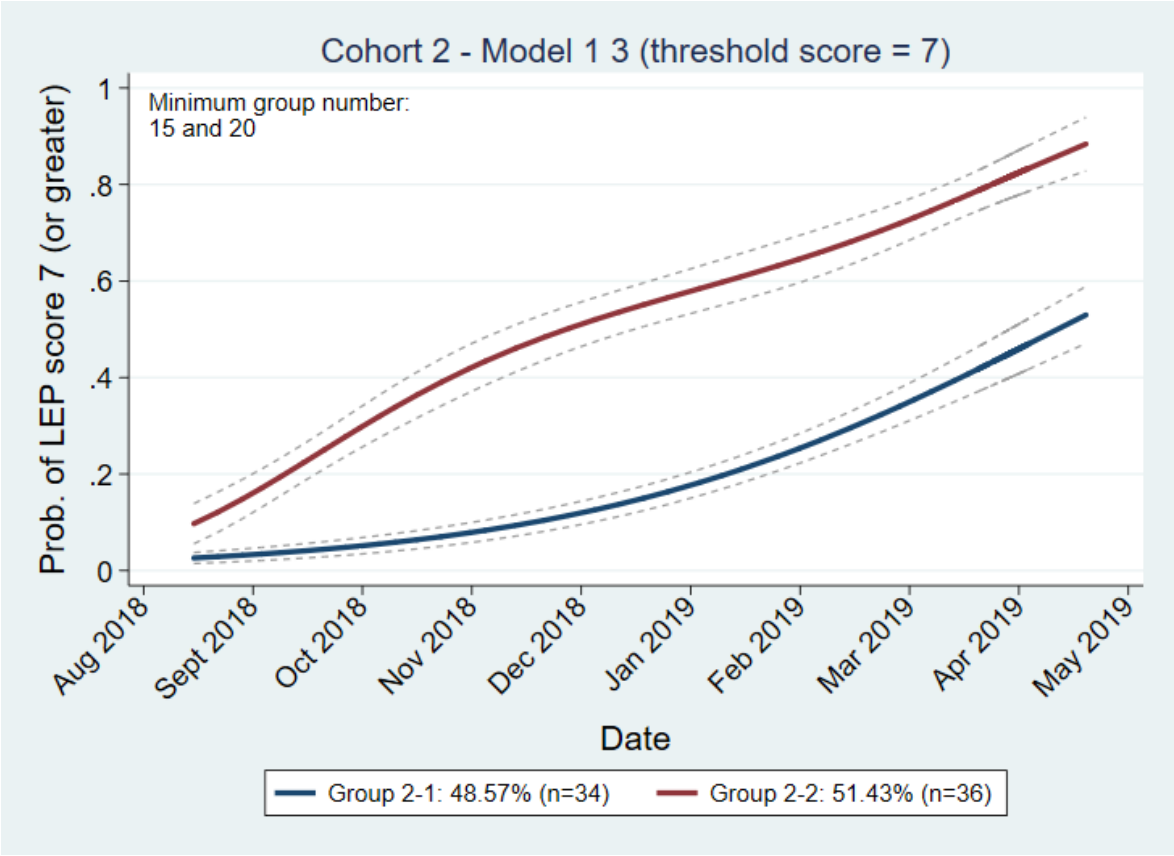


Figure 6.19 - Cohort 2: Trajectory groups for Bernoulli distribution model 1 3. NOTE: 95% confidence intervals around estimated group probabilities are depicted by the dotted lines around each trajectory.

Both groups increased their probability of scoring ≥7 over the duration of VDT. However, the trajectory of group 2-2 (51.4%; n = 36) displayed a greater and more rapid increase in development than group 2-1 (48.6%; n = 34). By the end the VDT year, group 2-2 had almost a 40% higher chance of achieving a score ≥7

compared to group 2-1 despite there being little difference in the starting probability between both groups (group 2-1 = 0.01|group 2-2 = 0.10). Therefore, group 2-2 were identified as the "best performing" overall. This model was similar to the 1 1 model chosen for cohort 1 (Figure 6.17). However, although the groups in cohort 2's 1 3 model appeared to perform poorer over the duration of VDT compared to their cohort 1 counterparts.

Each trajectory in this model displayed tight confidence intervals around their estimated group probabilities.

Cohort 3 - Threshold = 7; Minimum number VDPs per group = 5

Figure 6.20 presents the trajectories for model 1 3 0 1, which was selected to represented cohort 3's LEP data when a minimum group member restriction of 5 was imposed along with a threshold score of 7.

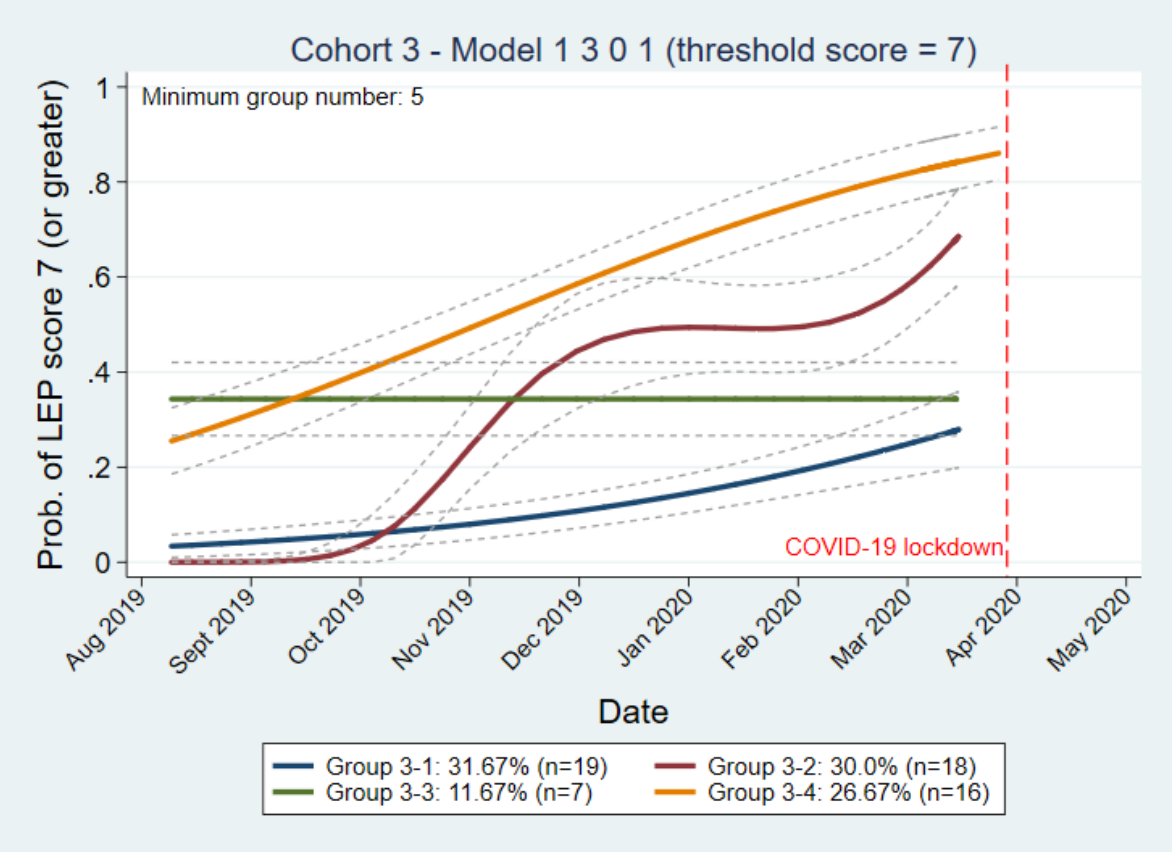


Figure 6.20 - Cohort 3: Trajectory groups for Bernoulli distribution model 1 3 0 1. NOTE: 95% confidence intervals around estimated group probabilities are depicted by the dotted lines around each trajectory.

Three VDP trajectory groups increased their probability of scoring ≥7 over the VDT year. Group 3-1 (linear; 38.3%; n = 23) and group 3-2 (cubic; 30.0%; n = 18)

both increased their probability from <0.05 to 0.27 and 0.66, respectively. Group 3-4 (linear; 26.7%; n = 16) also increased their probability from 0.27 to 0.82 and were regarded as the "best performing" overall. Group 3-3 (11.67%; n = 7) followed a zero-order trajectory and maintained a probability of 0.37 throughout the duration of the VDT year.

Tight confidence intervals around the estimated group probabilities were observed for groups 3-1 and 3-4. Wider confidence intervals were observed from groups 3-2 and 3-4.

Cohort 3 - Threshold = 7; Minimum number VDPs per group = 10

Model 3 1 3 1, which was selected to represent cohort 3's LEP data if a minimum group number restriction of 10 was applied, is shown in Figure 6.21.

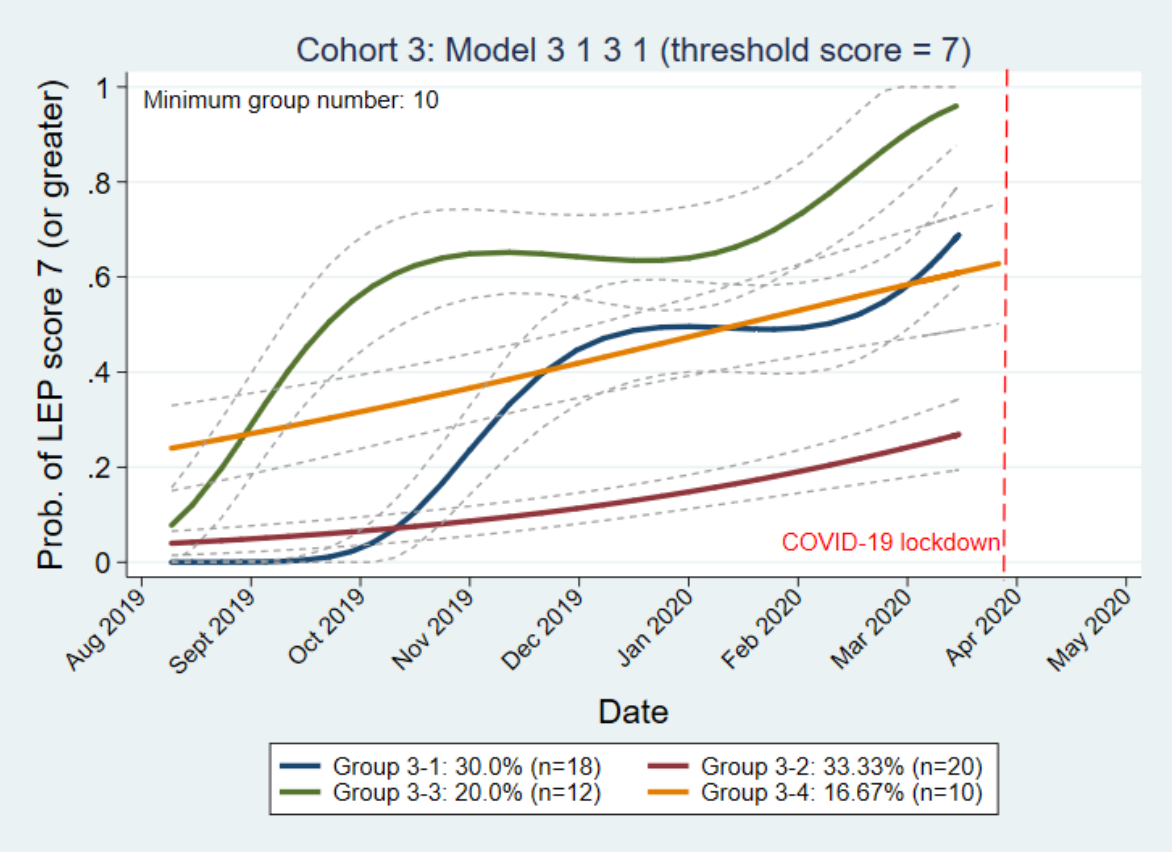


Figure 6.21 - Cohort 3: Trajectory groups for Bernoulli distribution model 3 1 3 1. NOTE: 95% confidence intervals around estimated group probabilities are depicted by the dotted lines around each trajectory.

All trajectory groups showed an increase in probability of scoring ≥7 over the course of the VDT year. Groups 3-1 (cubic; 30.0%; n = 18), 3-2 (linear; 33.3%; n = 20) and 3-3 (cubic; 20.0%; n = 12) all began VDT with a probability <0.1 before

increasing their probability to 0.65, 0.22 and 0.95 (respectively) – with group 3-3 demonstrating the largest increase. Group 3-4 (linear; 16.7%; n = 10) started VDT with the highest probability (~0.23) out of all the groups but finished with the third highest (~0.60). Overall, group 3-3 were the "best performing".

Narrow confidence intervals were observed around estimated group probability for group 3-2. Wider confidence intervals were observed for groups 3-1, 3-3 and 3-4.

Cohort 3 - Threshold = 7; Minimum number VDPs per group = 15

Figure 6.22 presented the trajectories for model 1 3 1, which was selected to represent cohort 3's LEP data if at least 15 VDPs were required per trajectory group.

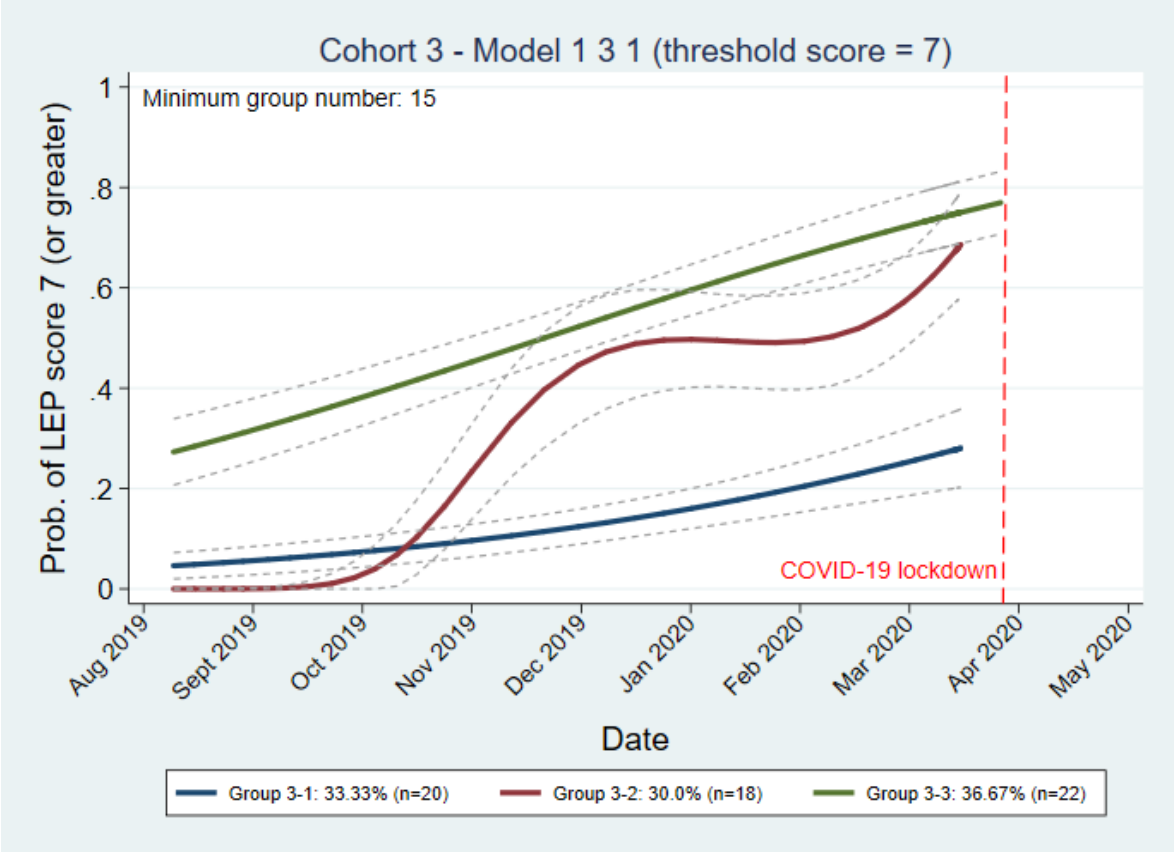


Figure 6.22 - Cohort 3: Trajectory groups for Bernoulli distribution model 1 3 1. NOTE: 95% confidence intervals around estimated group probabilities are depicted by the dotted lines around each trajectory.

All three groups improved their probability of scoring ≥7. Groups 3-1 (linear; 33.3%; n = 20) and 3-2 (cubic; 30.0%; n = 18) both increased their probability from <0.1 to 0.25 and 0.65, respectively. Group 3-3 (36.7%; n = 22)

demonstrated a linear increase in their probability from 0.29 to 0.78 and were the “best performing”.

Narrow confidence intervals were observed around estimated group probability for groups 3-1 and 3-3. Confidence intervals were observed for group 3-2 at the beginning of the VDT year, however they appeared wider from the middle to the end of the VDT year.

Cohort 3 - Threshold = 7; Minimum number VDPs per group = 20

Figure 6.23 illustrates the trajectories for model 1 2 – the model selected to represent cohort 3’s LEP data if a minimum group number restriction of 20 was applied.

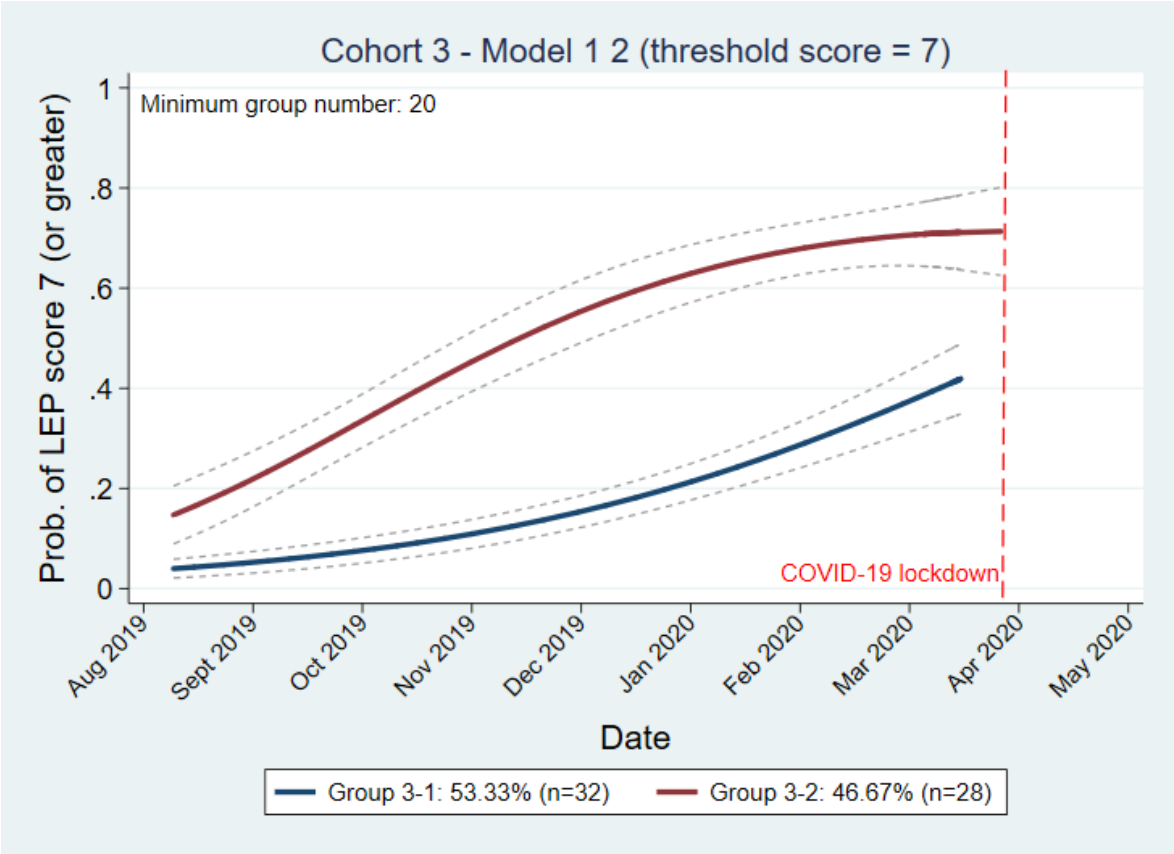


Figure 6.23 - Cohort 3: Trajectory groups for Bernoulli distribution model 1 2. NOTE: 95% confidence intervals around estimated group probabilities are depicted by the dotted lines around each trajectory.

Both groups improved their probability of scoring ≥7. Group 3-1 (53.3%; n = 32) produced a linear trajectory tracking an increase in probability from 0.05 to 0.41, whereas Group 3-2 (46.7%; n = 28) demonstrated a quadratic trajectory

which illustrated an increase in their probability from 0.15 to 0.70. Overall, group 3-2 performed “best”.

Narrow confidence intervals were observed around estimated group probability for both groups. However, the confidence intervals for group 3-2 appeared to widen towards the end of the VDT year.

6.5 Summary

This chapter has reported on data descriptions and GBTMs based on postgraduate longitudinal clinical assessment of three VDP cohorts who completed Scottish VDT schemes between 2017-18 and 2019-20.

There was a wide range in the number of clinical assessments completed by VDPs in both cohorts. Some VDP’s clinical skills were assessed less than 42 times (i.e., the minimum number of LEP assessments required) whereas others were assessed more. The distribution of LEP scores awarded was similar in both cohorts and the most awarded score in both cohorts was 6.

Trajectory models were successfully generated from LEP data using Bernoulli data distributions. It was established that using the two lowest “satisfactory” LEP scores (4 and 5) as a threshold for “competence” did not sufficiently distinguish different patterns of VDP performance across all cohorts. Differing VDP trajectories were detected if cut off scores $\geq$6 were adopted.

LEP data generated by all cohorts produced multiple models of equally good fit, suggesting a lack of stability in LEP scoring. This observation was not exclusive to any specific cut off score (6 or 7) or minimum group number restriction (5, 10, 15 and 20 VDPs) and, therefore, the number models to be used for comparisons with LIFTUPP© data could not be refined based on these criteria.

As a result, all threshold 6 and 7 models selected in this chapter will be used for comparisons with GBTMs generated (based on Bernoulli data distributions) from LIFTUPP© data (chapter 5) to investigate the predictive subtype of criterion validity of longitudinal clinical assessment (chapter 7). These comparisons,

however, may also signify if some models (both LIFTUPP© and LEP) are more stable than others based on minimum group number restrictions (see chapter 7).

The implications of the findings presented in this chapter and their subsequent comparisons with undergraduate examination and LIFTUPP© data are discussed in greater detail in chapter 9.

Chapter 7 - Exploring criterion validity of undergraduate longitudinal clinical assessment: associations with undergraduate examination outcomes and postgraduate longitudinal clinical performance

7.1 Introduction

The following chapter presents the evidence generated to investigate the criterion validity of longitudinal clinical assessment and builds towards a validity argument for its use in assessing the development of clinical competence.

Evidence is generated from comparisons between LIFTUPP© data (previously analysed in chapter 5) and undergraduate examination results (previously analysed in chapter 4) and postgraduate LEP data (previously analysed in chapter 6). Results of reliability testing for the panel of assessments used within each undergraduate BDS year (longitudinal clinical assessment and undergraduate examinations) are also presented.

7.2 Aim

To investigate the criterion validity of longitudinal clinical assessment of undergraduate dental students by comparing it with previously validated and reliably tested outcomes of undergraduate BDS degree examinations and postgraduate longitudinal assessment.

Specifically, comparison with undergraduate examinations will investigate the concurrent validity (a subtype of criterion validity) for longitudinal clinical assessment of undergraduate dental students and aims to address research question 2b: What is the association between undergraduate longitudinal clinical assessment and standalone assessment methods? – (chapter 2, section 2.2).

Comparisons with postgraduate LEPs will investigate the predictive validity (another subtype of criterion validity) and aims to address research question 2c: What is the association between undergraduate longitudinal clinical assessment and postgraduate assessment? – (chapter 2, section 2.2).

7.3 Method

LIFTUPP© GBTM data (chapter 5) were linked together with undergraduate examination analyses (chapter 4) and LEP GBTM data (chapter 6) using the linkage processes previously described in chapter 3 (section 3.5.3.3). Linked data sets for each cohort were saved as Stata (v.15.1) statistical software files.

7.3.1 Statistical analysis: LIFTUPP© vs undergraduate examination results (concurrent validity)

Generation of GBTMs for LIFTUPP© data and the subsequent results have previously been described in chapters 3 (section 3.5.3.6) and 5 (section 5.4.2), respectively. Chapter 5 (section 5.4.3) also summarised that GBTM based on Bernoulli data distributions (i.e., threshold models) GBTMs were more appropriate and preferable for modelling LIFTUPP© data than those based on censored normal distributions.

The results in chapter 5 also showed that Bernoulli models which used a threshold LIFTUPP© performance indicator of 4 only produced a single trajectory. Whilst these models were used to investigate content validity, they did not discriminate different groups of student clinical performance, which was required to investigate criterion validity (see chapter 5, section 5.4.2.2). Models which used a LIFTUPP© performance indicator of 5 were able to provide the necessary distinction between different student groups (see chapter 5, section 5.4.2.3).

For these reasons, LIFTUPP© data Bernoulli models with a threshold score of 5 (previously outlined in chapter 5 (section 5.4.2.3)) were used for comparisons between LIFTUPP© and undergraduate examination and LEP data to investigate criterion validity.

To recap, two models (3 2 and 0 1) were selected for cohort 1, and one model was selected for each of cohort 2 (model 1 3) and cohort 3 (model 3 2). The first LIFTUPP© trajectory group was deemed as the better performing group for cohort 1's 3 2, cohort 2's 1 3 and cohort 3's 3 2 models. In model 0,1, the second trajectory group performed best (see chapter 5, section 5.4.2.3).

Research question 2b was addressed in two ways. Firstly, the association between LIFTUPP© trajectory group membership and mean aggregated examination scores for all BDS years was assessed using independent sample (student) t-tests. Secondly, the association between LIFTUPP© trajectory group membership and a "top 20%" aggregated performance across each of the BDS1-5 examinations was assessed using cross tabulations with Fisher's exact tests. The probability of trajectory group membership (for each trajectory) per participant – which was required for these two analytical approaches - was calculated as part of the GBTM generation process by the traj plugin for Stata® statistical software (see chapter 3 section 3.5.3.6). A fictional example of this statistical output is shown in Table 7.1, where student 42 has a 0.99 (i.e., 99.7%) probability of belonging in group 1-2 of model 3 1 3 2 based on their individual clinical performance data.

Table 7.1 – Example of the traj plugin statistical output detailing the probability of student membership to each group within a group-based trajectory model (GBTM).

			Probability of membership			
Model	***Student ID***	***Assigned group***	***Group 1-1***	***Group 1-2***	***Group 1-3***	***Group 1-4***
3 1 3 2	42	1-2	.0000889	.9968286	.0030825	0

The rational for comparing trajectory group memberships with fifths of examination performance has previously been described in chapter 4 (section 4.4.3).

It should be noted that BDS1/2 examinations took place before student clinical work was assessed using LIFTUPP©. In BDS3/4/5, examinations and LIFTUPP© assessments occurred within the same academic years (see chapter 3, section 3.5.3.2). Examinations were scheduled for the end of each academic year (except for the BDS3 anatomy examination) and LIFTUPP© was used throughout the duration of the BDS3/4/5 years.

As per chapter 3 (section 3.5.3.1), although t-tests and Fisher's exact tests were performed as part of the investigations, and the results are presented in sections 7.4.1 and 7.4.2 below, caution must be taken when considering the p-values due to sample size limitations.

7.3.2 Statistical analysis: Reliability

Chapter 3 (section 3.5.3.7) described how the reliability of LIFTUPP© was investigated using Cronbach's alpha coefficient (Cronbach, 1951). These investigations form part of the approach for constructing a validity argument (see chapter 3, section 3.5.1.3).

7.3.3 Statistical analysis: LIFTUPP© vs longitudinal evaluations of performance (predictive validity)

Generation of GBTMs for LEP data and the subsequent results have previously been described in chapters 3 (section 3.5.3.6) and 6 (section 6.4.2), respectively.

Since Bernoulli distribution models were proposed as more appropriate and parsimonious for LIFTUPP© data, this chapter will only compare the Bernoulli models selected for both LEP and LIFTUPP©. Therefore, the same two-group LIFTUPP© data models that were selected for comparison with the undergraduate examination results in section 7.3.1 were compared with the LEP data.

The LEP models previously outlined in chapter 6 depended on the threshold score chosen and any imposed minimum group membership restrictions. The use of threshold scores 4 and 5 was previously rejected as they did not adequately distinguish VDP performance (see chapter 6, sections 6.4.2.1 and 6.4.2.2). Therefore, the LEP models selected for comparisons with LIFTUPP© data were those which had adopted 6 and 7 as the threshold (see chapter 6, sections 6.4.2.3 and 6.4.2.4).

To recap the LEP models (based on Bernoulli data distributions) selected for each cohort were:

- Cohort 1: 3 3 3 2 and 1 1 (for threshold = 6); 1 1 3 and 1 1 (for threshold = 7).

- Cohort 2: LEP - 2 3 1 3 and 3 2 (for threshold = 6); 1 1 3 2 and 1 3 (for threshold = 7).

- Cohort 3: LEP – 3 0 3 and 1 3 (for threshold = 6); 1 3 0 1, 3 1 3 1, 1 3 1 and 1 2 (for threshold = 7).

Research question 2c was addressed through cross tabulations between trajectory group memberships for each GBTM selected to represent each cohort's LIFTUPP© and LEP data. GBTMs were cross tabulated according to any minimum group number restrictions imposed. For example, the model selected to represent a cohort's LIFTUPP© data - when a minimum trajectory group number of 5 students per group was applied - was only cross tabulated against the model selected to represent LEP trajectory data chosen under the same minimum group number restriction. Like for LIFTUPP© data, predicted LEP trajectory group memberships for each VDP were produced as part of the statistical output for GBTMs generated by the traj plugin.

As per chapter 3 (section 3.5.3.1), Fisher's exact tests were performed as part of the investigations but, due to sample size limitations, caution must be taken with interpretation of their returned p-values.

7.4 Results

7.4.1 Association between mean undergraduate BDS examination scores and LIFTUPP© trajectories (Bernoulli/threshold) – concurrent validity

Table 7.2 presents the mean aggregated scores for the BDS1-5 examinations according to LIFTUPP© trajectory groups (based on Bernoulli data distribution with a threshold score of 5).

There were no clear and consistent differences in mean aggregated examination scores between LIFTUPP© trajectory groups for all BDS examinations in all three cohorts. The largest difference in examination performance between trajectory groups was seen in BDS1 for cohort 1, where group 1-1 from model 0 1 scored, on average, 4.3% higher than group 2 (64.4% vs. 68.6%). A similar difference was observed for model 3 2. For both models, higher scores were observed for the better performing trajectory. Details on how better performing trajectories were identified have previously been provided in chapter 5 (section 5.4.2).

Similarly, there were only very small differences in mean aggregated examination scores for BDS2/3/4/5 between trajectory groups for all three cohorts. The greatest differences were observed in BDS4 (3.2%) and BDS5 (2.4%) for cohort 2. The results returned showed the best performing LIFTUPP© trajectory groups from each model scored, on average, higher in the BDS examinations for each year. Two exceptions to this observation were found in BDS3 for both models selected to represent cohort 1's LIFTUPP© data, where the second-best performing trajectory groups returned the higher mean aggregated examination scores. A further exception was seen in BDS4 for cohort 3.

Figures 7.1–7.4 provide illustrated summaries of mean aggregated BDS examination scores per LIFTUPP© trajectory group.

Table 7.2 – Summary data for aggregated BDS1-5 examination scores per trajectory group in threshold group-based trajectory models (GBTMs).

Cohort	Model	Trajectory group	n	Academic BDS Year									
				BDS1		*BDS2*		*BDS3*		*BDS4*		*BDS5*	
				Mean (%); Standard deviation (%)	Minimum score (%); Maximum score (%)	Mean (%); Standard deviation (%)	Minimum score (%); Maximum score (%)	Mean (%); Standard deviation (%)	Minimum score (%); Maximum score (%)	Mean (%); Standard deviation (%)	Minimum score (%); Maximum score (%)	Mean (%); Standard deviation (%)	Minimum score (%); Maximum score (%)
1	3 2	**1-1***	63	**68.38**; 6.95	53.30; 86.07	**75.14**; 5.68	62.92; 87.32	72.51; 6.98	58.79; 88.21	74.99; 6.33	60.88; 89.38	71.47; 4.51	62.86; 81.43
		1-2	17	64.32; 4.78	53.64; 72.53	74.87; 3.37	53.64; 78.83	**73.61**; 6.00	59.54; 81.14	**76.95**; 4.57	66.75; 83.75	**72.81**; 3.55	65.61; 78.21
		p-values (t-test)		0.03		0.86		0.56		0.24		0.26	
	0 1	**1-1**	21	64.35; 5.95	53.30; 79.32	74.53; 4.19	53.30; 79.32	**73.12**; 6.04	59.54; 84.71	**76.52**; 5.30	64.50; 84.75	**72.56**; 3.89	65.61; 79.66
		1-2*	59	**68.64**; 6.67	53.90; 86.07	**75.28**; 5.61	62.92; 87.32	72.62; 7.04	58.79; 88.21	75.01; 6.26	60.88; 89.38	71.47; 4.48	62.86; 81.43
		p-values (t-test)		0.01		0.58		0.77		0.33		0.33	
2	1 3	**2-1***	21	**73.68**; 6.82	58.07; 89.09	**78.19**; 5.57	65.80; 87.64	**78.48**; 6.39	69.95; 89.79	**74.70**; 5.61	66.25; 84.63	**73.53**; 5.22	61.32; 81.61
		2-2	65	71.16; 7.24	53.26; 88.57	76.72; 5.08	66.76; 87.75	75.56; 7.41	60.50; 91.67	71.55; 5.39	57.50; 82.50	71.18; 4.20	61.93; 80.37
		p-values (t-test)		0.16		0.26		0.11		0.02		0.04	
3	3 2	**3-1***	43	**76.20**; 7.48	59.42; 91.00	**75.43**; 5.22	61.78; 85.74	**78.27**; 6.59	51.92; 88.29	79.12; 6.79	61.13; 92.25	**75.39**; 3.77	66.53; 81.99
		3-2	25	75.23; 6.60	62.77; 86.79	73.69; 5.98	59.94; 87.19	78.06; 5.23	63.62; 88.30	**80.77;** 5.12	69.50; 91.88	74.48; 4.26	63.10; 82.08
		p-values (t-test)		0.60		0.21		0.90		0.29		0.36	

Best performing LIFTUPP© trajectory groups are marked with *

Highest mean (aggregated) examination scores presented in bold.

Blue = Trajectory group 1; Red = Trajectory group 2

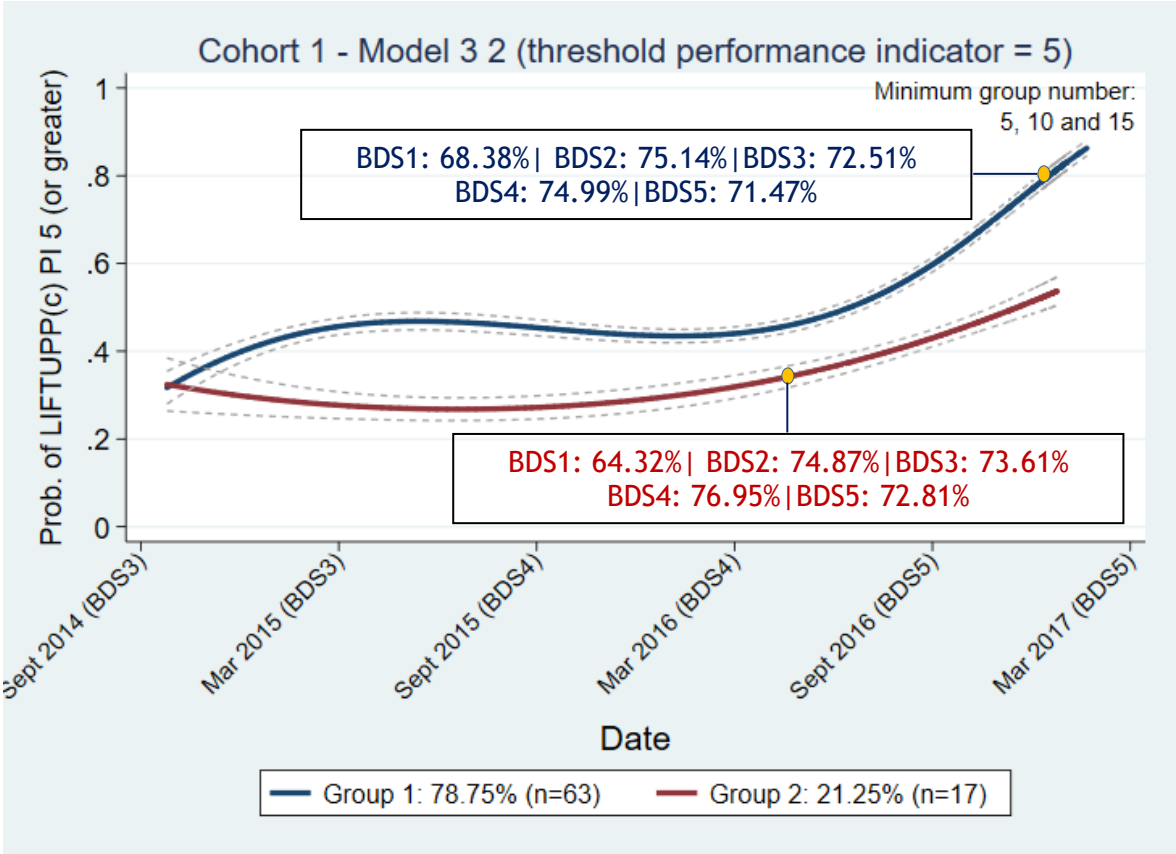


Figure 7.1 – Cohort 1: Mean aggregated BDS examination scores per trajectory group in model 3 2. PI = Performance indicator.

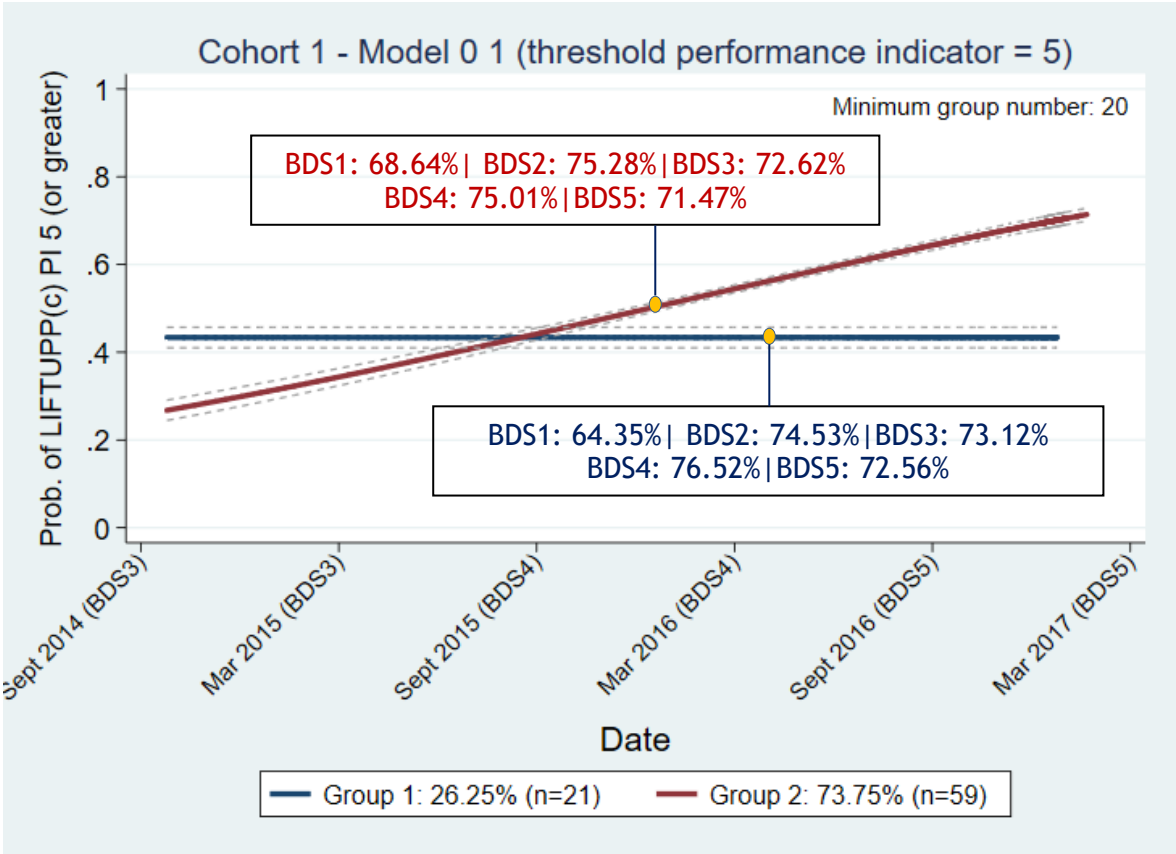


Figure 7.2 – Cohort 1: Mean aggregated BDS examination scores per trajectory group in model 0 1. PI = Performance indicator.

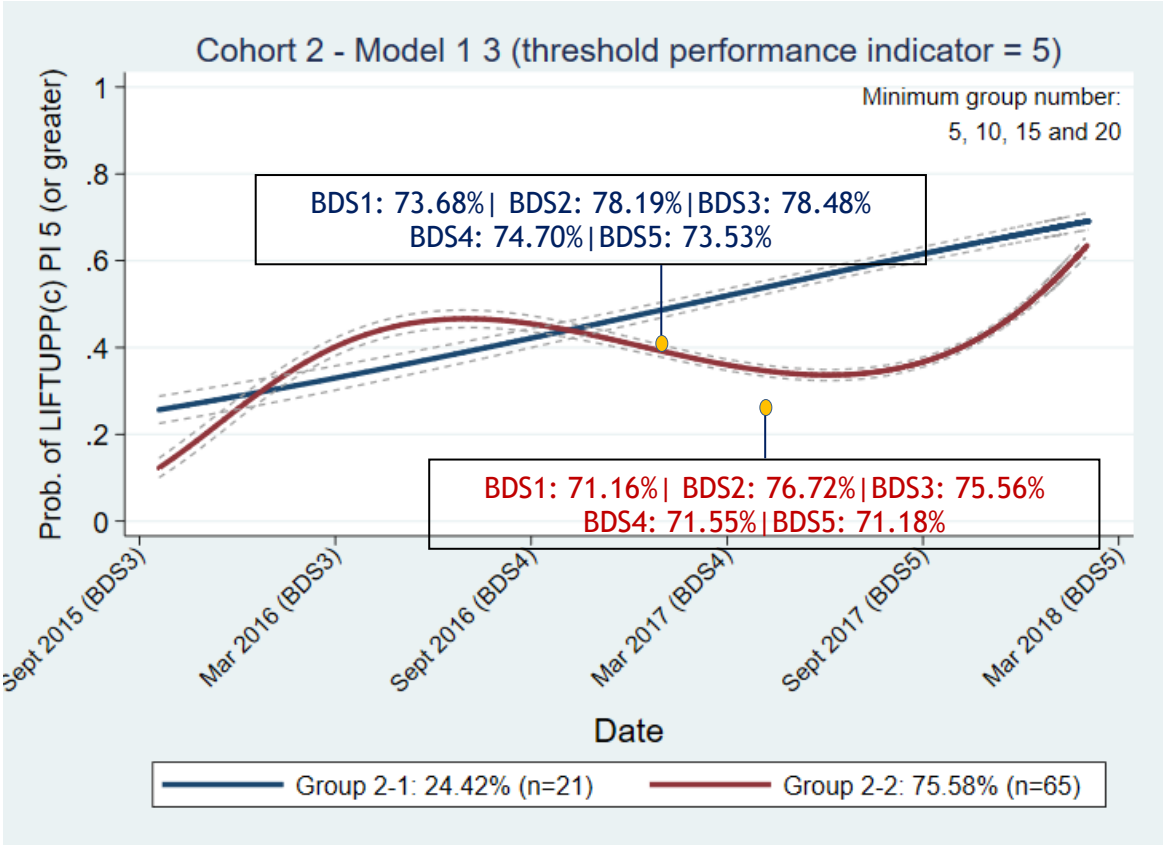


Figure 7.3 – Cohort 2: Mean aggregated BDS examination scores per trajectory group in model 1 3. PI = Performance indicator.

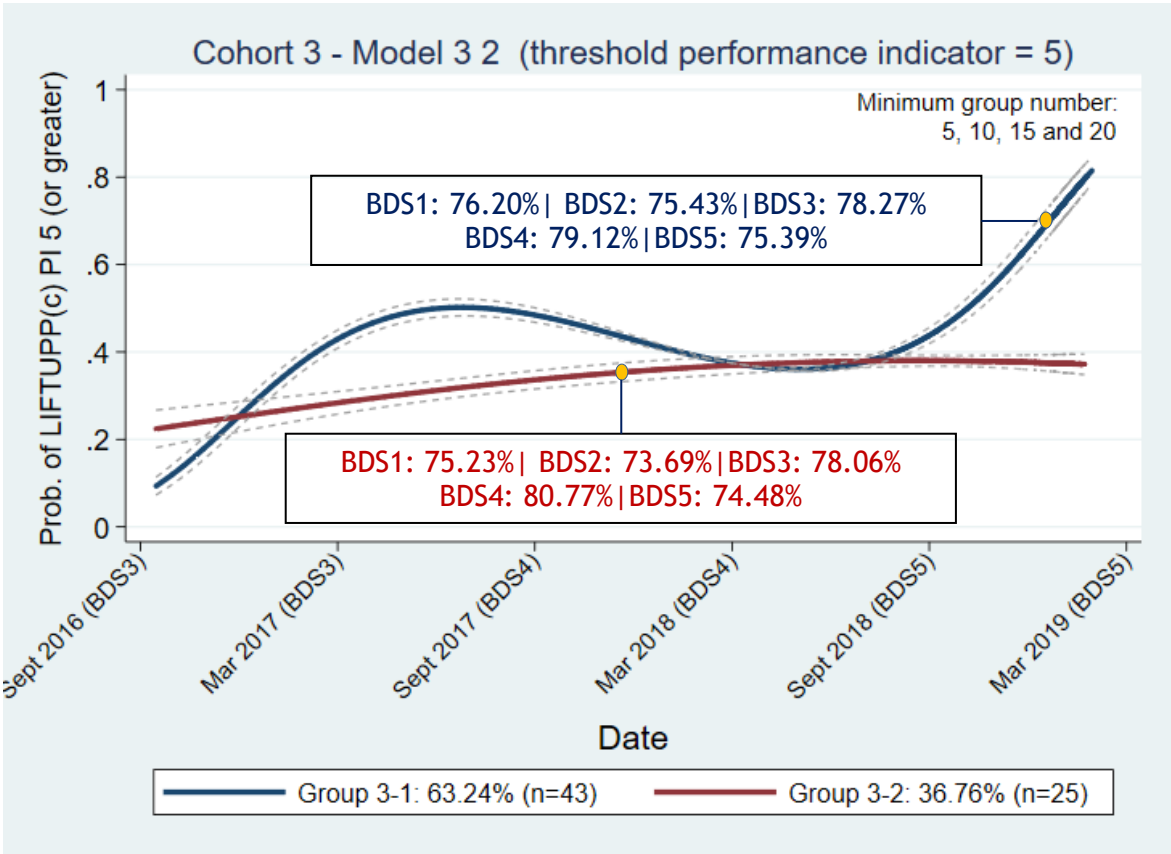


Figure 7.4 – Cohort 3: Mean aggregated BDS examination scores per trajectory group in model 3 2. PI = Performance indicator.

7.4.2 Association between "top fifth" undergraduate BDS examination scores and LIFTUPP© trajectories (Bernoulli/threshold) – concurrent validity

Cross-tabulations between the top fifth (20%) of mean aggregated BDS1-5 examination performance and trajectory group memberships for threshold models are shown in Table 7.3. Figures 7.5–7.8 provide illustrated summaries of the proportion of LIFTUPP© trajectory groups who achieved a top fifth examination performance in each BDS year.

In general, the proportion of students scoring in the top fifth of each BDS year was higher for the better performing trajectory groups. This observation was most consistent for cohorts 2 and 3.

For both cohort 1 models (3 2 and 0 1), a greater proportion of students in the "better" performing trajectory group scored in the top fifth for the BDS1 and BDS2 examinations compared to the second-best performing group. The most pronounced example was seen in model 3 2 where 25.2% of the best performing group scored in the top fifth for BDS1 compared to 0% of the second-best group (Table 7.3 and Figure 7.5). However, in contrast, a greater proportion of the second-best trajectory group were in the top fifth for the BDS3/4/5 examinations compared to the best trajectory group. The most prominent example of this observation was seen in model 0 1 - where 23.8% of the second-best trajectory group scoring in the top fifth for BDS5 compared to 18.6% of the best trajectory group (Table 7.3 and Figure 7.6).

For cohort 2's 1 3 model, a greater proportion of students who were members of the best performing trajectory group (i.e. group 2-1) scored in the top 20% across all five BDS years, although there was little difference between both trajectory groups in BDS2 (3.1%).The greatest difference between the groups was observed in BDS5 whereby 33.3% of the best performing trajectory (group 2-1) produced a top fifth examination performance compared to 15.4% of the second-best trajectory (group 2-2) (Table 7.3 and Figure 7.7).

Similar findings were observed for cohort 3's 3 2 model. A higher proportion of students following the best performing trajectory group (group 3-1) scored in the top 20% in all five BDS years compared to the second-best performing trajectory

(group 3-2). However, little difference was observed between the trajectory groups in BDS2 (0.93%), BDS4 (2.60%) and BDS5 (4.93%) (Tables 7.3 and Figure 7.8).

Table 7.3 - Cross tabulations between LIFTUPP© trajectory group membership and the top fifth (20%) of BDS1-5 examination performance.

Cohort	Model	Trajectory group	n	Academic BDS Year									
				BDS1		*BDS2*		*BDS3*		*BDS4*		*BDS5*	
				Out with top 20% (Q1 to Q4) [% (n)]	In top 20% (Q5) [% (n)]	Out with top 20% (Q1 to Q4) [% (n)]	In top 20% (Q5) [% (n)]	Out with top 20% (Q1 to Q4) [% (n)]	In top 20% (Q5) [% (n)]	Out with top 20% (Q1 to Q4) [% (n)]	In top 20% (Q5) [% (n)]	Out with top 20% (Q1 to Q4) [% (n)]	In top 20% (Q5) [% (n)]
1	3 2	**1-1***	63	75.81 (47)	24.19 (15)	76.19 (48)	23.81 (15)	80.95 (51)	19.05 (12)	82.54 (52)	17.46 (11)	80.95 (51)	19.05 (12)
		1-2	17	100.00 (17)	0.00 (0)	94.12 (16)	5.88 (1)	76.47 (13)	23.53 (4)	82.35 (14)	17.65 (3)	76.47 (13)	23.53 (4)
		p-values (Fisher's exact)		0.03		0.17		0.74		1.00		0.74	
	0 1	**1-1**	21	95.24 (20)	4.76 (1)	95.24 (20)	4.76 (1)	76.19 (19)	23.81 (5)	80.95 (17)	19.05 (4)	76.19 (16)	23.81 (5)
		1-2*	59	75.86** (44)	24.14 (14)	74.58 (44)	25.42 (15)	81.36 (48)	18.64 (11)	83.05 (49)	16.95 (10)	81.36 (48)	18.64 (11)
		p-values (Fisher's exact)		0.10		0.06		0.75		1.00		0.75	
2	1 3	**2-1***	21	71.43 (15)	28.57 (6)	76.19 (16)	23.81 (5)	71.43 (15)	28.57 (6)	71.43 (15)	28.57 (6)	66.67 (14)	33.33 (7)
		2-2	65	81.54 (53)	18.46 (12)	80.00 (52)	20.00 (13)	81.54 (53)	18.46 (12)	81.54 (53)	18.46 (12)	84.62 (55)	15.38 (10)
		p-values (Fisher's exact)		0.6		0.76		0.36		0.36		0.11	
3	3 2	**3-1***	43	76.74 (33)	23.08 (10)	79.07 (34)	20.93 (9)	74.42 (32)	25.58 (11)	81.40 (35)	18.60 (8)	79.07 (34)	20.93 (9)
		3-2	25	87.50 (21)	12.50 (3)	80.00 (20)	20.00 (5)	88.00 (22)	12.00 (3)	84.00 (21)	16.00 (4)	84.00 (21)	16.00 (4)
		p-values (Fisher's exact)		0.34		1.00		0.23		1.00		0.75	

Best performing LIFTUPP© trajectory groups are marked with *

Blue = Trajectory group 1; Red = Trajectory group 2

Group with greatest proportion of membership in top 20% for BDS year examinations highlighted.

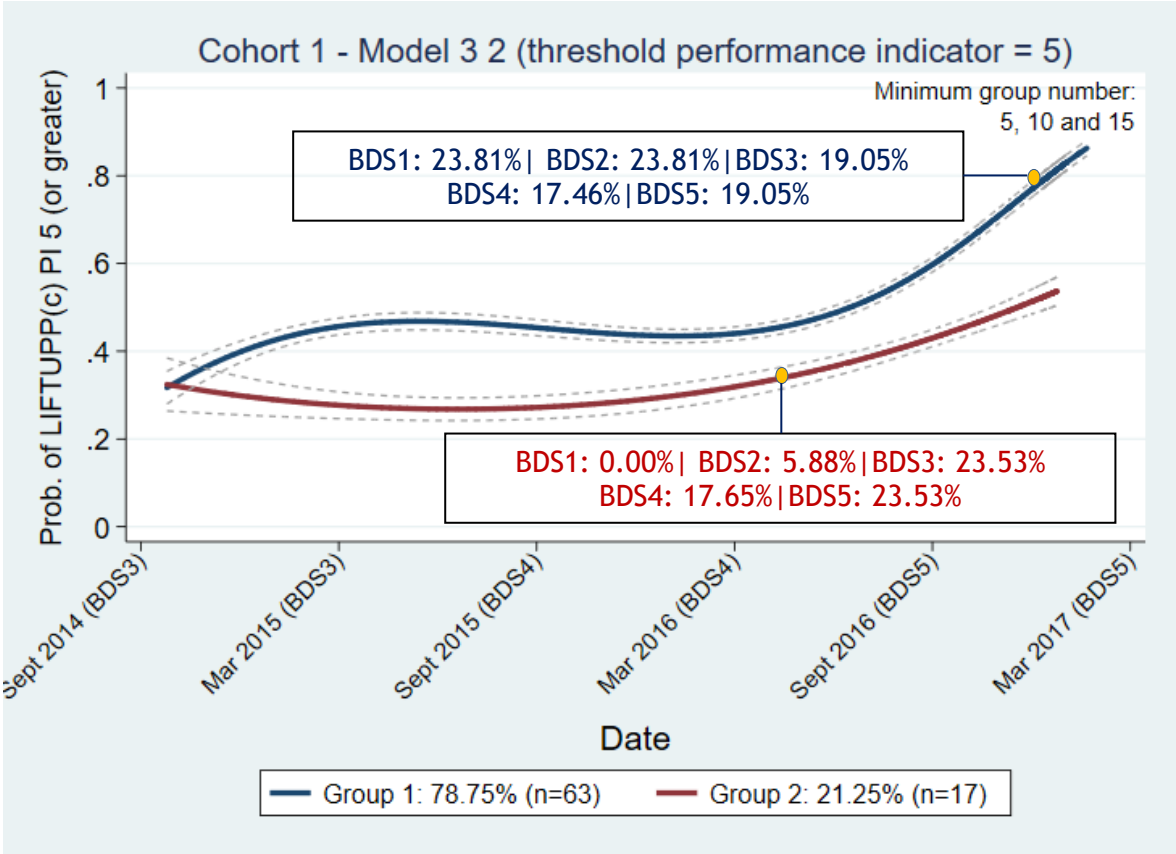


Figure 7.5 – Cohort 1: Proportion of trajectory group who achieved a top fifth performance in the BDS examinations (model 3 2). PI = Performance indicator.

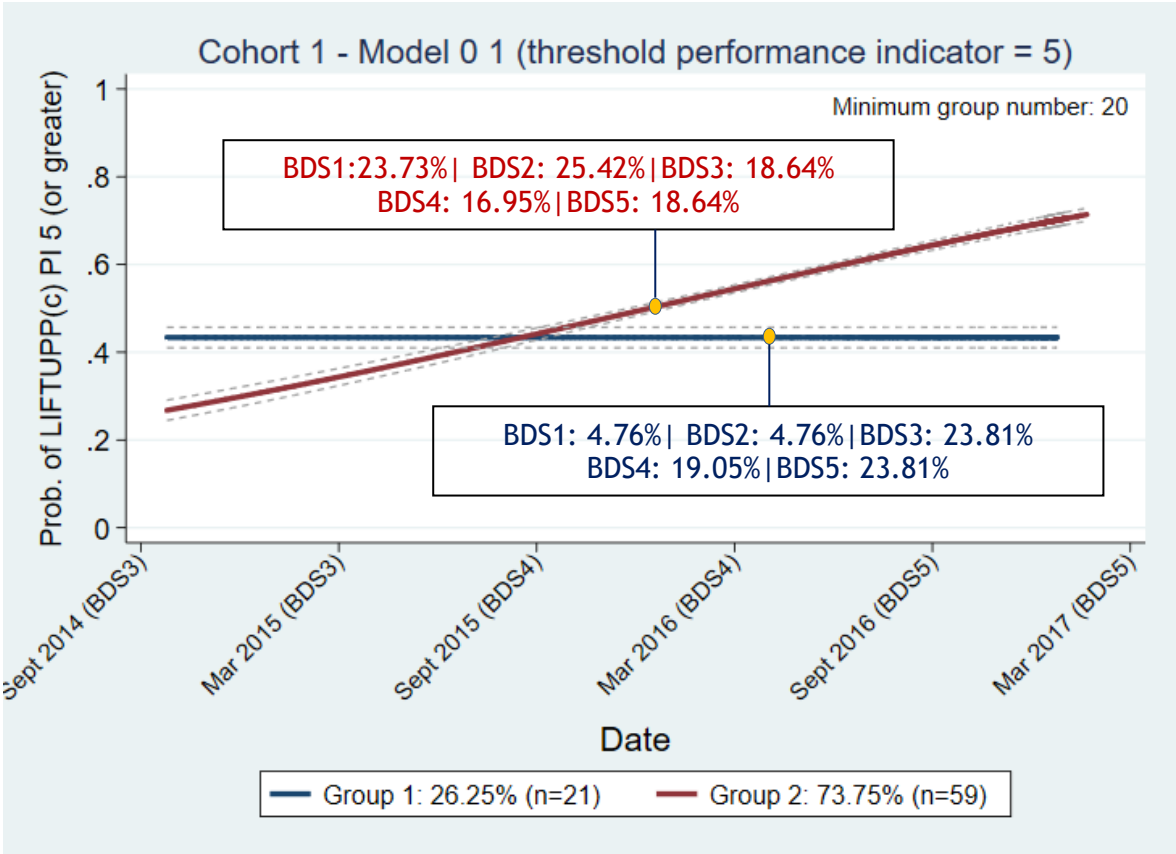


Figure 7.6 – Cohort 1: Cohort 1: Proportion of trajectory group who achieved a top fifth performance in the BDS examinations (model 0 1). PI = Performance indicator.

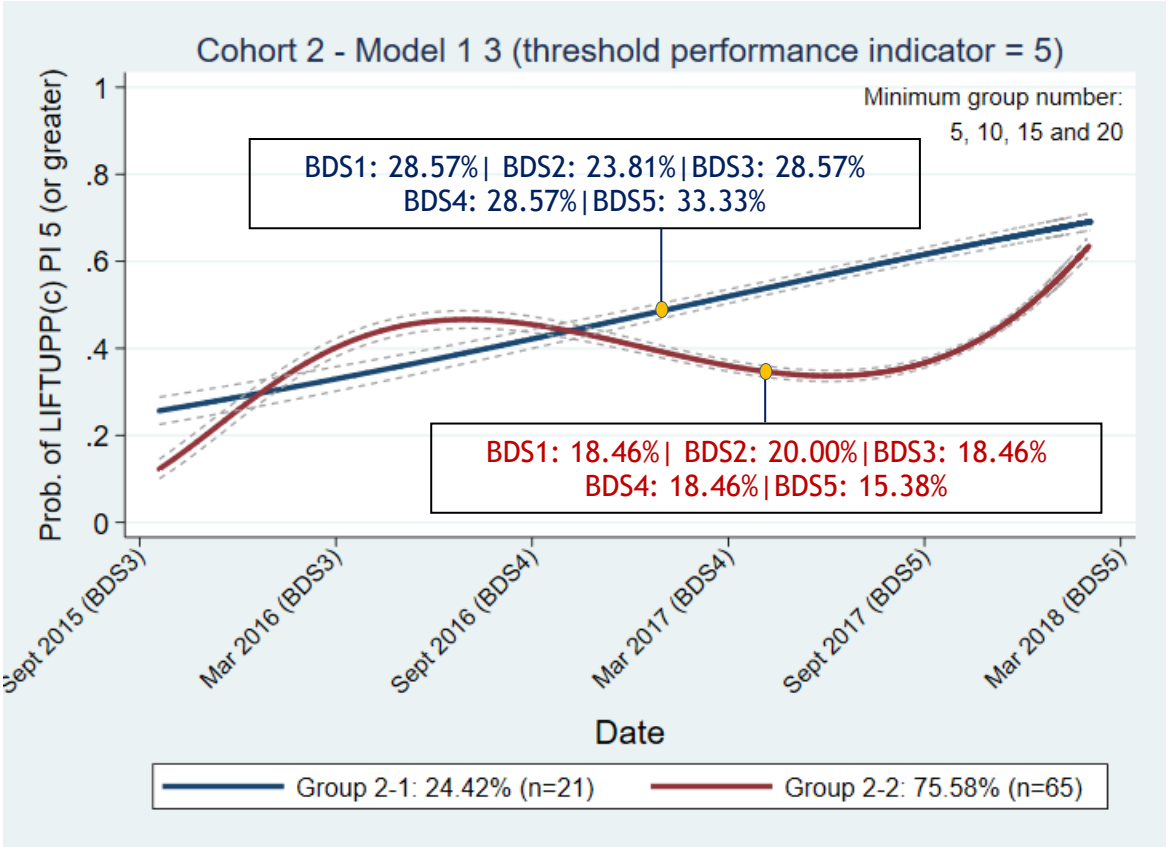


Figure 7.7 – Cohort 2: Proportion of trajectory group who achieved a top fifth performance in the BDS examinations (model 1 3). PI = Performance indicator.

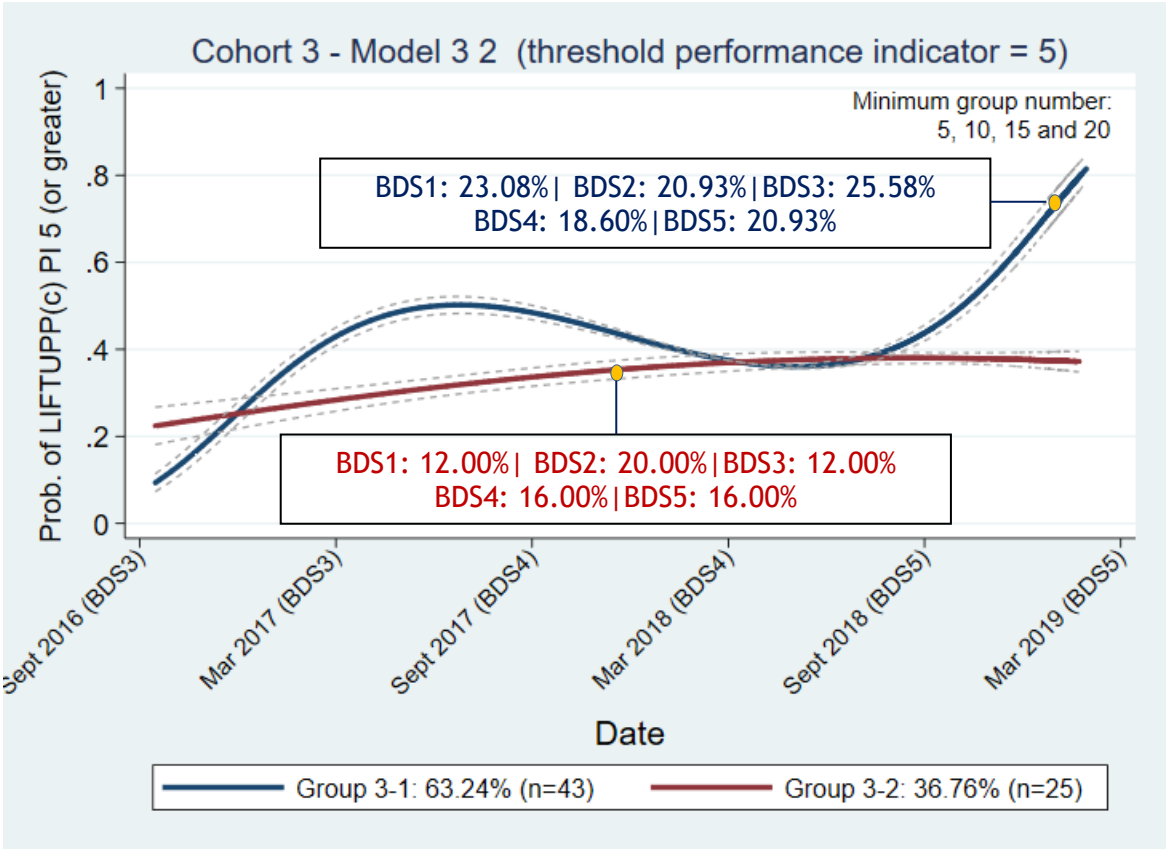


Figure 7.8 – Cohort 3: Proportion of trajectory group who achieved a top fifth performance in the BDS examinations (model 3 2). PI = Performance indicator.

7.4.3 Reliability

Only LIFTUPP© trajectory models with a threshold performance indicator of 5 could be included as part of the Cronbach's alpha calculations since no discrimination between student clinical performance was found using a threshold performance indicator of 4 (see above).

Overall, the panel of assessments displayed high reliability across all three cohorts since all overall Cronbach's alpha scores were $\geq$0.90. However, exclusion of LIFTUPP© data to the panel of assessments led to a marginal increase in reliability in all three cohorts (Cohort 1 - 0.90 to 0.92; Cohort 2 – 0.92 to 0.93; Cohort 3 - 0.90 to 0.92). Removal of any of the other assessments decreased reliability (Table 7.4).

Table 7.4 – Cronbach's alpha coefficients across all BDS1-5 examinations and LIFTUPP© trajectory group membership probability per cohort. Coefficient values are based on the removal of the assessment method from the panel of assessments. MCQ = Multiple-choice question. MSA = Multiple-short answer. OSCE = Objective structured clinical examination.

BDS year	**Assessment**	**Cohort 1**		**Cohort 2**		**Cohort 3**	
		Observations (n)	***Cronbach's alpha***	***Observations (n)***	***Cronbach's alpha***	***Observations (n)***	***Cronbach's alpha***
1	***MCQ***	90	0.8925	92	0.9152	73	0.8845
	OSCE	89	0.8911	92	0.9187	73	0.8823
2	***MCQ***	88	0.8942	92	0.9128	72	0.8881
	MSA	88	0.8895	92	0.9140	72	0.8815
	OSCE	88	0.8995	92	0.9245	72	0.8931
3	***Anatomy***	87	0.8921	92	0.9138	72	0.8829
	MCQ	87	0.8925	92	0.9167	72	0.8843
	MSA	87	0.8912	92	0.9113	72	0.8824
	OSCE	87	0.8997	92	0.9183	72	0.8919
4	***MSA***	84	0.8912	91	0.9147	72	0.8950
5	***OSCE***	82	0.8953	93	0.9236	69	0.8909
3/4/5	***LIFTUPP© group prob.***	80	0.9226	86	0.9335	68	0.9152
		Overall Cronbach's alpha	0.9042	***Overall Cronbach's alpha***	0.9246	***Overall Cronbach's alpha***	0.8980

7.4.4 Association between LIFTUPP© GBTM membership and LEP GBTM membership – Predictive validity

The following section presents the results of cross tabulations between trajectory group memberships for the LIFTUPP© and LEP GBTMs selected to represent clinical assessment data.

Due to the number of models previously selected, 16 cross tabulations were conducted to cater for any minimum group membership restrictions or threshold scores applied to the selected LIFTUPP© and LEP models. However, in this chapter, only models which used a minimum group number restriction of 20 students/VDPs will be presented. This is because, after comparing the selected LIFTUPP© models with undergraduate examination outcomes (see sections 7.4.1 and 7.4.2 above) and the selected LEP models, there was still no clear evidence to suggest models using particular minimum group restriction numbers were better than others – from a statistical perspective. However, GBTMs with a minimum group number of 20 appeared more stable and parsimonious since the trajectories produced were easy to understand and gave clearer explanations regarding educational significance when LIFTUPP© and LEP models were compared with one another and when LIFTUPP© models were compared with undergraduate examinations (sections 7.4.1 and 7.4.2).

Comparisons between LIFTUPP© and LEP models for minimum group numbers of 5, 10 and 15 are provided in appendix 11.

To recap, the threshold performance indicator for LIFTUPP© GBTMs was consistently set at 5, whereas threshold scores of 6 and 7 were investigated for LEP GBTMs (see sections 7.3.1 and 7.3.3 above).

7.4.4.1 LIFTUPP© threshold ≥5 models vs. LEP threshold ≥6 models (minimum group number = 20)

Cohort 1

Of the 80 study participants who graduated BDS from the University of Glasgow in 2017, 67 (83.8%) registered with a Scottish VDT scheme. Therefore, a total of

67 participant LIFTUPP© and LEP GBTM trajectory group memberships were compared for cohort 1.

Table 7.5 displays cross tabulations between LIFTUPP© model 0 1 and LEP model 1 1. A greater proportion of the best LIFTUPP© group (i.e., LIFTUPP© group 1-2) became members of the best performing LEP group (i.e., LEP group 1-2) compared to the second-best performing LIFTUPP© group (i.e., LIFTUPP© group 1-1) (66.0% vs 60.0%).

Table 7.5 – Cross tabulations between the trajectory group memberships for LIFTUPP© group-based trajectory model (GBTM) 0 1 and longitudinal evaluation of performance (LEP) GBTM 1 1. NOTE: The best performing groups for both LIFTUPP© and LEP GBTMs are marked with an *.

Cohort 1				
Minimum number participants per group: 20		**LEP model 1 1 group (threshold score = 6)**		**Fisher's exact (p)**
		1-1	***1-2****	
LIFTUPP© model 0 1 group (threshold performance indicator = 5)	***1-1*** (n = 20)	n = 8 40.00%	n = 12 60.00%	0.78
	1-2* (n = 47)	n = 16 34.04%	n = 31 65.96%	

Cohort 2

In cohort 2, 70 out of 86 (i.e., 81.4% of) study participants from the Glasgow's graduating BDS class of 2018 enrolled in a Scottish VDT scheme. Therefore, 70 participant's LIFTUPP© and LEP GBTM trajectory group memberships were available for comparison.

Comparisons between LIFTUPP© model 1 3 and LEP model 3 2 are presented in Table 7.6. A greater proportion of members of the best performing LIFTUPP© group (i.e., LIFTUPP© group 2-1) became members of the best performing LEP group (i.e., LEP group 2-2) compared to the second-best performing LIFTUPP© group (i.e., LIFTUPP© group 2-2) (72.2% vs 53.9%).

Table 7.6 – Cross tabulations between the trajectory group memberships for LIFTUPP© group-based trajectory model (GBTM) 1 3 and longitudinal evaluation of performance (LEP) GBTM 3 2. NOTES: The best performing groups for both LIFTUPP© and LEP GBTMs are marked with an *. Both LIFTUPP© model 1 3 and LEP model 3 2 were also the models selected if a minimum group number restriction of 15 was used.

Cohort 2				
Minimum group number: 15 and 20		LEP model 3 2 (threshold score = 6)		Fisher's exact (p)
		2-1	***2-2****	
LIFTUPP© model 1 3 group (threshold performance indicator = 5)	***2-1**** (n = 18)	n = 5 27.78%	n = 13 72.22%	0.27
	2-2 (n = 52)	n = 24 46.15%	n = 28 53.85%	

Cohort 3

Of the 68 study participants who graduated BDS from the University of Glasgow in 2019, 60 (88.2%) registered with a Scottish VDT scheme. Therefore, a total of 60 participant LIFTUPP© and LEP GBTM trajectory group memberships were compared for cohort 3.

Table 7.7 – Cross tabulations between the trajectory group memberships for LIFTUPP© GBTM 3 2 and Longitudinal Evaluation of Performance (LEP) GBTM 1 3. NOTES: The best performing groups for both LIFTUPP© and LEP GBTMs are marked with an *. Both LIFTUPP© model 3 2 and LEP model 1 3 were also the models selected if minimum group number restrictions of 10 and 15 was used.

Cohort 3				
Minimum group numbers: 10, 15 and 20		LEP model 1 3 (threshold score = 6)		Fisher's exact (p)
		3-1	***3-2****	
LIFTUPP© model 3 2 group (threshold performance indicator = 5)	***3-1**** (n = 37)	n = 17 45.95%	n = 20 54.05%	0.30
	3-2 (n = 23)	n = 14 60.87%	n = 9 39.13 %	

Table 7.7 shows cross tabulations between trajectory group memberships for the LIFTUPP© model 3 2 and LEP model 1 3. Compared to the second-best performing LIFTUPP© group (i.e., LIFTUPP© group 3-2), 14.9% more members of

the best performing LIFTUPP© group (i.e., LIFTUPP© group 3-1) became members of the best performing LEP group (i.e., LEP group 3-2).

7.4.4.2 LIFTUPP© threshold ≥5 models vs. LEP threshold ≥7 models

Cohort 1

Table 7.8 displays cross tabulations between LIFTUPP© model 0 1 and LEP model 1 1. The comparison between these models showed that a near equal proportion of participants who were in either the second-best performing LIFTUPP© group (i.e., LIFTUPP© group 1-1) or the best group (LIFTUPP© group 1-2) were found in the best performing LEP group (i.e., LEP group 1-1) (55.0% vs 55.3%).

Table 7.8 – Cross tabulations between the trajectory group memberships for LIFTUPP© GBTM 0 1 and Longitudinal Evaluation of Performance (LEP) GBTM 1 1. NOTE: The best performing groups for both LIFTUPP© and LEP GBTMs are marked with an *.

Cohort 1				
Minimum group number: 20		**LEP model 1 1 group (threshold score = 7)**		**Fisher's exact (p)**
		1-1	***1-2****	
LIFTUPP© model 0 1 group **(threshold performance indicator = 5)**	***1-1*** (n = 20)	n = 9 45.00%	n = 11 55.00%	1.00
	1-2* (n = 47)	n = 21 44.68%	n = 26 55.32%	

Cohort 2

Table 7.9 shows cross tabulations between trajectory group memberships for the LIFTUPP© model 1 3 and LEP model 1 3. Compared to the second-best performing LIFTUPP© group (i.e., LIFTUPP© group 2-2), 20.5% more members of the best LIFTUPP© group (i.e., LIFTUPP© group 2-1) belonged in the best performing LEP group (i.e., LEP group 2-2).

Table 7.9 – Cross tabulations between the trajectory group memberships for LIFTUPP© GBTM 1 3 and Longitudinal Evaluation of Performance (LEP) GBTM 1 3. NOTES: The best performing groups for both LIFTUPP© and LEP GBTMs are marked with an *. Both LIFTUPP© model 1 3 and LEP model 1 3 were also the models selected if a minimum group number restriction of 15 was used.

Cohort 2				
Minimum group numbers: 15 and 20		**LEP model 1 3 (threshold score = 7)**		**Fisher's exact (p)**
		2-1	***2-2****	
LIFTUPP© model 1 3 group	***2-1****	n = 6	n = 12	0.11
	(n = 18)	33.33%	66.67%	
(threshold performance indicator = 5)	***2-2***	n = 28	n = 24	
	(n = 52)	53.85%	46.15%	

Cohort 3

Finally, Table 7.10 shows cross tabulations between trajectory group memberships for the LIFTUPP© model 3 2 and LEP model 1 2. A greater proportion of the best performing LIFTUPP© group (i.e., LIFTUPP© group 3-1) belonged to the best performing LEP group (i.e., LEP group 3-2) compared to the second-best performing LIFTUPP© group (i.e., LIFTUPP© group 3-2) (51.4% vs 39.1%).

Table 7.10 – Cross tabulations between the trajectory group memberships for LIFTUPP© GBTM 3 2 and Longitudinal Evaluation of Performance (LEP) GBTM 1 2. NOTE: The best performing groups for both LIFTUPP© and LEP GBTMs are marked with an *.

Cohort 3				
Minimum group number: 20		**LEP model 1 2 (threshold score = 7)**		**Fisher's exact (p)**
		1	***2****	
LIFTUPP© model 3 2 group	***3-1****	n = 18	n = 19	0.43
	(n = 37)	48.65%	51.35%	
(threshold performance indicator = 5)	***3-2***	n = 14	n = 9	
	(n = 23)	60.87%	39.13%	

7.4.5 A note on Longitudinal Evaluations of Performance (LEP) models: Threshold scores 6 and 7.

Based on the comparisons presented in section 7.4.4, the use of either LEP scores of 6 and 7 did not appear to influence the relationship between LIFTUPP© and LEP models. However, it can be argued a score of 7 is a high threshold to set for clinical competence since it is categorised as a "superior" level of performance in the LEP assessment criteria (chapter 3, section 3.5.3.2). As a result, only Bernoulli LEP models using a threshold score of 6 were taken forward for presentation to key stakeholders in the qualitative component of the study (chapter 8).

7.5 Summary

This chapter has reported on investigations undertaken to compare undergraduate longitudinal clinical assessment data against well-established assessment methods within dental education (BDS undergraduate examinations and LEPs). It has further developed a validity argument for the use of longitudinal clinical assessment data in assessing the development of clinical competence in undergraduate dental students by investigating two subtypes of criterion validity (concurrent and predictive). Concurrent validity was investigated through comparisons between undergraduate longitudinal clinical assessment data and undergraduate examination outcomes (BDS1-5). Predictive validity was investigated by comparing undergraduate longitudinal clinical assessment data with postgraduate longitudinal clinical assessment data.

When considering mean undergraduate examination scores, there was little distinction between the LIFTUPP© trajectory groups for all cohorts, perhaps reflective of the narrow spread (small variance) in examination scores within each cohort. However, when considering students scoring in the top fifth of their year, there were clear associations with the better performing trajectories - particularly for cohorts 2 and 3. The results from the two more recent cohorts indicate LIFTUPP© data have a degree of concurrent validity.

The lack of a stable association between LIFTUPP© trajectories and examination performance in cohort 1 may have been caused by poorer calibration among

assessors following the initial adoption of LIFTUPP© into the BDS curriculum. This topic is discussed further in chapter 9.

Comparing participant LIFTUPP© trajectory group memberships (established in chapter 5) against LEP trajectory group memberships (established in chapter 6) suggested that participants who performed well clinically at undergraduate level were more likely to perform well in VDP. This observation was seen in almost all cross tabulations between LIFTUPP© and LEP trajectory group memberships, the only expectation was found in the comparison between LIFTUPP© GBTM 3 2 and LEP GBTM 1 1 3. The association between "better" undergraduate longitudinal clinical performance and "better" postgraduate longitudinal clinical performance suggests LIFTUPP© data have a degree of predictive validity. These comparisons also revealed there was little change in the relationship between each of the datasets if either minimum group number restrictions of 5, 10, 15 and 20 were used for both LIFTUPP© and LEP GBTMs. However, GBTMs with a minimum group number of 20 appeared more stable and parsimonious. Furthermore, the use of either LEP scores 6 and 7 did not impact upon interpretation of the relationship between LIFTUPP© and LEP data.

In terms of reliability, the inclusion of LIFTUPP© data in the panel of undergraduate assessments led to a marginal decrease in the overall reliability in all three cohorts. However, the probability of LIFTUPP© trajectory group membership may have not been the best metric to use in the calculation of reliability.

Further discussion on the findings presented in this chapter and their implications is provided in chapter 9.

Chapter 8 - Consultations with key stakeholders in dental education

8.1 Introduction

The following chapter reports the findings obtained from thematic analyses of focus group discussions between key stakeholders within dental education.

As previously highlighted in chapter 3 (section 3.5.2), qualitative input was sought to provide complementary data for triangulation with the quantitative analyses conducted in the study. By ensuring the findings generated from the quantitative component of the study reflect the assessment experiences of key stakeholders in dental education, additional support for content validity can be attributed to longitudinal clinical assessment. Qualitative input from key stakeholders was also required to identify areas for enhancement of assessment in dental education and generate future lines of enquiry for subsequent research.

Collectively, the findings generated for both purposes (together with the quantitative results) may assist dental schools in reviewing their assessment methods to ensure they are adopting best practice according to the available evidence. Ultimately this protects patients through ensuring dental students are appropriately trained and assessed before they are registered as a qualified practitioner.

8.2 Aim

To canvas the opinions of key stakeholders on the analyses of undergraduate longitudinal clinical assessment data and their comparison with more established forms of assessment – namely undergraduate examination outcomes and postgraduate longitudinal clinical assessment.

This aim is aligned with two research questions (chapter 2, section 2.2):

Research question 2a: What are the main patterns of longitudinal clinical assessment over time within a BDS year, across BDS years and across cohorts? (Content validity)

Research question 3: According to key stakeholders in dental education, how might the findings of research questions 2a, 2b and 2c be used to enhance assessment in dentistry?

8.3 Method

The overall study design, setting, recruitment, data generation and analysis conducted in this chapter were the same as those described in chapter 3 (section 3.5.4).

The quantitative results presented to the recruited key stakeholders during the focus group discussions included:

- Figures 5.15-5.17 (chapter 5, section 5.4.2.2) – LIFTUPP© trajectories (threshold performance indicator = 4).

- Figures 5.19-5.21 (chapter 5, section 5.4.2.3) - LIFTUPP© trajectories (threshold performance indicator = 5).

- Simplified versions of Figures 7.6-7.8 (chapter 7, section 7.4.2) - LIFTUPP© trajectories (threshold performance indicator = 5) vs. top 20% performance in final (BDS4/5) examinations.

- Annotated side-by-side comparisons of Figures 5.19 and 7.6, Figures 5.20 and 7.7, and Figures 5.21 and 7.8 – LIFTUPP© trajectories (threshold performance indicator = 5) vs. LEP trajectories (threshold score = 6).

These results were selected for presentation at the focus groups since, following comparisons between LIFTUPP© and undergraduate examination and LEP data, the GBTMs for LIFTUPP© and LEP which used minimum group number restrictions of 20 students/VDPs appeared more stable and parsimonious (see chapter 7, section 7.4.4). It was also important to show how the use of threshold

performance indicators 4 and 5 impacted the number of trajectories produced for LIFTUPP© data.

8.4 Results

8.4.1 Recruitment

As previously described in chapter 3 (section 3.5.4), two separate focus groups were conducted. Seven participants were ultimately recruited for the first focus group, which was to consist of undergraduate dental students, VDPs and recent (2017-19) dental graduates. In total, three dental students (one BDS3 and two BDS5), two VDPs, and two recently (2017-19) qualified dentists were recruited. From this point onwards, these individuals will be referred to as "student focus group participants" as the VDPs and recent graduates were previously members of the dental student cohorts whose assessment data were used for this study.

Five participants were finally recruited for the second ("faculty") focus group, conferring experienced staff representation from senior dental school academia (both internal and external to the University of Glasgow), BDS course co-ordination, dental school clinical supervision, and VDP assessment from NES. All participants had some responsibility for elements of delivery of teaching and assessment in their respective dental curricula.

8.4.2 Themes identified from focus group discussions

As previously described in chapter 3 (section 3.5.4.5), the content of the transcripts from both focus groups were initially coded to reduce and refine the volume of data so that categories and key themes could be readily identified. Codes generated and assigned to comments made by the focus group participants during the thematic analysis process could be organised into 13 categories. From the list of categories, five overarching themes were identified across both focus groups:

1. Attitudes towards longitudinal clinical assessment data trajectories.

2. Scepticism on LIFTUPP© data quality and consistency.

3. Attitudes towards longitudinal clinical assessment.

4. Enhancement of assessment practice.

5. Future research.

A brief example of the analysis process for the first theme (attitudes towards longitudinal clinical assessment data trajectories) is provided in Figure 8.1.

Tables 8.1-8.5 summarise how focus group participant comments were categorised and arranged under the overarching themes. More detailed summaries and examples of comments aligned with each theme are provided over the following sections (8.4.2.1-8.4.2.5).

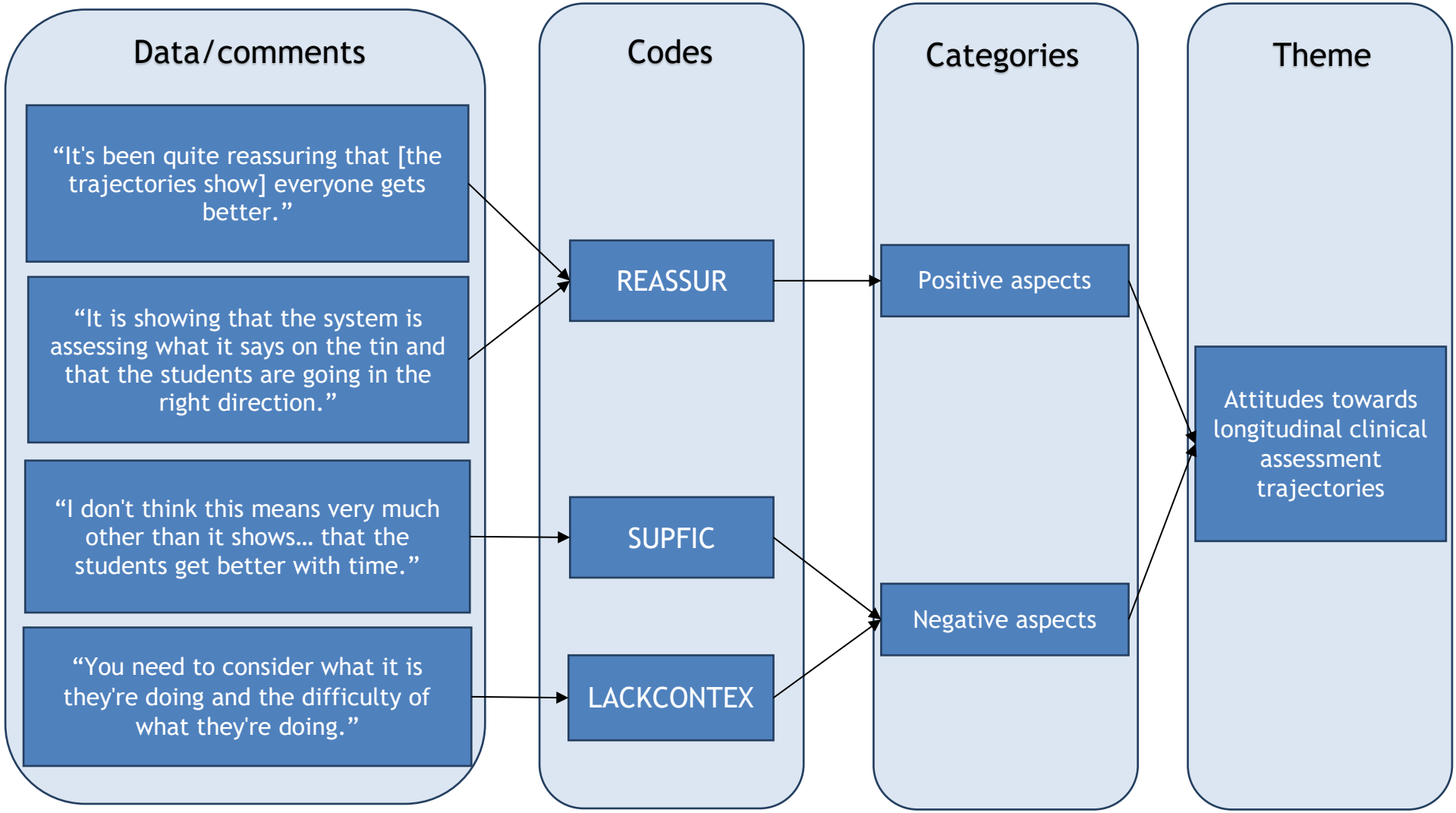


Figure 8.1 - Example of thematic analysis process - from coding comments made by participants to identification of categories and an overarching theme. Codes: REASSUR = "Reassurance"; SUPFIC = "Superficial"; LACKCONTEX = "Lack of context".

Table 8.1 - Thematic analysis of comments on attitudes towards longitudinal clinical assessment data trajectories.

THEME 1: Attitudes towards longitudinal clinical assessment data trajectories	
Category 1a: Making sense of patterns	***Category 1b: Positive aspects***
- Reflect experiences of clinical training and assessment - Reflect experiences of student progression through the undergraduate BDS curriculum	- Show progressive development of student clinical performance - Provide reassurance that student clinical performance improves over the duration of Glasgow Dental School's BDS curriculum - Provide reassurance that longitudinal clinical assessment is capturing improvement in student clinical performance - Act a starting point for research into longitudinal clinical assessment within dental education
Category 1c: Negative aspects	***Category 1d: Relationship with other forms of assessment***
- Superficial - Lack of context of the assessment/clinical encounters - Cannot identify poor performing students - Cause of confusion	- Strong/Strongest relationship - Weak relationship - No Relationship

Table 8.2 - Thematic analysis of comments signifying scepticism on LIFTUPP© data quality and consistency.

THEME 2: Scepticism on LIFTUPP© data quality and consistency	
Category 2a: Assessor factors	***Category 2b: Student factors***
- Standardisation - Adapting to new assessment method - Time constraints and convenience - Failure to fail - Assessor bias	- Student approaches to learning - Comfort within clinical environment(s) - Self-confidence
Category 2c: Other factors	
- LIFTUPP© system IT problems	

Table 8.3 - Thematic analysis of comments signifying attitudes towards longitudinal clinical assessment.

THEME 3: Attitudes towards longitudinal clinical assessment	
Category 3a: Value of longitudinal clinical assessment	***Category 3b: Role of longitudinal clinical assessment***
- Longitudinal assessment is valuable	- Replace standalone competence assessments - Part of panel of assessment methods

Table 8.4 - Thematic analysis of comments on how the study results could be used to enhance assessment practice.

THEME 4: Enhancement of assessment practice	
Category 4a: Student assurance	***Category 4b: Data collection***
- Standardisation and moderation - Multiple/Variety of assessors	- Streamlining and ensuring key information collected - Ensuring procedure difficulty and complexity are recorded. - Previous procedural experience

Table 8.5 - Thematic analysis of comments on how the study results could be used to inform future dental education studies.

THEME 5*: Future research*	
Category 5a: Larger studies	***Category 5b: Expanded studies***
- Small participant numbers in current study	- Beyond vocational dental training (VDT) year - Need to provide context of the assessment/clinical encounters within longitudinal clinical assessment (e.g., procedure difficulty) - Student insight and confidence on clinical performance - Compare longitudinal clinical assessment data with case presentation assessments

8.4.2.1 Longitudinal clinical assessment data trajectories

i) Making sense of patterns

Participants from both focus groups commented they were able to make sense of the GBTM patterns based on their experiences of clinical training, assessment, and how Glasgow Dental School's undergraduate BDS curriculum is structured. For example, some comments referring to "dips" in trajectories patterns at the beginning of BDS4 were:

> "That might be your first crown prep or your first endo so you then might get a lower score because you've not had as much practice… so I think that does make sense that it would dip at that point."

> "…in fourth year, they do the simulated sessions to make them safe to progress with the more difficult procedures, and then they go into the clinic without their stabilisers on. They do initially need help and advice, and getting that advice and input brings their performance indicator down a bracket or two"

ii) Positive aspects

All participants felt the undergraduate longitudinal clinical assessment data trajectories signified students' clinical skills were improving over time. Some student focus group participants signified that seeing progressive clinical development presented in this format was valuable. Others commented they had expected to see the upward trends displayed.

> "I think the fact that you've got an overall curve is much better."

> "It's no surprise that is the trend, but it's good *to* see it."

Faculty focus group participants also felt the trajectory patterns were somewhat valuable and interesting. More importantly, it "reassured" them students' clinical skills were progressing as hoped within Glasgow's BDS curriculum and their development was being captured by the LIFTUPP© system.

> "It shows that you run a valid training programme in that the students get better with time."

> "It is showing that the system is assessing what it says on the tin and that the students are going in the right direction."

Regardless of their drawbacks (see *iii* below), faculty agreed the trajectories and their comparisons with more established forms of assessment (see *iv* below) provided a good "building block" for further research on longitudinal clinical assessment within dental education.

iii) Negative aspects

Despite the reassurances, faculty felt the trajectories did not provide enough detail for identifying students who were struggling – which they suggested should be one of the main focuses of longitudinal clinical assessment.

> "My experience of LIFTUPP© is that it's looking for patterns of failure or where development isn't as strong as it should be".

Some faculty commented they had hoped to see distinct patterns of poor performance in the presented results. This indicated some participants may have misunderstood the purpose of the study as Glasgow Dental School already has a process for identifying weak students individually. This process was acknowledged by another faculty focus group participant, who subsequently suggested data presentation formats available within the LIFTUPP© system were "really good for identifying people who are not performing quite well". The potential misunderstanding of the study purpose and Glasgow Dental School's process for identifying poorly performing students in LIFTUPP© is discussed further in chapter 9 (section 9.3.7).

Faculty also felt that more context needed to be given to longitudinal clinical assessment data patterns – e.g., procedure difficulty and complexity and one participant from the student focus group also found the trajectories difficult to interpret at several points in the discussion.

iv) Relationship with other forms of assessment

The consensus between all focus group participants was that a positive trend/relationship could be seen between undergraduate longitudinal clinical assessment data trajectories and more established forms of assessment (i.e., undergraduate examination outcomes and postgraduate LEPs). However, there were differing opinions on how strong these trends/relationships were. All faculty believed the relationships were "weak"/ had a "loose correlation"/had "little correlation".

Students had mixed opinions. Some felt there was a "strong relationship" between undergraduate longitudinal clinical assessment data and examination

outcomes – although one student was "not convinced that there [was] a relationship" between these two assessment methods.

Others felt there was a strong relationship between undergraduate longitudinal clinical assessment data and LEPs and, out of all the assessment method comparisons, this relationship was the strongest. Some students offered explanations as to why they believed this relationship was the strongest.

> "…that's largely down to the fact that LIFTUPP© is very similar to LEP."

> "they're probably the ones that are more easily comparable in terms of type of assessment".

Participants from both focus groups also commented they had anticipated seeing a relationship between LIFTUPP© and LEPs but it was "not as strong" as expected. This topic is discussed further in chapter 9 (section 9.5.3).

8.4.2.2 Scepticism on LIFTUPP© data quality and consistency

A sense of mistrust of LIFTUPP© data was a key theme identified from the focus group discussions – especially among student focus group participants. A variety of factors which may have impacted and influenced the quality and consistency of the longitudinal clinical assessment data collected via LIFTUPP© were highlighted.

i) Assessor factors

Standardisation of assessors for longitudinal clinical assessment was one of most predominant discussion points in both focus groups. Student focus group participants recalled experiences of inconsistent awarding of LIFTUPP© performance indicators and variation among assessors.

> "There is so much variation between different assessors so it can be quite difficult to see where you're at… just because it is subjective."

> "I remember, just from personal experience, I had done a certain piece of work, I showed it to two different staff members, and I got completely different results from both."

Faculty were also aware of variation between assessors, which can lead to some being labelled as “hawks” and “doves” – the latter of which tend to be “targeted” by students. Concerns of overly strict assessors (i.e., “hawks”) - who did not apply the assessment criteria correctly and may be over penalising students - were expressed within both focus groups. One faculty participant surmised:

> “Sometimes it's the students that are more inquiring and more inquisitive and wishing to develop further that get the lower score because they ask more questions, and it doesn't mean that they couldn't do it. And you really need a well-trained and well-calibrated assessor to be able to know: “actually, that’s still a 5” 'cause they've done it! They’re now just wanting to do it even better.”

Faculty acknowledged that, despite regular efforts to calibrate assessors and familiarise them with the assessment criteria, there were ongoing challenges relating to standardisation.

> “I've done a calibration session for staff myself and even with playing a video of an event and having the descriptors on the slide, it's amazing how broad variation staff have on their award of grades, even though the criteria are very explicit and even printed in front of them. So, it's… it's a very difficult issue to address.”

Assessors adapting to a new clinical assessment method was also discussed as potential cause for variation in assessment. Student focus group participants suspected it may have taken time for assessors to get used to the assessment criteria and standard against which students were to be assessed.

> “Most of them were used to… giving scores that were quite different, so when… they changed the criteria, we had a lot of… teething problems.”

Student focus group participants also highlighted incidences where they suspected assessors may not award appropriate performance indicators.

> “[It] depends on how busy the clinician is…sometimes they’re giving scores for six students and… they’re just scoring [to just get] them all done to get out on time or whatever.”

> "When you're talking about scores of 3 and 4 - when 3 requires a written comment, but a 4 doesn't…so… [for] convenience, you might

> end up with all 4s when in actual fact a 3 with a comment would have been a lot more beneficial."

Finally, several student focus group participants commented that LEPs may be particularly prone to assessor bias.

> "You do have the possibility of a trainer being more favourable to somebody in [VDT]."

> "Your [VDT] trainer is more likely to be nice to you and I think they all (laughs)… it's true! Because their own practice is on the line."

> "…that's going to be one single assessor the whole time, who… is quite invested in their [VDP] doing well."

ii) Student factors

Variation in student approaches to learning was recognised as a potential reason for differences in assessment performance between cohorts.

> "It might just be that cohort 3 were a really good year."

> "Some year groups may have people that naturally just pick up things a lot quicker than others. So, you always have a variation every year on how people progress, I think."

> "Some people work harder, study more, are more interested, more motivated so therefore they will naturally get better. But the other people that aren't as interested… may just… may not get better at a faster rate."

One participant suggested student comfort and familiarity within clinical environments may also impact which LIFTUPP© performance indicators they are awarded since they may need "to ask more questions" and, therefore, they may be awarded lower performance indicators.

Participants from both focus group also suspected the degree of "confidence" students display in their clinical performance could also affect which performance indicators are awarded.

i) *Other factors*

Technical difficulties during the early stages of integration of the LIFTUPP© system at Glasgow Dental School were also cited as a problem which may have influenced some of the study results since some students may not have had some of their assessments recorded.

"Some clinicians didn't have passwords."

"[The iPads] weren't always being put on to charge."

8.4.2.3 Attitudes towards longitudinal clinical assessment

i) *Value of longitudinal clinical assessment*

All focus group participants believed longitudinal clinical assessment was a valuable form of assessment. Some participants even suggested it had become "more important" as a result of dental trainees having "fairly limited exposure to patients due to COVID".

ii) Role of longitudinal clinical assessment

Despite a desire for improved assessment practices (see section 8.4.2.4 below) and the need for stronger evidence on its use (see section 8.4.2.5 below), several student focus group participants advocated longitudinal clinical assessment is "a better measure" of competence then standalone tests of clinical competence. Some even suggested longitudinal clinical assessment should replace standalone clinical competence tests. However, students also acknowledged there was a need for multiple assessment formats within the BDS curriculum.

In contrast, faculty commented they felt longitudinal clinical assessment was not ready to replace other currently used assessment formats. It was suggested longitudinal clinical assessment may have the potential to supersede standalone clinical competence tests but "a lot of work" around assessor calibration and training was required for faculty to "be comfortable" with this proposition.

The need for a panel of assessment methods was strongly supported by faculty, and longitudinal clinical assessment was deemed as a key contributor to the overall assessment process. One faculty participant proposed the results of this study emphasised the importance of using multiple assessment formats.

> "I think you really do need triangulation. Some types of assessments fit certain skills better than other types of assessment. And I think the safety lies in a mixed bag of multiple assessments over a period of time."

> "[Longitudinal clinical assessment must be] part of the jigsaw of assessments that we need to have".

> "[Longitudinal clinical assessment] has to be part of a suite… of a range of assessments that you can use to assess. I think longitudinal assessment of your clinical performance has to be a key element of assessing the capability of a student and someone who's ready to progress or graduate."

> "I think probably the most valuable thing to come out of this for me is the importance of the other assessments and everything working together."

8.4.2.4 Enhancement of assessment practice

i) Student assurance

Student focus group participants suggested their trust in longitudinal clinical assessment (and therefore the results of this study) would be improved if they were convinced assessors are well calibrated.

> "I would like to see issues with [standardisation] dealt with first before we go on to say that it can show whether a student is competent or not."

The need for improved calibration was also recognised by faculty participants.

> "[the study results] show us that we need to do even more with calibration."

In addition, although faculty focus group participants stressed that students receive "multiple inputs from different staff members, in different locations, with different ideas", student focus group participants said they would also like

to see proof that all students are assessed by a range of different assessors over the duration of BDS course.

ii) Data collection

Several student focus group participants stressed they would like the number of assessment options available within the LIFTUPP© system to be "streamlined". They felt the large number of criteria on which they can be assessed may have occasionally resulted in them not receiving assessment and feedback on key aspects of a clinical procedure.

> "I think just because there [were] so many other variables that they put in there, a lot of the time they were missing things, and things weren't be assessed."

Faculty were more concerned with obtaining accurate data on the difficulty and complexity of clinical procedures undertaken by students. Recording the number of times students had performed a procedure was also suggested.

8.4.2.5 Future research

i) Larger studies

The small number of participants eligible for this study was acknowledged in both focus groups. This suggests larger studies would be welcome in future to increase the strength of evidence on longitudinal clinical assessment.

ii) Expanded studies

According to faculty, one of the most important focuses for future studies on longitudinal clinical assessment should be incorporation of context of the assessment/clinical encounters, especially regarding procedural difficulty and complexity.

> "Going forward that would be the driver… to make sure that complexity is added in for everyone, and then that is considered in further assessment of the system and further research."

> "But it's just getting your method into shape to be able to take into account all the nuances of complexity of procedure, making sure they have multiple assessors, continued calibration of your assessors, to make it useful in a valid tool."

Faculty suggested future studies should incorporate tracking student progression beyond vocational training "to places like the Royal Colleges" and felt "it was a "shame [Glasgow Dental School previously] didn't have the scoring system" that would have permitted comparisons between longitudinal clinical assessment and clinical case presentations assessments as it "would've been quite good to see [longitudinal clinical assessment] compared to… more clinical performance type data".

Finally, student focus group participants proposed they would like to see studies on whether longitudinal clinical assessment data patterns correlate with student confidence and insight into their clinical performance.

8.5 Summary

This chapter has presented the findings obtained via thematic analysis of transcripts of focus group discussions with key stakeholders in dental education, from which five overarching themes were identified.

Several discussion points under themes 1 (attitudes towards longitudinal clinical assessment data trajectories) and 3 (attitudes towards longitudinal clinical assessment) appeared to offer additional support for longitudinal clinical assessment having content validity. The participants felt the trajectories shown to them reflected their clinical assessment experiences and demonstrated that students' clinical skills improved over the BDS course. However, it should be noted that the participants did not explicitly say they believed longitudinal clinical assessment had content validity – even when they were asked directly. Instead, they suggested the results produced from the quantitative analyses appeared promising but wished to see further evidence before commenting further.

The unwillingness of some participants to directly comment on the evidence for validity may have been due to their unfamiliarity and/or uncertainty on the

LIFTUPP© data analysis approaches adopted in this study - particularly GBTM. Alternatively, some may have been reluctant to present strong views due to potential operational issues with the LIFTUPP© assessment method they had previously experienced.

Participants were keen to discuss operational issues with the LIFTUPP© system, even though they were not asked about them directly. However, the discussions provided were useful as they allowed another overarching theme (theme 2: Scepticism on LIFTUPP© data quality and consistency) to be identified and contributed suggestions on how assessment practices could be improved (theme 4: Enhancement of assessment practice).

Discussions under themes 4 and 5 (future research) provided valuable data on how assessment within dental education could be enhanced. The former identified that students were seeking assurances on longitudinal assessment practice and there was a need to improve data collection and/or accessibility for LIFTUPP© and other undergraduate assessment methods – namely standalone tests of competence. The latter provided suggestions for future lines of inquiry and how stronger evidence on longitudinal assessment might be obtained.

Further discussions on the findings presented in this chapter, and how they related to other results presented within this thesis, are provided in the subsequent chapter (9).

Chapter 9 - Discussion, conclusions, and recommendations

9.1 Introduction

In terms of results, this study has contributed early evidence to a validity argument on the use of longitudinal data for assessing the development of clinical competence in undergraduate dental students.

It is appreciated that assessors in dental education wish to see evidence that longitudinal data can be used to identify students who consistently underperform, as it could serve as a means of supporting students through identifying learning needs, directing training resources towards addressing these needs, and monitoring performance following delivery of any remedial training or other interventions (Field et al., 2017). However, these potential benefits can only be fully realised if longitudinal data have been shown to have good psychometric properties and benchmarks for varying degrees of longitudinal performance (e.g., “good”, “satisfactory” and “poor”) have been established. This study has taken an initial step towards investigating the former as it has shown longitudinal clinical assessment appears to have a degree of content and criterion validity since its data patterns reflected how students would be expected to develop over the BDS programme and could be related to other well-established assessment methods which assessors have confidence in. This should increase confidence in the use of longitudinal assessment data in determining development of clinical competence.

The final chapter of this thesis will now provide further discussions on key findings previously presented in chapters 4-8. Sections are primarily presented in the same order as the results over the previous chapters, however (where applicable), qualitative findings generated through the focus groups will be triangulated and discussed in tandem with findings from statistical analyses and data modelling. Limitations and strengths of the study are also discussed.

The thesis concludes with a list of recommendations for current assessment practices within dental education and suggestions for future research, which have been established through the work conducted in this study.

9.2 Undergraduate examinations

9.2.1 Examination outcomes and dental student cohorts

Analyses of the undergraduate examination outcomes revealed that very few dental students fail their academic assessments (chapter 4, section 4.4.1). This may, in part, be because a significant portion of dental student cohorts consist of individuals who are high achievers and very capable academically, as most dental students have had to satisfy rigorous course entry requirements to obtain a place on BDS programmes. For example, pupils from Scottish secondary schools applying to study dentistry at the University of Glasgow typically must have obtained at least four "A" grades and one "B" grade in their Higher examinations to be eligible for admission (University of Glasgow, 2021b). They must also perform well in the University Clinical Aptitude Test (UCAT) and a joint interview/admissions test (known as Multiple Mini Interviews (MMIs)) (Eva et al., 2004; Cleland et al., 2012; Eva et al., 2012; Husbands and Dowell, 2013). Applicants are also expected to have obtained some work experience/shadowing in dental practice to gain insight into whether dentistry is a career they wish to pursue.

It is, however, also worth acknowledging that meeting "traditional" entry requirements is no longer the only means thorough which applicants can gain access to dental school. Universities operate a contextual admissions system for applicants (Dental Schools Council, 2021; University of Glasgow, 2021c), which enables them to receive adjusted entry requirements (e.g., grade concessions and/or UCAT concessions). Furthermore, there are now "access courses" available for "adult returners" (who were previously unable to progress to university directly from secondary school), which offer an alternative route of entry to dental school (Greater Brighton Metropolitan College, 2021; University of Glasgow, 2021a; Medical Schools Council and Dental Schools Council, Accessed 2021).

Due to small numbers ($n<5$), the exact number of failures could not be presented per examination, BDS year and/or cohort since it may compromise anonymity. However, across all three cohorts, nine individuals left the BDS course due to unsuccessful examination outcomes, all of which were recorded in the early BDS

(1-3) years. Others who were unsuccessful in the examinations chose to repeat the relevant BDS years and subsequently passed. This included all students who failed in either of the final examinations (held in BDS4/5). These results may imply that most of those who are not suited to studying dentistry leave the BDS course during the earlier stages of undergraduate training. However, it should be noted that some students may leave the course for reasons other than a lack of academic attainment. In the case of this study, the number of students who left the course for reasons other than a lack of academic achievement was also very small (<5) and therefore, to preserve anonymity, exact figures could not be presented.

This finding may provide some reassurance since, if individuals are unfit to practise or no longer wish to pursue dentistry as a career, it is best they leave the course at the earliest stage possible. This increases patient safety, allows individuals to move onto another course of higher education or career pathway sooner, and helps keep financial losses incurred due to uncompleted training as low as possible.

The results may also be indicative of “failure to fail”, a well-established problem within HCP education, including dental education (Yepes-Rios et al., 2016). “Failure to fail” occurs when assessors are reluctant to fail students once they are admitted onto the course (Dudek, Marks and Regehr, 2005; Cleland et al., 2008) and can occur for a variety of reasons (Chambers, 1993; Chambers, 1998; Licari and Chambers, 2008; Bush, Schreiber and Oliver, 2013). A systematic review by Yepes-Rios et al. (2016) summarised the main causes of “failure to fail” into six categories: 1) evaluator’s professional considerations; 2) evaluator’s personal considerations; 3) trainee related considerations; 4) unsatisfactory evaluator development and evaluation tools; 5) institutional culture; and 6) consideration of available remediation for the trainee. This review initially screened over 5000 publications from across medical, dental, and nursing literature before evaluating 28 publications in detail and, to date, appears to the most comprehensive review of literature available on this topic for HCP education.

Although investigations into potential occurrences of “failure to fail” was not an intention of this thesis, it is important, in the absence of further information,

not to rule out this long running issue within dental education as a potential explanation for the returned results as it can lead to serious implications regarding patient safety (Cleland, Arnold and Chesser, 2005; Eva et al., 2009).

Further analysis of the data revealed that the most frequently awarded grade across most of the undergraduate dental examinations typically alternated between "B" and "C". However, the most frequently awarded grade for the BDS2 OSCE was "A" in all three cohorts. This could raise questions with respect to the validity of the assessment but is likely explained by the relatively small number of clinical skills available for assessment at this stage of the course. As a result, students may have more time to concentrate their learning and practise a select number of skills compared to subsequent BDS years (3-5), which introduce a greater range of skills, and this appears to be reflected through better performance in the BDS2 OSCE. It is also worth noting that, even though students appear to perform very well in this assessment, the BDS2 OSCE ensures they can proficiently perform core clinical skills before gradually building their skillset further over BDS3-5.

9.2.2 The relationship between early (BDS1-3) examination performance and finals (BDS4/5)

Comparisons between early undergraduate examinations and finals showed moderate to high positive correlations between each of the early examinations and the finals (see chapter 4, section 4.4.2). BDS3 examination performance displayed the strongest correlation with finals and, based on the c-statistics returned (appendix 8), its outcomes appeared to be the most indicative (predictive) of finals performance compared to other early examinations. However, there were exceptions to these findings. In cohort 3, the BDS2 examinations displayed the strongest correlation with the final MSA (BDS4) examination, and, in cohort 2, the BDS2 examinations displayed the highest predictability for the final MSA (BDS4) examination. Potential causes for these exceptions could not be readily explained from the available data.

Increased predictability of examination performance between BDS3 and the final examinations may be due to increased similarity between the content of the BDS3 and final examinations compared to BDS2. The earlier years of Glasgow's

BDS course have a greater focus on various “basic” scientific subjects (such as biochemistry, immunology, and physiology), before gradually transitioning to a greater focus on clinical sciences and practice in the latter years. Therefore, BDS2 examinations contain more items assessing comprehension and application of “basic” scientific subjects compared to BDS3 examinations, which have a greater emphasis on clinical sciences and practice. BDS2 examinations still contain clinical items, and scientific questions may still appear in the BDS3 and final examinations but are usually only a small component of a clinical scenario.

The transition from school to university might explain the lower correlation and predictive capacity between the BDS1 and final examinations. This transition can be an extremely challenging and stressful time for students and failure to adapt well can result in poorer academic performance (Yorke and Longden, 2004; Hommes et al., 2012; McMillan, 2013; Bowman, 2017; Hassel and Ridout, 2017; van Herpen et al., 2020). Therefore, it may take time for some students to settle into life at dental school before they start performing well in the assessments.

Cross tabulations between early and final examinations indicated dental students who performed well in early examinations were considerably more likely to do well in finals, and those who performed less well in the early assessments were less likely to perform well in finals. However, the degree of variation between early and the final examinations suggested there was scope for students to improve their performance over time, i.e., students performing less well can develop into high achieving students over the duration of the BDS course. It is also worth remembering that each of the undergraduate examinations are “one-off”/standalone tests which come down to performance on the day. The “good-day, bad-day” phenomenon has previously been acknowledged within medical education subjects (van der Vleuten and Schuwirth, 2005; Rauf, 2021) and it is not unreasonable to assume that some of the results used in the analysis for this study were due to some usually high performing students simply delivering a poor performance at the time of assessment (and vice versa).

There may be a variety of underlying reasons explaining how students may improve their examination performance (e.g., approaches to learning (Good,

Ramos and D'Amore, 2013; Ghazivakili et al., 2014; İlçin et al., 2018), assessment literacy (Price et al., 2012) and examination readiness (Weinstein and Wu, 2009; Heinicke, Zuckerman and Cravalho, 2017)) but it was beyond the scope of this study to investigate these in detail. Therefore, further investigation using additional student cohorts is required to confirm the findings presented here. There would also be merit in exploring if there is a point in the BDS course where students who are performing less well begin to improve and develop into better students. This point of student development could be regarded as a "threshold concept" (i.e., a concept which, once understood, transforms one's perception and thinking of something (Land, Meyer and Smith, 2008)), which is a topic that is being increasingly discussed in both medical (Neve, Lloyd and Collett, 2017) and dental (Kinchin et al., 2011; Bowman, 2017; Green and Rasmussen, 2018) educational literature. However, at present, publications on threshold concepts appear to focus primarily on describing what threshold concepts are, how they can be identified and how they might be used to develop curricula in higher education. There are few which present robust evidence on thresholds being "crossed" by students. This lack of evidence has led some authors to question the degree of support for threshold concepts and highlight that there are still significant definitional and methodological problems surrounding the theory which are hindering research (Stopford, 2020; Salwen, 2021).

9.3 Undergraduate longitudinal clinical assessment - LIFTUPP©

9.3.1 Student clinical experience

Due to the rigorous inclusion criteria (chapter 3, section 3.5.3.4), the number of LIFTUPP© clinical assessments eligible for this study only reflected 60.5%, 57.1% and 57.8% of student clinical experience in cohorts 1, 2 and 3 respectively (chapter 5, section 5.4.1.1). For a procedure to be eligible, an individual student had to have performed (and been assessed on) all the key stages of a treatment item (appendix 2) on an individual patient.

Since students are timetabled to rotate between different treatment centres (especially during the BDS5 outreach programme, where they gain the most

patient contact), some students will not complete all the relevant stages for procedures spread across multiple visits (e.g., root canal treatments, crown and bridgework, and dentures) on the same patient. Treatment may be commenced by one student, who is then replaced by another student timetabled to attend the same treatment centre. The replacement student then completes the treatment for the patient.

It can be argued students undertaking parts of a procedure is reflective of working in practice and, therefore, a worthwhile training experience. For example, a patient may attend a clinic in an emergency due to toothache and is seen by someone who is not their regular dentist. The replacement dentist may commence treatment to provide initial pain relief before referring the patient back to their regular dentist for the treatment to be completed later. Alternatively, some procedures may not have been completed due to patients failing to return for continued treatment. Based on the analysis conducted in this study, it would appear scenarios like these, which lead to individual students not overseeing single procedures in their entirety, may occur approximately 40% of the time.

The LIFTUPP© system can easily record student experience in performing individual procedural stages (or the components of a procedure). However, determining how competent students are in performing a procedure becomes more complicated if assessors are required to piece together parts of an assessment. It stands to reason that a clearer indication of student competence can be drawn from procedures which have been completed from beginning to end, hence why this study opted to focus on procedures completed in their entirety by the same student. Imposing these strict inclusion criteria also facilitated investigation of student competence under the premise that their overall performance for individual treatment items was only as good as their lowest performance indicator for a single procedural stage (see chapter 3, section 3.5.3.4).

Additionally, in cohort 1, some clinical procedures were not assessed and recorded during the early stages of LIFTUPP©'s implementation, and, in both cohort 1 and 2, some procedures were not attributable to individual patient encounters and had instead been assigned per clinical session (as previously

described in chapter 3, section 3.5.3.4). The inclusion criteria resulted in procedures affected by these circumstances being discounted, further reducing the number of eligible procedures and, therefore, the proportion of the students' overall clinical experience represented in the study.

9.3.2 Differences between the three student cohorts

Initial descriptions of LIFTUPP© data revealed a clear difference in the total number of eligible clinical assessments in cohort 1 compared to cohorts 2 and 3. Although each subsequent cohort provided more clinical data eligible for inclusion than the previous, the greatest increase was observed between cohorts 1 and 2. There were also differences in the distribution of LIFTUPP© performance indicators between cohort 1 and cohorts 2 and 3 – with the two most recent cohorts recording a greater number of slightly lower performance indicators. These observations were noted when analysing the data from both student and assessor perspectives (see chapter 5, sections 5.4.1.1-5.4.1.4), and may indicate assessors were, over time, becoming more familiar (and confident) with the assessment options available within LIFTUPP© and assigning performance indicators more in line with the guidance. Cohort 1 were the first group of students to have had all undergraduate clinical activity assessed via LIFTUPP© and a "settling in" period was to be expected. Students focus group participants acknowledged that variation in assessment within the first year may have been due to assessors adapting to a new format of clinical assessment (see chapter 8, section 8.4.2.2).

Furthermore, assessor calibration exercises were introduced following the system's adoption, as were training sessions specifically directed towards combating "failure to fail". These exercises were conducted every 6-12 months and may have resulted in alignment of the performance indicators with the guidance in the more recent cohorts. However, the impact of assessor training on scoring remains disputed. Some studies suggest examiner training does impact scoring and assessor variability (Holmboe, Hawkins and Huot, 2004), whereas others remain unconvinced (Cook et al., 2009; Gauthier, St-Onge and Tavares, 2016) (see section 9.6.1 for further discussion). Although the results presented in this thesis detected a difference in assessment patterns between cohort 1 and cohorts 2 and 3, it does not provide sufficient evidence to suggest

assessor calibration exercises influenced the LIFTUPP© data patterns since their impact was not investigated as part of the study.

9.3.3 The impact of changes to LIFTUPP© performance descriptors

It was initially anticipated that, as described in chapter 3 (section 3.5.3.2), modification of LIFTUPP© performance descriptors at the end of 2014-15 would have impacted on the frequency of performance indicators awarded between cohort 1 and cohorts 2 and 3. However, based on the frequencies presented in this study, this appears not to have been the case.

The most significant changes in the descriptors were centred around performance indicators 3 and 4 (see chapter 3, section 3.5.3.2, Table 3.1). Under the original descriptors, assessors could be forgiven for interpreting a performance indicator 3 as the minimal level of satisfactory clinical performance, even if they provided students with physical assistance to complete the procedure. By contrast, the updated descriptions suggest any physical assistance by a supervising clinical cannot be regarded as independent practice and, therefore, a performance indicator of 3 cannot not be considered as the minimal level of satisfactory clinical performance. However, the shift in frequencies of performance indicators between cohort 1 and cohorts 2 and 3 was not centred around the number of 3s and 4s awarded. Instead, it was centred around the awarding of 4s and 5s - with the former being awarded more and the latter being awarded less in the two more recent cohorts. The frequency of 3s appeared consistent across all three cohorts (see chapter 5, sections 5.4.1.2 and 5.4.1.4). This suggests the modifications to the descriptors appear to have had little impact on the frequency of performance indicators awarded.

9.3.4 LIFTUPP© assessment experience among assessors

Analysis of LIFTUPP© data revealed a wide range of clinical assessment experience among assessors. Some assessors had assessed every student in a cohort and/or provided thousands of procedural stage assessments. Others had only assessed a single student in a cohort and provided a single LIFTUPP© assessment entry (see chapter 5, section 5.4.1.3). The latter could be

problematic since assessor experience may play a role in assessment reliability (Baker et al., 2008; Suto and Nadas, 2008; Suto, Nadas and Bell, 2011).

Ideally, no assessors should be assessing only once as they would be unable to build a point of reference for their own assessment practice. The results of this study showed 25% of the assessor cohort, on average, assess up to ten students and provide up to 150 assessments (chapter 5, section 5.4.1.3), therefore, assessors who have extremely little assessment experience were rare. However, it may be that incidences of assessors with very little assessment experience were due to LIFTUPP© log-in issues, which can arise spontaneously, but were especially prevalent following the initial introduction of the system. Some assessors had not been registered as users of the system and others had problems gaining access despite being registered. As a result, some assessors resorted to using another's log-in details and, therefore, the degree of assessment activity for some assessors will be diluted and exaggerated accordingly. Alternatively, very low assessment activity may simply be aligned with "test" and/or redundant assessor profiles within the LIFTUPP© system. However, this could not be investigated since assessor names had be removed from LIFTUPP© data sets by data management staff at LIFTUPP© Ltd prior to their transfer to the third-part data analyst involved in this study.

Although this study found, on average, each assessor had assessed approximately 30 different students per cohort, it did not investigate how many different assessors students had been assessed by. These are data that students recruited for the focus groups said they wished to a see as an assurance that all students are assessed by a range of different assessors (see chapter 8, section 8.4.2.4).

Ensuring students are assessed by a range of assessors has been recommended by various authors (Norcini et al., 2011; Carraccio and Englander, 2013; Harris et al., 2017). It may counter-balance potential extremes of assessor, i.e., "hawks" (very strict) and "doves" (lenient) (McManus, Thompson and Mollon, 2006; Hodges, 2013; Lockyer et al., 2017), facilitate triangulation of information on performance so an accurate representation of a student's skills and abilities can be formed (Hodges, 2013; Hoang and Lau, 2018) and increase assessment validity (Downing, 2003; Harris et al., 2017; Royal, 2017).

Based on the focus group discussions, variation among assessors (i.e., “hawks” and “doves”) is an issue that students are suspicious of, hence why they wanted to see proof of all students being assessed by multiple assessors. The desire for students to be assessed by multiple assessors echoes suggestions from a previous study by Uma et al. (2017)). However, simply increasing the number of assessors does not guarantee that students will be assessed reliably if all the assessors are poorly calibrated against the desired standard. It may be better to have a small group of well calibrated assessors overseeing student assessment instead of a large group of poorly calibrated assessors, although evidence to support this supposition appears to be scarce within the existing literature.

Faculty focus group participants asserted all students are assessed by a range of assessors and that they use data available within LIFTUPP© to help ensure this. They also suggested that proof of this is shown to students who believe they have been unfairly assessed. Therefore, there may be a case for making these data more accessible to all students to provide them with the reassurances they request. Potential variability and bias among assessors has been well documented and discussed within the wider literature (McManus, Thompson and Mollon, 2006; Gormley, 2011; Bartman, Smee and Roy, 2013; Berendonk, Stalmeijer and Schuwirth, 2013; Jonge et al., 2017; Patel et al., 2018; Coetzee and Monteiro, 2019; Desy et al., 2019), which may propagate student suspicion of their occurrence. However, both the degree of calibration among assessors and the incidence of assessor extremes (“hawks” or “doves”) in relation to longitudinal clinical assessment have yet to be meaningfully explored.

9.3.5 Application of LIFTUPP© assessment criteria and the threshold for competent clinical performance

Although assessors participate in team calibration exercises in relation to the application of LIFTUPP© descriptors/performance indicators, in practice there is some disagreement between groups of assessors as to which descriptor/ performance indicator marks the threshold for competent clinical performance. Some staff still consider performance indicator 4 as the threshold whilst others suggest the threshold is 5. Having a well-established threshold is important since it creates a target or benchmark which must be met. The Quality Assurance Agency for Higher Education (QAA) have described the target as the “threshold

standard" within their benchmark statements for various subject areas, including dentistry (QAA, 2002). This is "the minimum acceptable level of achievement which students must demonstrate to be eligible for award of an academic qualification" (QAA, 2018; Heriot-Watt University, 2019), which is relevant for dentistry and the BDS course since students must develop to the level of the "safe beginner" (GDC, 2015a) by the end of BDS5. The role of the GDC in defining the threshold standard is acknowledge by the QAA in their benchmark statement for dentistry (QAA, 2002).

In this study, performance indicators 4 and 5 were both used as part of the investigations into undergraduate longitudinal clinical assessment data patterns. However, as described in chapter 5 (section 5.4.2.2), a threshold performance indicator of 4 could only be used to investigate content validity since it did not distinguish different groups of student clinical performance in all three cohorts. Threshold models using a performance indicator of 5 were able to provide distinction between groups and therefore could also be used to investigate criterion validity (see section 9.5).

The data presented in chapter 5 (section 5.4) may add further to the debate on the validity of using performance indicator 4 as the threshold for competence performance since less than 10% of all undergraduate student patient work was assigned a performance indicator <4 in all three cohorts. This means over 90% of clinical procedures across all BDS years were performed by students to the desired standard with minimal assistance – a finding which could be challenged, especially by some assessors who have significant experience of supervising dental students. However, it should be remembered that this study has only investigated student clinical work undertaken between BDS3 and BDS5. Prior to BDS3, students complete a series of pre-clinical skills courses and must pass practical competence tests on simulated/mannequin patients (known as "Phantom Heads") before they are permitted to perform procedures on members of the public. Therefore, students who are not yet safe to perform these procedures should not be attending BDS3-5 clinics, which might account for the small number of performance indicators <4 recorded.

Further investigations on student development over the duration of the pre-clinical skills courses may provide additional information on this topic. Glasgow

Dental School began using LIFTUPP© to assess student clinical performance in pre-clinical skills settings in 2015-16, but it was initially not used in the same manner as it was for patient clinics (students were using the system for self-assessment instead of being assessed by a qualified dentist). This was changed for the 2017-18 academic term, but it was not possible to include pre-clinical skills in this study since the input of these data were not consistent across all three cohorts. However, they could be included in future studies based on LIFTUPP© data from 2017-18 onwards. Additionally, data from BDS2 patient clinics could also be included. These data were excluded from this study since they had not been recorded for the first cohort.

Alternatively, the low number of performance indicators <4 may once again reflect the long-standing problem of "failure to fail" (Dudek, Marks and Regehr, 2005; Cleland et al., 2008) within dental education (see section 9.2.1). Even though the LIFTUPP© system seeks to make assessment more objective through provision of descriptors which define the awarding of performance indicators (see chapter 3, section 3.5.3.2), there is no guarantee assessors will always abide by the criteria or they may interpret the assessment criteria differently based on their own clinical and assessment experience (Wilby et al., 2019). Berendonk, Stalmeijer and Schuwirth (2013) also suggested assessors may be tempted to make comparisons between individual student performances (i.e., students are benchmarked against one another instead of the assessment criteria). Furthermore, Gingerich, Regehr and Eva (2011) have suggested assessor's judgements can be influenced by a variety of cognitive factors (such as mood, impression formation, and interactions with previous individuals), which (again) may result in assessors (unconsciously) deviating from the application of strict assessment criteria.

Within the focus group discussions, student participants were suspicious of assessors not following LIFTUPP©'s assessment criteria as intended (see chapter 8, section 8.4.2.2). LIFTUPP© has a large list of skills and attributes on which students can be assessed and entering performance indicators <4 requires assessors to enter written feedback into the system to complete the assessment. These features of the LIFTUPP© system means completing assessment entries can be time consuming, which is problematic on clinics which are running late or

have little time remaining. As a result, assessors may be more selective on what students are assessed on and be tempted to quickly award performance indicators ≥4 (which don't require written feedback) to ensure the clinic finishes on time, even though lower performance indicators were merited. Inevitably this will lead to an inflation in recordings of performance indicators ≥4. This supposition was echoed in the focus group discussions, where students raised the issue that they may sometimes be awarded performance indicators ≥4 out of convenience instead of those that were warranted.

Further discussion on the awarding of performance indicators and assessor calibration is provided in section 9.6.1.

9.3.6 Creating trajectories of LIFTUPP© clinical performance

Despite the small number of participants and cohorts in this study, the LIFTUPP© system provided very large data sets derived from robust (almost real time) electronic data capture. It was possible to apply statistical modelling methods to these data and it appears that the GBTM method shows promise in modelling the large amounts of data (varying over time) gathered by LIFTUPP© since it provided simple - but meaningful – graphical and tabular summaries. These summaries not only allowed content validity for longitudinal clinical assessment data to be explored, but also facilitated comparisons with other assessments (undergraduate examinations and LEPs) to investigate criterion validity since student trajectory group memberships could be considered as outcomes of clinical performance.

9.3.7 Content validity of LIFTUPP©/longitudinal clinical assessment

One of the key objectives of this study was to investigate the evidence for content validity of longitudinal clinical assessment. Content validity can be investigated via a variety of methods - most of which typically require "subject experts" to submit opinions/judgements on an assessment having reviewed the items that make up - or may potentially make up - the assessment. Subject experts' opinions can be obtained using various approaches (e.g., Likert-type questionnaires, rating scales and item similarity ratings (Sireci and Faulkner-Bond, 2014)) and have been used to provide evidence of content validity for

assessment methods used within medical education subjects (including dentistry).

Examples of such studies within dental education include Walters, Osborn and Raven (2005), Macluskey et al. (2011), Barry, Bradshaw and Noonan (2013) and Zijlstra-Shaw, Roberts and Robinson (2017). The former three studies investigated and found evidence for content validity for OSCEs through obtaining the opinions of clinicians/assessors/academics through discussion panels. However, aside from Macluskey et al. (2011), no exact numbers for panel participants were reported. The opinions obtained by Macluskey et al. (2011) stemmed from five assessors. The latter of these studies (Zijlstra-Shaw, Roberts and Robinson (2017)), investigated and concluded there was content validity for an assessment system for professionalism which was designed to encourage student reflection and explanation of their observed behaviours. Evidence was based on opinions obtained from focus groups of academics and students, which each consisted of 4-8 participants.

The opinions of subject experts provide qualitative evidence and are the most frequently adopted approach for determining if assessments have content validity (Ding and Hershberger, 2002; Utkin, 2006). This thesis also used expert (i.e., key stakeholder) opinions to confirm content validity for longitudinal clinical assessment data. However, it also sought to obtain quantitative data which would a) contribute their own evidence for content validity, and b) provide key stakeholders with information on which they could potentially base their opinions on content validity (see chapter 3, section 3.5.2).

Studies which present evidence based on quantitative methods appear to be rarer (Ding and Hershberger, 2002). However, some studies within medical education subjects have attempted to provide evidence for content validity using statistical tests which summarise subject expert opinion, such as the content validity index (CVI). Examples of such studies include Delfino et al. (2013) and Hengameh et al. (2015), both of which investigated the content validity of DOPS (chapter 1, section 1.4.10). The former obtained opinions from six one-to-one interviews with anaesthesia teaching staff and 41 consensus survey responses (11 from staff and 31 from “resident” medics). The latter obtained the opinions of ten academic nursing staff, although the exact means

through which their opinions were obtained was not specified. Both studies calculated the CVI index from the obtained responses and concluded that DOPs had content validity since the CVI was >0.7. However, it should be noted that each of these studies were based on the assessment of either one (Delfino et al., 2013) or two (Hengameh et al., 2015) clinical procedures conducted at single institutions.

In this study, a different approach to investigating content validity was required. LIFTUPP© is a form of authentic assessment since, in accordance with Mueller's (2005) definition, it records student clinical performance on patients in real clinical environments. Therefore, assessors should not be assessing students on something that is not authentic and they (the assessors) should serve as the "subject experts" since they are qualified clinical practitioners who should be familiar with the standard of clinical work expected of a practising dentist. However, LIFTUPP© also produces large data sets (over 300,000 data points across multiple dental subjects per cohort – of which over 100,000 are clinical assessments) and, therefore, to investigate the entirety of the data across three cohorts, student clinical performance needed to be modelled. This allowed the behaviour of LIFTUPP© data to be investigated across all students, in all BDS years, in all cohorts. Individual trajectories could have been produced but they would not have been useful for the purpose of this study (to validate longitudinal clinical assessment data) as general student performance patterns would have been difficult to identify. GBTM offered a means of data reduction to provide a clear illustration of the data, from which the data patterns could be analysed to determine if they behaved as expected (i.e., they were reflective of expected clinical developmental as students progressed through dental school). The patterns were also presented to key stakeholders (i.e., subject experts) in the focus groups to provide triangulation and further determine if the patterns were representative of expected behaviour.

Selection of the best fitting GBTMs was predominantly guided by the BIC. In general, models with the highest (i.e., least negative) BIC are the best fitting (Dekker et al., 2007; Broadbent, Thomson and Poulton, 2008) and this also appeared to be the case in this study. None of the models with the highest BIC in each cohort - stratified by the model data distribution, threshold (for Bernoulli

models) and minimum group restriction numbers - were rejected following additional testing for statistical adequacy or because they appeared less parsimonious than other models when their graphical appearance was scrutinised. However, it remains a possibility that more appropriate models may have been missed despite this approach being consistent with guidelines on model selection suggested by the developers of the GBTM technique (Nagin, 2005; Nagin and Odgers, 2010).

Generally, the patterns produced by the selected GBTMs appeared to suggest that LIFTUPP© - and therefore longitudinal clinical assessment - data exhibit a degree of content validity for measuring development of dental student clinical competence. Regardless of data distribution (censored normal or Bernoulli), threshold performance indicator (for Bernoulli models), or minimum group number restrictions, almost all the trajectories displayed an upward trend between the beginning of BDS3 and the end of BDS5. This indicated students were - on average - being awarded higher LIFTUPP© performance indicators as they progressed through the BDS course, which suggested their clinical skills were improving and they were becoming more independent practitioners over time. Therefore, the upward trends reflected patterns of clinical development that are expected of students as they progress through the curriculum towards the level of the "safe beginner" (GDC, 2015a). These trends and expected behavioural patterns were also acknowledged and confirmed by participants in both focus groups.

As previously detailed in chapters 7 and 8, out of all the models generated, only those based on Bernoulli data distributions and with at least 20 students per trajectory group were presented to the focus group participants. Models with these criteria were regarded as more appropriate for presenting LIFTUPP© data (see chapter 5, section 5.4.3) and the most parsimonious (see chapter 7, section 7.4.5). Therefore, one of the models presented to the focus groups (~Bernoulli model 0 1 for cohort 1) contained a zero-order trajectory. Regardless, the trajectory patterns presented reassured faculty who participated in the focus groups that students' clinical skills were progressing (as anticipated) over the BDS course and that this information was being captured by the LIFTUPP© system (see chapter 8, section 8.4.2.1), which provided additional evidence for

longitudinal data having content validity. However, it should be noted that focus group participants did not directly commit to saying whether they felt the results of the study provided sufficient evidence for content validity. Instead, they insisted they needed to see further studies with additional lines of inquiry before they could be convinced.

Faculty focus group participants suggested they wished to see context applied to longitudinal clinical assessment, as it may help explain the patterns observed within the data. For example, context may decipher whether "dips" in clinical performance were due to students undertaking more difficult procedures, a lack of experience, or genuinely poor performance at the time of assessment. Provision of such information may have helped focus group participants determine whether the LIFTUPP© trajectory patterns were reflective of their reality of clinical assessment, which, in turn, may give them more confidence in commenting on whether there was evidence for content validity for LIFTUPP© data.

Some faculty had expected to see more information from the data – especially with respect to identifying students who were performing poorly. It is appreciated this is a desirable application of longitudinal assessment, but it was not the purpose of this study to model individual trajectories. Therefore, it appears some faculty focus group participants misunderstood the study's intentions. Glasgow Dental School has formed a panel of staff – known as the Clinical Development Panel (CDP) - who meet three times per academic year to interpret LIFTUPP© data and monitor student development. The way in which LIFTUPP© data is automatically displayed for the panel reviews already allows students who are struggling to be easily identified – which was acknowledged by some of staff based at Glasgow Dental School during the focus group discussions.

Since there is currently very little literature on utilising longitudinal data for assessment of clinical development, not only within dentistry but across other disciplines (such as medicine and nursing), there is little opportunity to relate the findings of this study to others. However, the results presented reflect a study by Roudsari (2017) which modelled LIFTUPP© data using an alternative technique (Bayes theorem) to demonstrate an upward trend in student clinical performance (extractions in oral surgery) over time at cohort level. Although

both Roudsari's study and the results presented in this thesis illustrate upward trends within LIFTUPP© data, there remains some debate over what these findings ultimately mean and there are several questions which need some consideration. These include: Does where students start in their development really matter? Are the ways in which students progress to the end point important? At what point would students be considered to have developed clinical competence?

Answering these questions require the data to be analysed at individual student level. Roudsari's (2017) study originally explored modelling LIFTUPP© data at student level, however they were not able to do so due to the small number of tooth extractions performed by each student in each BDS year, which resulted in extremely wide confidence intervals within their model, making interpretation of data meaningless. The study presented in this thesis never intended to model data at individual student level (it sought to find and analyse groups of student performance within cohorts) and therefore was not designed to provide answers to these questions, but it is worth remembering that these are avenues which should be considered for future studies on longitudinal clinical assessment.

9.4 Postgraduate longitudinal clinical assessment

9.4.1 Variability in the number of clinical longitudinal evaluation of performance assessments across cohorts

Descriptive statistics of the LEP data revealed a wide range in the number of clinical assessments completed by VDPs in each cohort, with some VDPs completing as few as 19 clinical assessments (in cohort 1), and others completing as many as 56 (in cohort 2) (see chapter 6, section 6.4.1.1). However, although VDPs are required to complete at least 42 LEPs to satisfactorily complete the VDT year, they do not all need to assess hands-on clinical skills. Some LEPs may focus exclusively on communication skills, others on management and leadership skills etc. If LEPs which did not assess clinical skills were included, all VDPs completed at least 42 LEPs in cohorts 1 and 2.

The COVID-19 pandemic will have impacted upon on the number of assessments completed in cohort 3 since clinical work for all dentists was greatly reduced as

a part of the national lockdown in Scotland - which commenced 24th March 2021 (Scottish Chief Dental Officer, 2020; Scottish Government, 2020). However, the number of LEP clinical assessments completed in cohort 3 was not much less than cohort 1. Cohort 1 only completed 338 more assessments than cohort 3 and, on average, cohort 1 VDPs only completed 1.6 more clinical LEPs more than their cohort 3 counterparts (means 34.2 vs. 32.6). This was despite cohort 1 having seven more VDPs than cohort 3 (67 vs. 60). It was also noted that cohort 2 completed 545 more clinical LEPs than cohort 1 (means – cohort 1:34.2 vs cohort 2: 40.6), and 883 more than cohort 3.

Aside from the COVID-19 pandemic, it was difficult to account for the range in clinical assessment numbers between the cohorts from the available data. The pool of trainers involved in the VDT scheme can vary each year, and it may be that trainers (and their VDPs) in cohort 2 were eager to complete as many LEP assessments as possible compared to the other cohorts. Another possible explanation is that cohort 2 trainers may (by chance) have been able to commit more time to VDP assessment (e.g., due to their own patients cancelling or failing to attend appointments). A further possibility is that some VDPs in cohort 2 were not performing satisfactorily and therefore required additional LEP assessments. VDT trainers are instructed to repeat LEPs on procedures where VDPs obtained “needs improvement” scores (i.e., LEP scores 1 -3) so that further progression can be recorded and monitored. As previously discussed in chapter 3 (section 3.5.3.2), VDPs cannot successfully complete VDT if they have any outstanding “needs improvement” scores (NES, publication year unknown). VDPs who have not been able to address these by the end of block 3 are given a grace “year-end” period to repeat LEPs. No “year-end” LEPs were recorded in cohort 3 due to the COVID-19 lockdown, and 32 and 15 were recorded for cohorts 1 and 2 respectively. However, it should be noted that “needs improvement” scores may be addressed within each of the three assessment blocks of the VDP year, therefore a “year-end” grace period may not be required for some VDPs even if they obtained “needs improvement” scores. In this study, “year-end” LEPs were recorded within block 3 of each cohort.

The degree of variation in clinical assessment experience in LEPs is comparable to the variation seen with undergraduates in LIFTUPP©, where some

students/VDPs complete more than twice the number of clinical assessments than some of their peers (see appendices 9 and 10).

9.4.2 The longitudinal evaluation of performance scoring system and the threshold for competent clinical performance

Unlike LIFTUPP©, and as previously discussed in chapter 3 (section 3.5.3.4), LEPs do not currently have a set of clear descriptors associated with each of their scores. The 1-9 scoring scale is split over three categories of performance (unsatisfactory (1-3), satisfactory (4-6) and superior (7-9)) (Prescott-Clements et al., 2008; NHS Education for Scotland, Accessed June 2021) but the lack of descriptors means it is difficult to specify what level of performance merits the award of scores within each category, i.e., there is no clear difference between the award of a 7 and a 9. This places greater emphasis on the personal judgements of the assessors (both VDT trainers and external assessors), meaning the LEP scoring system is more subjective than LIFTUPP©.

Guidance from NES implies that a LEP score of 4 serves as the minimum standard of performance in VDT (NHS Education for Scotland, Accessed June 2021). However, it would also be of value to see improvement in VDP's clinical performance over the duration of the VDT year as a VDP who only received scores of 4 throughout the year would be brought to the attention of NES. This study investigated LEP data using four different threshold scores (4, 5, 6 and 7) and found a threshold of 4 did not distinguish different groups of clinical performance among VDPs in all three cohorts. This may imply that all VDPs are (on average) performing above the minimum standard required, but it may also (again) be a manifestation of "failure to fail".

LEPs may be more susceptible to "failure to fail" and assessor bias than LIFTUPP© assessments since they are predominantly (but not entirely) completed by each VDP's assigned trainer. The increased risk of single VDT trainers assessing their trainees more leniently was discussed by VDPs and recent graduates during the focus groups. VDPs and recent graduates stipulated that more lenient assessment may occur in LEPs because it was "in the interest of" VDT trainers for their trainees to appear to be performing well.

There is currently no readily available literature which supports or conflicts the assertion that “failure to fail” is more prominent in postgraduate training (for any HCPs) than undergraduate settings, nor that it occurs for reasons other than those previously summarised by Yepes-Rios et al. (2016) (section 9.2.1). However, suspicions that the relationship between a VDP and their trainer could be detrimental to the assessment process (and vice versa) were previously raised as part of a UK consultation report on DVT from 2003, which was based on questionnaire responses from 77 key stakeholders. Subsequent studies have found that trainee-trainer relationships can impact feedback given as part of the WBA process (which is similar to both LIFTUPP© and LEP assessment). A recent example of these studies, Bok et al. (2016), was based on 14 semi-structured interviews with clinical teaching staff for undergraduate medicine. When coupled with multiple other studies (from various HCPs) which have suggested that “failure to fail” may be caused by emotional challenges presented to assessors (Duffy, 2006; Duffy and Hardicre, 2007; Cleland et al., 2008; Gopee, 2008; Carr et al., 2010; Watling et al., 2010; Jervis and Tilki, 2011; Bush, Schreiber and Oliver, 2013; Larocque and Luhanga, 2013; DeBrew and Lewallen, 2014), it may explain why VDT trainers find it difficult to “fail” VDPs with whom they have formed a good personal and/or professional relationship and, therefore, why assessor bias can result.

NES attempt to mitigate for assessor bias (and improve reliability) by insisting that VDPs must also be assessed by an external assessor two- or three-times during VDT (Prescott-Clements, Hurst and Rennie, 2003; NHS Education for Scotland, Accessed June 2021). Although this ensures VDPs are evaluated by at least two different practitioners, they are assessed by a smaller range of assessors compared to undergraduates in dental schools and postgraduate trainees based in hospital settings – where a greater pool of potential assessors is available on-site. Obtaining a range of different assessors in general dental practice settings is much more challenging since external assessors need to travel to the different dental practices to conduct assessments – a difficulty which has previously been acknowledged within the literature (Grieveson, 2002; Prescott et al., 2002). It is also worth noting that fewer methods are employed for assessment of VDPs compared to undergraduates - LEPs being one of the examples. This is predominantly due to time and money constraints placed on

trainers since they work as general dental practitioners and VDP assessors in tandem.

9.4.3 Creating trajectories for longitudinal evaluations of performance

A greater number of Bernoulli GBTMs were returned from LEP data compared to LIFTUPP© data (Cohort 1: 116 [LEP threshold 6] and 195 [LEP threshold 7) vs 17 [LIFTUPP© threshold 5]; Cohort 2: 93 [LEP threshold 6] and 199 [LEP threshold 7] vs 19 [LIFTUPP© threshold 5]; Cohort 3: 30 [LEP threshold 6] and 182 [LEP threshold 7] vs 15 [LIFTUPP© threshold 5] - see appendices 9 and 10). This was because LEP GBTMs were based on less data points compared to those based on LIFTUPP© data. As a result, the statistical modelling plugin (traj) was unable to be as precise or accurate when categorising different patterns of VDP LEP performance data compared to student LIFTUPP© data and instead a greater number of potential models/trajectories could fit the LEP data. This also suggests the LEP models were less stable, with some VDPs potentially switching trajectory groups depending on the number and shape of trajectories requested as part of the data modelling process. A review of the probabilities of VDP trajectory group membership may present further evidence for this observation but it was beyond the scope of this study to review these data for 815 models.

Like for LIFTUPP© data, selection of the best fitting LEP GBTMs was predominantly guided by the BIC (see section 9.3.7) and there remains a possibility that more appropriate models may have been missed using the criteria outlined in chapter 3 (section 3.5.3.6) – even though they are consistent with guidelines on model selection suggested by the developer of the GBTM (Nagin, 2005; Nagin and Odgers, 2010).

Generally, all the trajectories within the selected models displayed an upward trend, which showed that VDP's clinical performance improved over the duration of the VDT year. These findings echo those from a previous study on the validity of LEPs by Prescott-Clements et al. (2008) who created trajectories of performance over the VDT year from the mean LEP scores of two VDP cohorts.

9.5 Comparisons between undergraduate longitudinal clinical assessment and examination outcomes/postgraduate longitudinal clinical assessment

9.5.1 Concurrent validity of LIFTUPP©/longitudinal clinical assessment

Comparisons between LIFTUPP© and undergraduate examinations suggested there was some evidence that LIFTUPP©/longitudinal data have a degree of concurrent validity (in measuring development of competent clinical performance) since students allocated to the better performing trajectories in cohorts 2 and 3 tended to perform better in standalone undergraduate assessments. Although there are no "gold standard" assessment methods within dental education, the validity of OSCEs (Brown, Manogue and Martin, 1999; Hodges, 2003; Park et al., 2004; Varkey et al., 2008; Taghva et al., 2010; Barry, Bradshaw and Noonan, 2013; Nickbakht, Amiri and Latifi, 2013) and MSAs (Sam et al., 2019) has been well publicised. There is less robust evidence to support the validity of MCQs, but they remain one of the most frequently used assessment methods within dental education (Albino et al., 2008; Williams et al., 2015; Roudsari, 2017). Therefore, comparing LIFTUPP© data with MCQ, MSA and OSCE outcomes has provided a starting point for investigating the concurrent validity of longitudinal assessment.

It was noted that the relationship between better LIFTUPP© performance and better examination performance was more pronounced in cohort 2 than in cohort 3. Potential causes for this observation are not yet fully understood. The lack of association in cohort 1 may (again) have been due to varying degrees of calibration among assessors following the initial adoption of LIFTUPP©. However, it could also be presenting a clear finding, as it is not unreasonable to assume some students may perform well clinically but less well in examinations and vice versa.

There is currently very little evidence on the relationship between clinical and academic performance at undergraduate level across the various HCPs. Park, Anderson and Karimbux (2016) serves as one of the few studies within dental education which has investigated this relationship. This study compared OSCE

and case presentation examination outcomes for 185 students (across 4 cohorts) at Harvard School of Dental Medicine and concluded there was a positive relationship between OSCE and case presentation examination outcomes and "quality points" – which were awarded to students based on their didactic and clinical course grades. However, the study did not provide details on how "quality points" were assigned to students and, therefore, it was difficult to establish how student clinical performance had been assessed and classified. Furthermore, the findings were not presented at cohort level, which meant differences between student cohorts could not be detected.

Other studies within dental education have compared clinical performance in pre-clinical skill settings with high school academic performance (Kothe, Hissbach and Hampe, 2014) and dental admissions tests (Gray and Deem, 2002). However, the findings of these studies cannot be directly compared to those presented in this thesis and by Park, Anderson and Karimbux (2016), which have explored relationships between both clinical performance on patients and academic performance within the undergraduate course. A recent systematic review of the literature which investigated the relationship between UCAT (admissions test) scores and assessments in undergraduate medical and dental training (Greatrix, Nicholson and Anderson, 2021) could also not be directly compared for the same reasons.

Cohort 1's results may serve as an example of a graduating class which contained more students who were better clinically than academically or signify that there is little relationship between clinical ability and examination outcomes. Further studies are required to determine if subsequent cohorts also return similar results or those presented by cohort 1 were unique and due to the recent the introduction of LIFTUPP©.

9.5.2 Predictive validity of LIFTUPP©/longitudinal clinical assessment

The results presented in chapter 7 (section 7.4.4) suggested there was evidence that LIFTUPP©/longitudinal data have predictive validity. Almost all cross tabulations between LIFTUPP© and LEP trajectory group memberships showed a greater proportion of students allocated to the better performing LIFTUPP©

trajectory group were more likely to become members of the better performing LEP trajectory group. Comparisons between these assessment methods provided a suitable starting point for determining the predictive validity of LIFTUPP© data since both LIFTUPP© and LEP are measures of longitudinal clinical performance and some evidence for content validity for LEPs had previously been established.

This evidence was provided by Prescott-Clements et al. (2008) using both quantitative and qualitative analyses. The support for content validity was based on the views of 200 VDT trainers (obtained via a comprehensive evaluation questionnaire) and because mean LEP scores increased over the duration of VDT. These results were obtained from only two cohorts of VDPs/VDT trainers; however, it remains one of the few studies across HCP education which has attempted to validate a form of longitudinal clinical assessment; therefore, it was a reasonable approach to compare LIFTUPP© with LEP data.

Other studies within the literature have focused on comparing performance in undergraduate examinations with postgraduate clinical performance. A systematic review by Hamdy et al. (2006), which was based on 19 studies across multiple HCP disciplines, concluded that OSCEs and grades awarded in pre-clinical courses appeared to be predictors for postgraduate clinical performance, but the correlation between them was low to moderate. This review reported widespread heterogeneity in the methods of analysis (e.g., some used Pearson's correlation coefficients, whilst others used regression analysis) and how postgraduate clinical performance was recorded across the included studies. Additionally, there was no evidence for validity for the various postgraduate clinical performance assessment methods. The study also only commented on OSCEs as a summative measure of undergraduate assessment and did not report on the relationship between postgraduate clinical performance and other forms of assessment (e.g., MCQs or MSAs), nor did it report on publications outside of medicine.

In comparison, a subsequent systematic review by Terry et al. (2017) compared various forms of undergraduate assessment with postgraduate clinical performance. This review, which was based on 18 studies across multiple HCP disciplines, also reported widespread heterogeneity on measurement of postgraduate clinical performance, examples of which included a "dental clinical

productivity value" (Graham, Zubiaurre and Anderson, 2013) and a "global rating instrument" (Wilkinson and Frampton, 2004). It also listed which studies had presented evidence for validity regarding their postgraduate clinical performance assessment methods. Not all studies included in the review provided evidence for validity and it was unclear which type of validity (face, content, criterion and/or construct) was associated with each assessment method. The review concluded that, based on the current evidence, OSCEs were (perhaps) the most appropriate means for identifying students that may perform poorly in a postgraduate clinical setting.

9.5.3 Focus group participant opinions on the relationship between LIFTUPP©/longitudinal clinical assessment and undergraduate examinations and postgraduate clinical performance

Most focus group participants felt the relationship between LIFTUPP© and undergraduate examinations and LEPs was a weak one, and some participants had expected to see a stronger relationship between LIFTUPP© and LEPs. These expectations and interpretations of the study results could further explain the hesitance of the focus groups participants to comment on the evidence for criterion validity.

It is perhaps not surprising that there was not a strong relationship between LIFTUPP© and undergraduate examinations and LEPs since the assessment formats are designed to measure different aspects of development in both undergraduate and postgraduate dental training – otherwise there would be no need for different assessment methods. Whilst it is appreciated that LIFTUPP© and LEPs are both forms of longitudinal assessment, the standard of performance against which undergraduate dental students/VDPs are being assessed is different. Dental students are measured against the standard of the "safe beginner" (GDC, 2015a) whereas VDPs are measured against the standard expected for safe independent practice as an associate general dental practitioner (NHS Education for Scotland, Accessed June 2021).

It should be remembered there are a range of factors which could influence an individual's future assessment performance (Shepard, 1993). Examples of factors which might influence a VDP's performance include the practice

environment/atmosphere, their relationship with their VDT trainer and other practice staff, the list of patients they treat etc. Additionally, some individuals may thrive more in a dental practice environment compared to dental school settings, although, having conducted this study, there is currently very little evidence to support this assertion. Overall, there is no guarantee clinical performance at undergraduate level will equate to clinical performance at postgraduate level, much like how good academic performance at high school does not guarantee good performance at university (Kumwenda et al., 2017). Some individuals find the transition from dental school into practice to be challenging and stressful, similar to how some find the transition from school to university difficult (Yorke and Longden, 2004; Hommes et al., 2012; McMillan, 2013; Bowman, 2017; Hassel and Ridout, 2017; van Herpen et al., 2020). It may take time for these individuals to adjust to a new environment, way of working, and their newfound responsibilities.

A previous qualitative study by Ali et al. (2016) found that the transition from undergraduate to working in practice can be challenging and difficult for some young dental professionals. These findings were based on responses from semi-structured interviews with 16 key stakeholders in dental education (academics, undergraduate students, postgraduate trainees and trainers, general dental practitioners, and the regional postgraduate dean). Although Ali et al.'s (2016) study only recruited key stakeholders who were based in south-west England, their findings have been echoed in a more recent qualitive study (Leadbeatter et al., 2020), which also conducted semi-structured interviews with recent dental graduates from the University of Sydney. Furthermore, similar findings have been reported in studies in medicine (Nicholson, 1984; Bogg, Gibbs and Bundred, 2001; Brennan et al., 2010) and nursing (Mitton et al., 2010; Walker et al., 2013; Regan et al., 2017).

Therefore, it was not surprising to see a weak relationship between LIFTUPP© and LEP performance. Regardless, it was important to establish if a relationship between the different forms of assessment existed and (if there was a relationship) whether it was positive or negative. A positive relationship would be indicative of predictive validity and that further investigations would be

merited, whereas a negative (or lack of) relationship would have given more cause for concern.

Some focus group participants commented they had expected to see correlations between the different assessment formats. However, it was not clarified during the focus group discussions whether these participants were referring to relationships between the assessment methods or statistical correlations. Correlations have been used in several other dental education studies investigating both the concurrent (Gerrow et al., 2003) and predictive validity (Foley and Hijazi, 2013; Kothe, Hampe and Hissbach, 2013; Lala, Wood and Baker, 2013; Christersson et al., 2015; Lambe, Kay and Bristow, 2018) of various assessment methods. Whilst this is an appropriate approach for measuring the relationship between continuous variables, it was not relevant for this study since it investigated concurrent and predictive validity using trajectory group memberships, which were a form of discrete data.

9.5.4 Reliability of LIFTUPP©/longitudinal clinical assessment

The way LIFTUPP© data were analysed did not clarify whether longitudinal clinical assessment was reliable. Incorporating each student's longitudinal clinical performance into a single metric to allow the calculation of a reliability coefficient was limiting as simply assigning a binary value of "1" or "0" to each student did not provide the necessary variation. Instead, the probability of membership of trajectory groups was used for each student, which appeared to be the best available metric at the time of study. However, this approach still had limitations and could explain why the reliability coefficient went down when LIFTUPP© was removed (chapter 7, section 7.4.3).

Further consideration needs to be given on how reliability could be explored for LIFTUPP© and which performance outcomes could be used in such investigations. This study used the Cronbach's alpha coefficient, which is the most common means for determining the reliability of assessments in medical (Beckman et al., 2004; Sullivan, 2011; Tavakol and Dennick, 2011; Sharma, 2016), dental (Sharma, 2016), and nursing education (Adamson and Prion, 2013; Sharma, 2016). Cronbach's alpha could be considered again in future studies that adopt a

different metric for measuring student LIFTUPP© performance. Alternatively, reliability could be explored via other approaches, such as:

- "test-retest analysis" – advocated by Leppink and Pérez-fuster (2017) for future investigations on reliability of academic assessments, but still rarely used.

- correlations coefficients (e.g., Pearson's and Spearman's correlations) - used by Al-Osail et al. (2015) to measure the reliability of an 18-station medical OSCE.

- inter-rater reliability (e.g., Kendall's coefficient of concordance (W)) – used by Lin et al. (2013) to investigate the reliability of clinical performance assessment for a subgingival root planing procedure in dentistry.

However, once again, there needs to be further consideration on which metric(s) for LIFTUPP© performance could be used if reliability were to be investigated via any of these means.

9.6 Focus groups with key stakeholders

9.6.1 Standardisation and assessor calibration

During the focus group discussions, student group participants revealed they had experienced - or were aware of - variation between assessors. This seemed to prevent them from fully trusting longitudinal clinical assessment data - even though they were aware that assessors participate in calibration exercises. The same issue was discussed among faculty focus group participants, who acknowledged that assessor calibration remains a constant challenge for all forms of assessment within dental education and there was a need for improvement.

However, although this study summarised three cohorts of longitudinal clinical assessment data from the perspective of assessors (see chapter 5, sections 5.4.1.3 and 5.4.1.4), the current degree of calibration among assessors is not

known. Therefore, there would be merit in conducting more enhanced studies primarily focused on standardisation and the degree of calibration among assessors.

If calibration among assessors is good, with no or few outliers, then students could be provided with proof that assessors are well-aligned with the assessment criteria. If calibration is poor, with wide-spread variation among assessors, then there would be a need for improvement before dental schools can convince students that they are being fairly assessed. Studies investigating the impact of assessor calibration exercises on longitudinal clinical assessment data patterns would also be desirable but could be challenging since their findings may only be attributable to a group of assessors who were evaluated at a specific time.

Previous studies on assessor training have produced variable results in terms of improving calibration. Most studies, across both medical and dental education, have found that, despite using various approaches, assessor training/calibration exercises have very little impact on variability among assessors (Williams, Klamen and McGaghie, 2003; Silber et al., 2004; Boursicot, Roberts and Pell, 2007; Cook et al., 2009; Kogan, Holmboe and Hauer, 2009; Lurie, Mooney and Lyness, 2009; Weitz et al., 2014; Kogan et al., 2015; Gauthier, St-Onge and Tavares, 2016; Crossley et al., 2019). There are some studies in medical education which have reported improvement in scoring consistency among assessors, however the degree of reduction in assessor variation appeared to be minimal. Holmboe, Hawkins and Huot (2004) demonstrated there was only a reduction in the range of assessor scores, and (Wong, Roberts and Thistlethwaite, 2020) found that overall variation among assessors was reduced by <8%.

Despite the current lack of evidence, assessor training/calibration is advocated as good assessment practice by high profile HCP associations, such the Association for Medical Education in Europe (AMEE) (Boursicot et al., 2011), and professional regulatory bodies, including the GDC (2015b).

9.6.2 Refining LIFTUPP© assessment

Some student focus group participants felt LIFTUPP© assessment on some aspects of clinical work – such as "dental chair position" – were of little value and suggested that the number of assessment options with the system should be streamlined to ensure they are assessed on areas of clinical work considered to be of greatest importance (namely their performance for each stage of a clinical procedure). However, retaining a diverse range of assessment options can facilitate precise feedback for other students who require improvement in these areas. Additionally, dental schools are required to provide the GDC with as much proof as possible that their graduates have been adequately trained and are consistently able to perform to the standard of the safe beginner (GDC, 2015a) (chapter 1, section 1.1.2). This body of evidence would be diminished if the options available within LIFTUPP© were to be reduced. Instead of removing assessment options from LIFTUPP©, other approaches could be considered to ensure students receive assessment and feedback on areas of practice they feel are more relevant to them individually.

Assessor training could once again play a role. Parts of clinical practice generally considered to be of greater importance could be identified and a greater focus on the completion of assessment(s) could be encouraged as part of assessor training/calibration exercises. These areas of greater importance could also be emphasised within the LIFTUPP© system itself (e.g., using different colours to highlight key procedural stages etc.). Additionally, greater emphasis could be placed on self-regulated learning (SRL), whereby students drive their own learning through keeping track of their development via LIFTUPP© and seek assessment and feedback in areas in which they require improvement or have not been previously assessed. The importance of SRL in higher education has been well documented within the literature, and multiple studies have advocated its use in undergraduate curricula for medicine (Quirk, 2006; Archbold Hufty Alegría et al., 2014; Cho et al., 2017), dentistry (Bowman, 2017) and nursing (Cadorin, Bressan and Palese, 2017). SRL has been closely affiliated with skills, attributes, and qualities associated with lifelong learning (Zimmerman, 1986; Schraw, Crippen and Hartley, 2006; Dignath and Büttner, 2008; Lüftenegger et al., 2016; Tekkol and Demirel, 2018), and, due to the nature of

HCPs, professional regularly bodies (like the GDC) propose it is vital that graduates are lifelong learners (GDC, 2019b).

Glasgow dental students are advised to use their LIFTUPP© data for SRL. At present, however, this recommendation is delivered relatively informally for BDS2-4 as part of the induction to each academic year. In recent years, more formal training on SRL in relation to LIFTUPP© has been delivered at the beginning of BDS5, but how well students perform and engage with this process has not yet been robustly investigated. A systematic review of 19 dental education articles found there was little information available regarding organised training of students on how to perform self-assessment, nor was there much evidence on how self-assessment impacted clinical performance (Mays and Branch-Mays, 2016). Turan, Demirel and Sayek (2009) previously suggested SRL improves and develops naturally in students as they progress through medical school, although Cho et al.'s (2017) review proposed that Turan et al.'s (2009) results should be met with caution since the study explored changes in SRL across separate student cohorts rather than tracking development within the same cohort. Cho et al.'s (2017) review also suggested that, based on broader literature, SRL does not always develop in some students if they are left to develop these skills themselves (Kruger and Dunning, 1999; Hodges, Regehr and Martin, 2001; Kornell and Bjork, 2007; Kornell and Bjork, 2008).

Therefore, if SRL were to be used as a solution for "streamlining" LIFTUPP© assessment to ensure richer, more valued, longitudinal clinical assessment data sets are acquired, students may need to receive more formal training on how LIFTUPP© can be used for SRL. Protocols for the LIFTUPP© assessment process could also be refined to further encourage student initiative. However, it is also worth noting that all assessment cannot be led by students as they may not have yet developed sufficient knowledge of dentistry to recognise which aspects of clinical practice they need to be assessed on.

9.6.3 Other lines of investigation – future possibilities and current barriers

Focus group participants were interested in whether context (such as procedural difficulty and complexity) had been considered as part of the study's

investigations, and the need for context has previously been discussed in section 9.3.7.

LIFTUPP© permits contextual data to be recorded (for some clinical procedures), and these data were considered for inclusion during the initial study design. However, how well assessors were trained on recording procedural difficulty was not fully known. For some treatment items, LIFTUPP© provides a numerical scale for recording procedural difficulty, however there are different scoring scales in use across the different dental subjects. For example, Restorative Dentistry procedures are rated on a difficulty scale of 1 to 3 and Oral Surgery procedures on a scale of 1 to 5. The difficulty of some clinical procedures (e.g., radiography) cannot be rated at all. As a result, incorporating the context of procedural difficulty across all types of clinical procedure would not be possible unless a uniform rating scale is adopted. Alternatively, procedures for different dental subjects could be analysed separately using the current rating scales – however, this would reduce the amount of data available in each field and the strength of the findings (unless larger studies are conducted).

Faculty focus group participants suggested the number of times individual students had performed clinical procedures could also be used as part of future analyses on longitudinal clinical assessments. These data are readily available within the LIFTUPP© system and Dawson et al.'s (2021) recent study used these data to investigate how valid they are in determining student clinical competence for provision of direct restorations (chapter 1, section 1.4.11). Whilst Dawson et al. (2021) used data on the number of times a procedure has been completed to investigate measurement of competence, the suggestions made by focus group participants in this study appear to refer to using these data to provide context as part of future analyses, which, in turn, may provide explanations for (and confidence in) trajectory patterns produced from LIFTUPP© data.

Participants from both focus groups also suspected the degree of self-confidence students display in their clinical performance could also affect which performance indicators are awarded. Whilst there are many studies which have investigated dental student's confidence in being able to perform clinical procedures (Hunter, Oliver and Lewis, 2007; Lynch et al., 2010; Yiu et al., 2012;

Davey, Bryant and Dummer, 2015; Gilmour et al., 2016; Coe et al., 2018; Hattar et al., 2021), there is currently limited literature on how it impacts performance outcomes. One study has indicated that an increase in self-confidence amongst postgraduate dental students improved their clinical practice (Fine et al., 2019). However, these findings were based on participant opinions and were not supported with clinical/performance outcome data. Another small-scale study demonstrated a weak correlation between self-confidence and examination (OSCE, oral/Viva, case report and written paper) outcomes (Rajan et al., 2020) but no statistical significance was attributable to the results, possibly due to small participant numbers (n = 58). Therefore, there remains a need for further research in this area, especially in relation to longitudinal performance outcomes.

Focus group participants also enquired whether communication, management and leadership, and professionalism data had been incorporated or analysed separately as part of the study. As previously discussed in chapter 3 (section 3.5.3.4), these data were omitted since they had not been aligned with individual student-patient encounters between BDS3 and BDS5 for the first two cohorts. However, the LIFTUPP© system was updated in the 2016-17 academic year to correct this issue and, therefore, subsequent studies could analyse these data for cohorts who graduate from 2019 onwards.

Comparisons between longitudinal clinical assessment and standalone tests of clinical competence (Table 9.1) were another line of inquiry raised within the focus groups. This was another investigation that was considered during the initial study design, however, following initial explorations for potential data sources that could be considered for the study, it became apparent that clinical competency assessment data could not be included. Ideally, data on the number of attempts individual students had taken to successfully pass each competence test were required as well as the dates of successful completion. These data are recorded by Glasgow Dental School but were not readily accessible for all three cohorts at the time of study.

Table 9.1 - University of Glasgow Dental School's clinical competence assessments.

Dental subject	Title of competence assessment
Oral medicine	Management of medical emergencies
Restorative dentistry	Mechanical non-surgical management of periodontal disease
	Provision of indirect restorations: Tooth preparation and provisional cover Gingival tissue management and impression taking Prescribing laboratory work
	Caries Management and bonding/sealing procedures
	Root canal preparation Access and instrumentation Obturation
	Clinical stages in denture provision Master Impressions Jaw Registration
Paediatric dentistry	History from the child patient
	Fissure sealant placement
	Restoration of a primary molar tooth
Orthodontics	Presentation and discussion of an orthodontic patient
	Orthodontic faults and emergencies
Oral surgery	Extraction of an erupted tooth
Radiology	Assessment of two intra-oral films and one panoramic film: Taking and processing radiographs Handling image receptors
Multidisciplinary	Local Anaesthesia

Although standalone assessments have been criticised earlier in this thesis for only providing a snapshot of performance at a single point in time (chapter 1, section 1.5.1), there would still be merit in investigating the relationship between "one-off" clinical competence tests and longitudinal clinical

assessment since previous studies have proposed competence tests have high face validity (Wilkinson et al., 2008; Marriott et al., 2011; Barton et al., 2012). However, if comparisons between longitudinal clinical assessment and standalone competence tests are to be made, the data for standalone tests of clinical competence need to be more accessible.
It is also worth acknowledging that student focus group participants made passing comments that they had expected LIFTUPP© to replace standalone clinical competence tests. This study was not focused on building evidence to suggest that LIFTUPP© should supersede standalone competence tests as the current premise is to ensure dental curricula adopt a variety of assessment methods to establish assessment of the range of skills and attributes expected of dental graduates. This need for multiple forms for assessment was supported by faculty focus group participants in this study, but has also been strongly advocated within the literature (Wilkinson, 2007; van der Vleuten and Verhoeven, 2013), by the ADEE (2010), and by the GDC (2015b).

9.7 Limitations of the study

9.7.1 Quantitative component

The findings of this study are based on data collected from a single dental school and, due to the relatively recent introduction of LIFTUPP© to Glasgow Dental School, only three cohorts of student data could be analysed. Therefore, as anticipated (chapter 3, section 3.5.3.1), the study lacked the power to demonstrate statistically significant differences due to its small sample size.

Additionally, some assessment methods used by Glasgow Dental School could not be included in the analyses - either because they had not been marked on comparable scales for all three cohorts (BDS4 clinical case presentation examination – alphabetic grades in cohorts 1 and 2 vs. numerical scores in cohort 3) or data were unavailable (standalone tests of clinical competence). Inclusion of these assessments may have impacted on the interpretation and strength of the findings or led to additional or alternative findings.

Categorising examination performance into thirds and quarters did not provide sufficient discrimination between groups of students. Thirds and quarters of

performance resulted in the best performing categories (i.e., the top third and top quarter) consisting of “As”, “Bs” and “Cs”, which was not representative of the best performing students. Although fifths of performance provided better discrimination (and ensured the top performing category consisted of students who had achieved the best (“A”) grades), there were either a small number or no students available in each category of performance, meaning it was not possible to discern findings from cross tabulations. These issues were mitigated by cross tabulating thirds of early examination performance with fifths of final examination performance (see chapter 4, section 4.3). However, despite this compromise, small numbers in each of the groups persisted across all three cohorts, resulting in very wide confidence intervals which were not useful for interpretation. This indicates larger group analyses (with involvement from other dental schools) are required to further explore or confirm the findings presented in chapter 4.

Issues with data collection also prevented some lines of investigation being pursued. As previously detailed in chapter 3 (section 3.5.3.4), LIFTUPP© communication, management and leadership, and professionalism data were not included since they were not aligned with single patient encounters for all three cohorts. This meant it was not possible to compare these data with their equivalents in LEP, which had been recorded for single patient encounters.

Whilst clinical domain data were the most prominent type of data collected by LIFTUPP©, aptitude in each of the other three domains are key to the development of the “safe beginner” (GDC, 2015a). Inclusion of data from these three other domains would have meant the study could have contributed evidence on whether longitudinal assessment is a valid method of assessment for the development of competence within and across all four GDC domains of competent practice, rather than just the clinical domain. However, LIFTUPP© was updated during the 2016-17 academic year to align communication, management and leadership, and professionalism data with single patient encounters – therefore, future studies will be able to include these data in their analyses.

A further data analysis limitation was identified during initial screenings of the LEP data sets. Descriptions of the procedure(s) being assessed in a LEP are

entered by VDP trainers using free text. Whilst some trainers provide detailed descriptions, others provide very little, making it difficult to decipher what treatment item(s) were completed by the VDP during the assessment. For example, some trainers simply entered “emergency treatment” on the LEP form – which could refer to various types of procedures (e.g., an extraction, root canal treatment etc.). As a result, it was not possible to analyse LEP data and compare them with LIFTUPP© data at a procedural level - even though LIFTUPP© data could have been analysed in this manner. As a result, the impact of some results may have slightly diminished since the number of data points viable for inclusion was reduced. If this level of investigation were to be pursued in future, the recording of more specific procedural details in all LEP assessments will need to improve (see recommendations - section 9.10.1).

The volume of data for analysis was further reduced by strict inclusion criteria (chapter 3, section 3.5.3.4). As previously reported (chapter 5, section 5.4.1.1), around 60% of all students’ clinical procedural experience were eligible for inclusion in each cohort. Part of the criteria were to ensure that only procedures in which an individual student had completed all key stages of a clinical procedure on an individual patient were included. These key stages (appendix 2) were identified and selected through discussions between the PhD researcher and one of their supervisor team. Whilst both are qualified dental practitioners and experienced clinical supervisors and assessors, there remains a possibility that an element of bias may have been introduced, which subsequently influenced which data were eligible. Furthermore, a panel or committee of other stakeholders in dental education may have produced a different list of key procedural stages, which in turn could have increased or decreased the number of procedures included in the study. However, it can also be argued that the strict inclusion criteria which were applied in this study resulted in more consistent and reliable data for analysis.

Other limitations stemmed from the use of GBTM to analyse both LIFTUPP© and LEP data. Like any statistical modelling technique, GBTM has associated limitations, most of which are due to GBTM functioning as a modelling method which summaries the average behavioural trend(s) of a collection of individuals to create distinct trajectory groups into which individuals can be classified. Data

classification processes are always susceptible to error (Roeder, Lynch and Nagin, 1999) and even though GBTM attempts to base its classifications on the available data, not every individual assigned to a group will follow the group trajectory perfectly. Instead, each trajectory group largely consists of individuals whose developmental patterns resemble one another (and the overall group trajectory) compared to other trajectories (Nagin and Piquero, 2010).

There may be a small number of individuals whose data do not follow the more typical trajectories produced by collections of other individuals within the cohort. If the number of individuals following more atypical trajectories is small, GBTM may not be able to identify the parameter estimates (Nagin and Piquero, 2010) required to generate a unique trajectory group for them to be classified into. Instead, these outlying individuals may be approximated into one of the more typical trajectory groups, resulting in some individuals being "miscategorised". Therefore, in this study, some students/VDPs may have been "upgraded" into a better performing trajectory group or "downgraded" into a group performing less well, meaning a more accurate representation of their development was lost, which, ultimately, will have influenced how the data were interpreted.

Also, the number, shape and accuracy of the trajectories identified through GBTM are strongly influenced by several factors, such as sample size, number of available data points and the timeframe over which data have been gathered (Nagin and Tremblay, 2001; Nagin and Piquero, 2010). Therefore, the trajectories themselves are not definitively fixed nor are individual's memberships to the trajectory groups (Nagin and Piquero, 2010). Although study power and precision of results can be increased through larger sample sizes (Nagin and Tremblay, 2005), GBTMs will always serve as approximations of more complex development behaviour(s) (Nagin and Piquero, 2010).

The notion of using longitudinal clinical assessment data to determine the development of competent practice amongst dental students was investigated through making observations based on existing (secondary) assessment data. There is a possibility that conclusions based on these data are not a true reflection of students' abilities and performance as the undergraduate and postgraduate clinical assessment data stemmed from judgements made by

assessors within a single dental school and postgraduate training scheme respectively.

Furthermore, subjective labels of performance were superimposed onto the trajectories (e.g., “best” performing”, “second-best” performing etc.) which could misrepresent how well students in each of the groups were performing. For example, labelling a trajectory group “second-best” could imply students belonging to that group are not performing well, when (in reality) they are performing well above the required standard.

In accordance with the post-positivist viewpoint, which formed part of the study’s methodological framework (see chapter 3, section 3.3), the introduction of human supposition casts an element of doubt on whether the study findings accurately depict reality. Even though there have been attempts to make assessments more objective by using descriptors to guide the award of performance indicators/scores, strict marking schedules, assessor calibration and double marking, there is no guarantee that the personal standards and conscientiousness of each assessor will not affect the results produced.

There was further scope for error in the study conclusions since they were based on observations made by the PhD researcher. Post-positivism acknowledges that findings may be a product of an investigator’s own theories, background, knowledge and values, all of which can have an influence on any conclusions made (Robson and McCartan, 2016). The PhD researcher attempted to be as objective as possible when analysing the data, but, in the absence of a “gold standard” for dental student assessment against which the study results can be compared, there cannot be complete assurance that an element of bias has not been incorporated into the conclusions reached.

9.7.2 Qualitative component

Further limitations stemmed from the qualitative investigations, which had to be adapted to comply with restrictions imposed by the Scottish and UK governments due to the outbreak of the COVID-19 pandemic at the time of study. Although the use of online video conferencing software allowed both focus groups to be conducted during the pandemic, face-to-face discussions may have resulted in a

different group dynamic between all participants (and between the participants and the moderator). Therefore, additional/different data and conclusions may have resulted.

Those recruited for the focus groups represented a good range of different key stakeholders within dental education; however, there were some key stakeholder groups which were not represented despite being invited to participate, e.g., students from the BDS1, BDS2 and BDS4 year groups and postgraduate VDT trainers. Input from these key stakeholders (especially VDT trainers) would have been valuable and may have also led to additional interactions with other focus group participants, which (again) may have resulted in different conclusions.

Even though participants were shown a small selection of the quantitative results, the figures presented may have initially overwhelmed some focus group participants. The GBTMs can appear simple at first but can take time to understand in full. Additionally, even though the moderator (i.e., the PhD researcher) provided an overview of the methods used to produce the results shown, focus group participants spent time asking the moderator questions on the intricacies of the methodology. Time had to be spent clarifying the purpose of the study since some of the focus group participants were expecting to see results which had identified poorly performing students. As previously detailed in chapter 3 (section 3.5.4.3), focus group participants were provided with an information leaflet in advance (see appendix 4), but perhaps it did not provide enough clarity on the study's aims and objectives.

The methodology questions from participants and the additional clarification on the study's purpose reduced the amount of time spent discussing the quantitative results. Each focus group took approximately two-hours to complete and, as a result, some participants may have become fatigued and were no longer contributing to the discussion as much as they had been during the earlier stages of the discussion.

To mitigate for these potential limitations in the future, a more comprehensive overview of the study's purpose and the methods used to generate the results shown could be provided within the participant information leaflets.

Finally, limitations associated with coding for the analyses of qualitative data may have affected the results presented in chapter 8. Since the study was designed to answer a specific research question on how assessment within dental education could be enhanced, coding was necessary for identifying common themes within the focus group transcripts which would serve as answers to the question. However, deducing and refining large volumes of qualitative data through coding risks stripping out important contextual data, which can occur if the researcher begins assigning the same codes to discussion points that may have merited their own unique code. Furthermore, the coding was conducted exclusively by the PhD researcher in this study. This adds further subjectivity to the study findings since the qualitative results were based on an individual's interpretation of the focus group transcripts, meaning the PhD researcher may have missed themes within the data which otherwise would have been identified. Although attempting to be as objective as possible, there remains a chance that the PhD researcher unconsciously introduced an element of bias in the analysis, which is a risk associated with adopting an interpretivist approach as part of the methodological framework of the study (chapter 3, section 3.3).

The input of a researcher's own beliefs, interpretations, and evaluations (whether consciously or subconsciously) has been referred to as "reflexivity" with qualitative research literature (Berger, 2015). As a concept, reflexivity acknowledges that the researcher is inadvertently part of the investigative process and, therefore, can influence the research outcomes.

9.8 Strengths of the study

This study adopted a novel means of modelling longitudinal clinical assessment data, and record-linked these data to both standalone undergraduate examination outcomes and postgraduate clinical performance to demonstrate content and criterion validity for three cohorts of dental students within a single dental school. It is the first time longitudinal clinical assessment data sourced from LIFTUPP© and LEP have been modelled in this manner. It is also the first-time both LIFTUPP© and LEP data have been linked to other forms of assessment used in dental education. This process created protocols for linking assessment data sets, which, if followed, should expediate subsequent comparative studies. Furthermore, obtaining and linking the different assessment data sets required

successful collaboration between Examsoft® (the software development company who own LIFTUPP©, who were recently acquired by Turnitin® (Examsoft, 2020; Turnitin, 2020)), the University of Glasgow and NES, providing a foundation for future research work focused on progression from dental school into postgraduate clinical practice.

Strong data security protocols ensured anonymity and risks to the participants were minimal, and there was no impact on their academic records or progression through the BDS curriculum. The data linkage process permitted secondary use of routinely collected assessment data - which is not often utilised within educational research - allowing the relationship between current and emerging assessment practices to be investigated.

The results have shown that additional research on the validity of longitudinal assessment would be worthwhile. They have provided a baseline from which future studies on longitudinal assessment can stem. Subsequent research should seek to contribute more evidence for both content and criterion validity and eventually build towards determining whether longitudinal assessment has construct validity.

Since it is important for dental schools to continually review their assessment methods and adopt the best practice in accordance with the available evidence, it stands to reason that stronger evidence on the use of longitudinal data (and its psychometric properties) is desirable. By certifying they are using the best assessment methods available, dental schools will safeguard patients through ensuring that dental students have been appropriately trained and assessed before being permitted onto professional registers (see chapter 1, section 1.1). Furthermore, since the training of dental students bares significant unit costs to the UK public purse, it is essential that dental schools are confident of the utility of their assessment methods and that they represent good value for money.

9.9 Conclusions

There is good evidence to suggest that longitudinal clinical assessment data obtained via the LIFTUPP© system possess both content and criterion validity for determining development of clinical competence.

Content validity is whether the assessment represents what it aims to measure. Evidence for content validity was demonstrated by trajectory patterns of clinical performance trending upwards as students progressed through the BDS course, which indicated students' clinical skills were improving over time. Upward trajectories should be expected because of continued teaching, student learning and development. Therefore, since the assessment approach (i.e., longitudinal clinical assessment) demonstrated this within its data, content validity can be attributed. Key stakeholders in dental education supported this interpretation of the trajectory patterns, providing further evidence for content validity. These findings provided answers for research question 2a.

Criterion validity is how the results obtained via an assessment method correspond to the results of a different assessment method measuring the same thing. There is evidence for criterion validity for longitudinal clinical assessment since there was an association between better undergraduate longitudinal clinical performance and better undergraduate examination outcomes (research question 2b) and postgraduate clinical performance (research question 2c) in the two most recent student cohorts (i.e., Glasgow Dental School's graduating classes of 2018 and 2019).

No conclusive evidence on the degree of reliability for LIFTUPP© data was obtained. This may have resulted from the metrics chosen to calculate reliability in this study. Although a degree of reliability is necessary for valid assessment methods, there was nothing to suggest that answering research questions on 2a-c was inhibited by this lack of conclusive evidence. However, further consideration on how reliability could be explored for LIFTUPP© is still required.

Although focus group participants acknowledged that LIFTUPP© data patterns were reassuring, behaved as expected, and related somewhat to other assessment outcomes (undergraduate examination outcomes and LEPs), the results were met with a degree of scepticism. Primarily, this appeared to stem from previous experiences of operational issues and assessment practices in relation to LIFTUPP©, and from an absence of contextual data for clinical assessments in the analyses. Operational issues and assessment practices have improved over subsequent years since LIFTUPP©'s introduction to Glasgow Dental School but it would be worth investigating if there are persistent

problems with assessor calibration before making further adjustments to the assessment processes. However, there are some changes to assessment practices and data collection which should be implemented based on the results of this study (especially regarding the recording of contextual data for each clinical assessment) to increase confidence in the results obtained. These suggested operational and assessment practice changes provided by focus group participants, along with their recommendations for future research avenues, provided answers for research question 3.

Data collection in relation to assessment methods included (and considered for) this study requires improvement. Data processing revealed that various lines of investigation were impeded due to a lack of detail or consistency across the various data sets. Issues regarding the recording of communication, management and leadership, professionalism and BDS2 data within LIFTUPP© have previously been resolved, however other problems with the collection of standalone clinical competency tests and LEP data remain. In terms of the former (standalone clinical competency tests) additional data on the number of attempts, date of attempts and date of successful completion should be recorded. For the latter (LEPs), text entries should be considered for removal to ensure sufficient details are obtained for clinical assessments. In addition, the LEP scoring system and assessment criteria needs to be refined so there is a clearer, more objective understanding of how VDPs perform in relation to the expected standard.

To build further evidence and strengthen the findings presented in this study, replication of the results in future student cohorts within Glasgow Dental School and comparisons across different dental schools are required. There is also a need to conduct further work using a variety of other research approaches, which will make additional contributions towards a validity argument on the use of longitudinal data for competence assessment.

Overall, the study has provided an early, valuable contribution of evidence towards a validity argument on the use of longitudinal assessment in determining the development of clinical competence in undergraduate dental students. No previous studies have used both quantitative and qualitative approaches to investigate the evidence for the validity of longitudinal assessment, and the conclusions of this study (both positive and negative) should serve as basis for

subsequent research in this area. A list of recommendations on which lines of inquiry should be pursued is provided in section 9.10.

Future studies will also need to adapt to potential upcoming changes in assessment terminology, which may be driven by amendments to the GDC's educational requirements and LOs.

9.10 Recommendations

9.10.1 Current assessment practices and data collection

As a result of this study there are some recommendations that relate to data management and collection, namely:

1. Contact LIFTUPP© developers to refine collection of procedural difficulty information by ensuring the scale on which these data are recorded is consistent across the various dental subjects.

2. Ensure that additional data on number of attempts, date(s) attempted and date of successful completion for standalone clinical competence assessments is made more accessible for research.

3. Refine LEP submission forms to ensure more precise and consistent procedural assessment data are collected. This could be facilitated through digital technology.

There are recommendations associated with assessment policy within Glasgow Dental School:

1. Liaise with senior academic staff at Glasgow Dental School to further refine policies on clinical assessment that further ensure rich and meaningful datasets are created, which can be utilised by both students and assessors for various purposes.

2. Collaborate with senior academic staff to establish how analyses conducted within this study can be integrated as part of Glasgow

Dental School's CDP process for reviewing student clinical activity, performance, and progression.

There are recommendations that are more widely applicable, namely:

1. Work with Glasgow Dental School staff to refine and further develop existing policies for using assessment for learning through feedback based on the validated models presented within this thesis.

2. Streamline and enrich assessment data by providing all BDS years with formal training on how LIFTUPP© can be used for SRL. Assessors will also need to be informed/trained to ensure students engage with this initiative, whilst also being mindful that not all assessment should be student-led.

3. Following appropriate investigation (see section 9.10.2), collaborate with Glasgow Dental School staff to create a data base on assessor calibration that helps maintain calibration records and identifies any increasing variability.

4. Provide students with documented evidence of assessor calibration and that all students are assessed by a range of assessors (following appropriate investigations – see section 9.10.2) to increase student confidence in longitudinal clinical assessment.

5. Liaise with NES to refine the LEP scoring system by creating descriptors aligned with each score. These descriptors should provide a summary of how VDPs performed in relation to assessment standards.

6. This study has highlighted there is potential to enhance the scope of assessment for postgraduates in the spirit of the continuum of lifelong learning, which NES may wish to consider. Clearly there are associated funding implications.

9.10.2 Future research

Based on the work and findings of this study, there are recommendations for subsequent research related to longitudinal assessment of dental students, namely:

1. Determine whether students are assessed by a range of different assessors and how well calibrated assessors are for longitudinal clinical assessments methods.

2. Conduct studies on how assessor training/calibration exercises may influence longitudinal clinical assessment patterns.

3. Explore how LIFTUPP© outcomes (other than group membership probabilities) can be used to determine LIFTUPP© reliability.

4. Repeat investigations described in this study for subsequent cohorts of dental students at the University of Glasgow, incorporating data from the BDS4 clinical case presentation (from the 2016-17 academic year onwards), the BDS4 unseen clinical case assessment (to be introduced in 2021-22), LIFTUPP© communication, management and leadership, and professionalism data (from 2016–17 onwards) and BDS2 LIFTUPP© data (from 2017-18 onwards). Contextual data (procedural difficulty, performance consistency and number of times a procedure has been performed by students) should also be incorporated into subsequent investigations.

5. Investigate longitudinal clinical assessment trajectory patterns for specific clinical procedures (e.g., direct restorations).

6. Conduct larger, expanded studies across multiple dental schools which have adopted longitudinal clinical assessment methods. Longitudinal assessment data from other disciplines (such as medicine, nursing, and veterinary medicine) could also be investigated.

7. Expand study to include pre-BDS course assessment data (e.g., MMI outcomes) and post-BDS course assessment data (e.g., Royal College examinations and WBAs conducted in postgraduate training posts).

Reflection

As a clinical lecturer, the exciting opportunity to complete a PhD study for both my professional and personal development was one I couldn't turn down and remain grateful for.

I have acquired new knowledge in disciplines outside of clinical dentistry, especially in relation to assessment methods, data protection and management, approval processes, and statistics. The study has also allowed me to learn new skills, particularly regarding the use of statistical software, and organising and moderating focus group discussions. Furthermore, my academic writing has undoubtedly improved from compiling this thesis.

The study has not been without challenges. Being from a clinical background, I initially found much of the educational literature difficult to read. However, I have become more accustomed to the style and concepts discussed within educational papers and textbooks and now feel less overwhelmed by them. It also took time for me to understand some statistical theories, how they can be applied and how they can be interpreted. Other early frustrations arose from the introduction of the EU's GDPR, which caused initial uncertainty and hesitancy among data protection personnel and, as a result, led to delays in obtaining the approvals required for the study. The data sources were also a cause of frustration since their content (or lack of) prevented some lines of investigation from being pursued. Finally, the national restrictions which were imposed due to the COVID-19 pandemic slowed progression of the study. Immediate contact with my supervisors was reduced, additional time was needed to draw up contingency plans for the qualitative component of the study and amendments to data management approvals were required. That said, I was fortunate compared to other postgraduate research students conducting clinical and/or laboratory-based studies, which were more significantly impacted and delayed. Patience and tenacity were certainly required to overcome these difficulties.

In hindsight, I feel some aspects of the study could have been approached differently. Firstly, despite providing participants with information leaflets in advance, I think the purpose of the study could have been re-iterated during the focus groups. I had anticipated that some participants would expect to see individual patterns of poor student clinical performance as part of the results,

and this proved to be the case during the discussions. Additionally, the participants could have also been shown the results of the investigations into the reliability of longitudinal clinical assessment. These small changes may have resulted in the accumulation of valuable additional qualitative data for analysis.

Regardless, I believe that, overall, I have designed and conducted a study which provides an important early contribution towards research on longitudinal assessment in education pertaining to HCPs and is aligned with the priorities set by the SOHRC. The knowledge and skills I have acquired will now allow me (either as principal or supervisory researcher) to conduct subsequent dental education research projects more efficiently.

Completion of this study should now allow me to return to full-time specialty training in Restorative Dentistry. This will present a series of new challenges resulting from reduced exposure to specialist level clinical practice over the past three-and-a-half-years and the prolonged effects of the COVID-19 pandemic on the dental profession. However, I look forward to meeting these challenges with an improved self-confidence which has gradually increased over the course of this study.

References

ABDULGHANI, H. M., AHMAD, F., IRSHAD, M., KHALIL, M. S., AL-SHAIKH, G. K., SYED, S., ALDREES, A. A., ALROWAIS, N. & HAQUE, S. 2015. Faculty development programs improve the quality of Multiple Choice Questions items' writing. *Scientific Reports,* 5**,** 9556.

ABDULGHANI, H. M., IRSHAD, M., HAQUE, S., AHMAD, T., SATTAR, K. & KHALIL, M. S. 2017. Effectiveness of longitudinal faculty development programs on MCQs items writing skills: A follow-up study. *PLoS One,* 12**,** e0185895.

ABDUS, S., RABEYA, Y., SHEIKH, B. & ABU, M. 2020. Multiple Choice Questions in Medical Education: How to Construct High Quality Questions. *International journal of human and health sciences,* 4**,** 79-88.

ADAMSON, K. A. & PRION, S. 2013. Reliability: Measuring Internal Consistency Using Cronbach's α. *Clinical simulation in nursing,* 9**,** e179-e180.

ADEE 2010. Curriculum structure, content, learning and assessment in european undergraduate dental education. Assocation for Dental Education in Europe.

AJJAWI, R., BARTON, K. L., DENNIS, A. A. & REES, C. E. 2017. Developing a national dental education research strategy: priorities, barriers and enablers. *BMJ Open,* 7**,** e013129.

AKBARI, M. & MAHAVELATI SHAMSABADI, R. 2013. Direct Observation of Procedural Skills (DOPS) in Restorative Dentistry: Advantages and Disadvantages in Student's Point of View. *Iranian Journal of Medical Education,* 13**,** 212-220.

AL-OSAIL, A. M., AL-SHEIKH, M. H., AL-OSAIL, E. M., AL-GHAMDI, M. A., AL-HAWAS, A. M., AL-BAHUSSAIN, A. S. & AL-DAJANI, A. A. 2015. Is Cronbach's alpha sufficient for assessing the reliability of the OSCE for an internal medicine course? *BMC Research Notes,* 8**,** 582.

ALBANESE, M. A. & DAST, L. C. 2014. Problem-based learning. *In:* SWANWICK, T. (ed.) *Understanding Medical Education: Evidence, Theory and Practice.* 2nd ed.: John Wiley & Sons, Ltd.

ALBINO, J. E. N., YOUNG, S. K., NEUMANN, L. M., KRAMER, G. A., ANDRIEU, S. C., HENSON, L., HORN, B. & HENDRICSON, W. D. 2008. Assessing Dental Students' Competence: Best Practice Recommendations in the Performance Assessment Literature and Investigation of Current Practices in Predoctoral Dental Education. *Journal of Dental Education,* 72**,** 1405-1435.

ALFARIS, E., NAEEM, N., IRFAN, F., QURESHI, R., SAAD, H., AL SADHAN, R., ABDULGHANI, H. M. & VAN DER VLEUTEN, C. 2015. A One-Day Dental Faculty Workshop in Writing Multiple-Choice Questions: An Impact Evaluation. *Journal of Dental Education,* 79**,** 1305-13.

ALI, K., TREDWIN, C., KAY, E. & SLADE, A. 2016. Transition of new dental graduates into practice: a qualitative study. *European Journal of Dental Education,* 20**,** 65-72.

ALI, S. H., CARR, P. A. & RUIT, K. G. 2016. Validity and Reliability of Scores Obtained on Multiple-Choice Questions: Why Functioning Distractors Matter. *Journal of the Scholarship of Teaching and Learning,* 16**,** 1-14.

ALKHATIB, H. S., BRAZEAU, G., AKOUR, A. & ALMUHAISSEN, S. A. 2020. Evaluation of the effect of items' format and type on psychometric properties of sixth year pharmacy students clinical clerkship assessment items. *BMC Medical Education,* 20**,** 190.

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION, AMERICAN PSYCHOLOGICAL ASSOCIATION & NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION 2004. Standards for educational and psychological testing. Washington DC: American Eduational Research Association.

AMIEL, G. E., TANN, M., KRAUSZ, M. M., BITTERMAN, A. & COHEN, R. 1997. Increasing examiner involvement in an objective structured clinical examination by integrating a structured oral examination. *American Journal of Surgery,* 173**,** 546-549.

AMINI, A., SHIRZAD, F., MOHSENI, M., SADEGHPOUR, A. & ELMI, A. 2015. Designing Direct Observation of Procedural Skills (DOPS) test for selective skills of orthopedic residents and evaluating its effects from their points of view. *Research and Development in Medical Education,* 4**,** 1452-1457.

ANASTAKIS, D. J., COHEN, R. & REZNICK, R. K. 1991. The structured oral examination as a method for assessing surgical residents. *American Journal of Surgery,* 162**,** 67-70.

ANDERSON, L. W., KRATHWOHL, D. R., AIRASIAN, P. W., CRUIKSHANK, K. A., MAYER, R. E., PINTRICH, P. R., RATHS, J. & WITTROCK, M. C. 2001. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's Taxonomy of Educational Objectives (Complete edition),* New York, Longman.

ARCHBOLD HUFTY ALEGRÍA, D., BOSCARDIN, C., PONCELET, A., MAYFIELD, C. & WAMSLEY, M. 2014. Using tablets to support self-regulated learning in a longitudinal integrated clerkship. *Medical Education Online,* 19**,** 23638-23638.

ASADI, K., MIRBOLOOK, A. R., HAGHIGHI, M., SEDIGHINEJAD, A., NADERI NABI, B., ABEDI, S. & DEHSARA, F. 2012. Evaluation of Satisfaction Level of Orthopedic Interns from Direct Observation of procedural Skills Assessment (DOPS). *Research in Medical Education,* 4**,** 17-23.

BABBIE, E. R. 2010. *The Practice of Social Research.,* Belmont, CA, Wadsworth Cengage.

BAIRD, A. S. 2010. The new Extended Matching Question (EMQ) paper of the MFSRH Examination. *Journal of Family Planning and Reproductive Health Care,* 36**,** 171-173.

BAKER, E., AYRES, P., O'NEIL, H. F., CHLI, K., SAWYER, W., SYLVESTER, R. M. & CARROLL, B. 2008. KS3 English Test Marker Study in Australia: Final Report to the National Assessment Agency of England. Sherman Oaks, CA: University of Southern California.

BAMBER, D. 1975. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of mathematical psychology,* 12**,** 387-415.

BARRY, M., BRADSHAW, C. & NOONAN, M. 2013. Improving the content and face validity of OSCE assessment marking criteria on an undergraduate midwifery programme: a quality initiative. *Nurse Education Practice,* 13**,** 477-480.

BARTMAN, I., SMEE, S. & ROY, M. 2013. A method for identifying extreme OSCE examiners. *The clinical teacher,* 10**,** 27-31.

BARTON, J. R., CORBETT, S., VAN DER VLEUTEN, C. P., ENGLISH BOWEL CANCER SCREENING, P. & ENDOSCOPY, U. K. J. A. G. F. G. 2012. The validity and reliability of a Direct Observation of Procedural Skills assessment tool: assessing colonoscopic skills of senior endoscopists. *Gastrointestinal Endoscopy,* 75**,** 591-597.

BEANLAND, C., SCHNEIDER, Z., LOBIONDO-WOOD, G. & HABER, J. 1999. *Nursing research: methods, critical appraisal and utilisation,* Mosby, Sydney, Harcourt Brace & Company.

BECKMAN, T. J., GHOSH, A. K., COOK, D. A., ERWIN, P. J. & MANDREKAR, J. N. 2004. How reliable are assessments of clinical teaching?: A review of the published instruments. *Journal of General Internal Medicine,* 19**,** 971-977.

BELLAMY, N. 2015. Principles of clinical outcome assessment. *In:* HOCHBERG, M. C., SILMAN, A. J., SMOLEN, J. S., WEINBLATT, M. E. & WEISMAN, M. H. (eds.) *Rheumatology*. 6th ed.: Mosby.

BELLARA, A. P. C. 2018. Modified Angoff Method. *In:* FREY, B. B. (ed.) *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation*. . Thousand Oaks: SAGE Publications, Inc.

BERENDONK, C., STALMEIJER, R. E. & SCHUWIRTH, L. W. T. 2013. Expertise in performance assessment: assessors' perspectives. *Advances in Health Sciences Education: Theory and Practice,* 18**,** 559-571.

BERGER, R. 2015. Now I see it, now I don't: researcher's position and reflexivity in qualitative research. *Qualitative research: QR,* 15**,** 219-234.

BEULLENS, J., STRUYF, E. & VAN DAMME, B. 2005. Do extended matching multiple-choice questions measure clinical reasoning? *Medical Education,* 39**,** 410-417.

BEULLENS, J., STRUYF, E. & VAN DAMME, B. 2006. Diagnostic ability in relation to clinical seminars and extended-matching questions examinations. *Medical Education,* 40**,** 1173-1179.

BEULLENS, J., VAN DAMME, B., JASPAERT, H. & JANSSEN, P. J. 2002. Are extended-matching multiple-choice items appropriate for a final test in medical education? *Medical Teacher,* 24**,** 390-395.

BIESTA, G. 2010. Pragmatism and the philosophical foundations of mixed methods research. *In:* TASHAKKORI, A. & TEDDLIE, C. (eds.) *SAGE handbook of mixed methods in social and behavioral research*. 2nd ed. Thousand Oaks, CA: SAGE.

BIGGS, J. 1996. Enhancing teaching through constructive alignment. *Higher Education,* 32**,** 347-364.

BILIC-ZULLE, L., FRKOVIC, V., TURK, T., AZMAN, J. & PETROVECKI, M. 2005. Prevalence of plagiarism among medical students. *Croatian Medical Journal,* 46**,** 126-131.

BINDAL, N., GOODYEAR, H., BINDAL, T. & WALL, D. 2013. DOPS assessment: a study to evaluate the experience and opinions of trainees and assessors. *Medical Teacher,* 35**,** e1230-e1234.

BLOOM, B. S. 1984. Taxonomy of Educational Objectives. *In:* MCKAY, D. (ed.) *The Cognitive Domain*. 1st ed. New York: Company Inc.

BLOOM, B. S., ENGELHART, M. D., FURST, E. J., HILI, W. H. & KRATHWOHL, D. R. 1956. *Taxonomy of educational objectives: Handbook I: Cognitive domain,* New York, David McKay.

BLOXHAM, S., DEN-OUTER, B., HUDSON, J. & PRICE, M. 2016. Let's stop the pretence of consistent marking: exploring the multiple limitations of assessment criteria. *Assessment & Evaluation in Higher Education,* 41**,** 446-481.

BOGG, J., GIBBS, T. & BUNDRED, P. 2001. Training, job demands and mental health of pre-registration house officers. *Medical Education,* 35**,** 590-595.

BOK, H. G. J., JAARSMA, D. A. D. C., SPRUIJT, A., VAN BEUKELEN, P., VAN DER VLEUTEN, C. P. M. & TEUNISSEN, P. W. 2016. Feedback-giving behaviour

in performance evaluations during clinical clerkships. *Medical Teacher,* 38, 88-95.
BOLLEN, K. A. & CURRAN, P. J. 2006. *Latent curve models: A structural equation approach,* Hoboken, NJ, Wiley.
BOULD, M. D., CRABTEE, N. A. & NAIK, V. N. 2009. Assessment of procedural skills in anesthesia. *British Journal Anaesthesia,* 103, 472–483.
BOURSICOT, K., ETHERIDGE, L., SETNA, Z., STURROCK, A., KER, J., SMEE, S. & SAMBANDAM, E. 2011. Performance in assessment: consensus statement and recommendations from the Ottawa conference. *Medical Teacher,* 33, 370-383.
BOURSICOT, K. A., ROBERTS, T. E. & PELL, G. 2007. Using borderline methods to compare passing standards for OSCEs at graduation across three medical schools. *Medical Education,* 41, 1024-1031.
BOURSICOT, K. A. M., ROBERTS, T. E. & BURDICK, W. P. 2014. Structured Assessments of Clinical Competence. *In:* SWANWICK, T. (ed.) *Understanding Medical Education: Evidence, Theory and Practice*. John Wiley & Sons, Ltd.
BOURSICOT, K. A. M., ROBERTS, T. E. & BURDICK, W. P. 2019. Structured Assessments of Clinical Competence. *In:* SWANWICK, T., FORREST, K. & O'BRIEN, B. C. (eds.) *Understanding Medical Education: Evidence, Theory, and Practice*. Hoboken, NJ: Wiley-Blackwell.
BOWMAN, M. 2017. The transition to self-regulated learning for first-year dental students: threshold concepts. *European Journal of Dental Education,* 21, 142-150.
BRADLEY, D. & HUSEMAN, S. 2003. Validating competency at the bedside. *Journal for Nurses in Staff Development,* 19, 165-173.
BRAME, C. 2013. *Writing good multiple choice test questions.* [Online]. Available: https://cft.vanderbilt.edu/guides-sub-pages/writing-good-multiple-choice-test-questions/ [Accessed 16th December 2020].
BRANNICK, M. T., EROL-KORKMAZ, H. T. & PREWETT, M. 2011. A systematic review of the reliability of objective structured clinical examination scores. *Medical Education,* 45, 1181-1189.
BRAUN, V. & CLARKE, V. 2006. Using thematic analysis in psychology. *Qualitative Research in Pyschology,* 3, 77-101.
BRENNAN, N., CORRIGAN, O., ALLARD, J., ARCHER, J., BARNES, R., BLEAKLEY, A., COLLETT, T. & DE BERE, S. R. 2010. The transition from medical student to junior doctor: today's experiences of Tomorrow's Doctors. *Medical Education,* 44, 449-458.
BRIDGE, P. D., MUSIAL, J., FRANK, R., ROE, T. & SAWILOWSKY, S. 2003. Measurement practices: methods for developing content-valid student examinations. *Medical Teacher,* 25, 414-421.
BROADBENT, J. M., THOMSON, W. M. & POULTON, R. 2008. Trajectory patterns of dental caries experience in the permanent dentition to the fourth decade of life. *Journal of Dental Research,* 87, 69-72.
BROWN, C., ROSS, S., CLELAND, J. & WALSH, K. 2015. Money makes the (medical assessment) world go round: The cost of components of a summative final year Objective Structured Clinical Examination (OSCE). *Medical Teacher,* 37, 653-659.
BROWN, G., MANOGUE, M. & MARTIN, M. 1999. The validity and reliability of an OSCE in dentistry. *European Journal of Dental Education,* 3, 117-125.
BROWN, J. D. 2000. What is construct validity? *Sjiken: JALT Testing and Evaluation SIG Newsletter,* 4, 8-12.

BRYMAN, A. 2016. *Social research methods,* Oxford, Oxford University Press.
BUCHWITZ, B. J., BEYER, C. H., PETERSON, J. E., PITRE, E., LALIC, N., SAMPSON, P. D. & WAKIMOTO, B. T. 2012. Facilitating long-term changes in student approaches to learning science. *CBE Life Sciences Education,* 11, 273-282.
BUSH, H. M., SCHREIBER, R. S. & OLIVER, S. J. 2013. Failing to fail: clinicians' experience of assessing underperforming dental students. *European Journal of Dental Education,* 17, 198-207.
BYRNE, J. & HUMBLE, A. 2006. An Introduction to Mixed Method Research.
CADORIN, L., BRESSAN, V. & PALESE, A. 2017. Instruments evaluating the self-directed learning abilities among nursing students and nurses: a systematic review of psychometric properties. *BMC Medical Education,* 17, 229-229.
CAPAN MELSER, M., STEINER-HOFBAUER, V., LILAJ, B., AGIS, H., KNAUS, A. & HOLZINGER, A. 2020. Knowledge, application and how about competence? Qualitative assessment of multiple-choice questions for dental students. *Medical Education Online,* 25, 1714199.
CARPENTER, J. L. 1995. Cost analysis of objective structured clinical examinations. *Academic Medicine,* 70, 828-833.
CARR, J., HEGGARTY, H., CARR, M., FULWOOD, D., GOODWIN, C., WALKER, W. W. & WHITTINGHAM, K. 2010. Reflect for success: recommendations for mentors managing failing students. *British journal of community nursing,* 15, 594.
CARRACCIO, C. L. & ENGLANDER, R. 2013. From Flexner to Competencies: Reflections on a Decade and the Journey Ahead. *Academic Medicine,* 88, 1067-1073.
CARSON, D., GILMORE, A., PERRY, C. & GRONHAUG, K. 2001. *Qualitative Marketing Research,* London, SAGE.
CASE, S. M. & SWANSON, D. B. 1993. Extended-matching items: a practical alternative to free-response questions. *Teaching and Learning in Medicine,* 5, 107-115.
CASE, S. M. & SWANSON, D. B. 2001. *Constructing Written Test Questions for the Basic and Clinical Sciences* [Online]. Philadelphia: National Board of Medical Examiners. Available: http://www.nbme.org/PDF/2001iwg.pdf [Accessed].
CHAMBERS, D. W. 1993. Toward a competency-based curriculum. *Journal of Dental Education,* 57, 790-793.
CHAMBERS, D. W. 1998. Competency-based dental education in context. *European Journal of Dental Education,* 2, 8-13.
CHESBRO, S. B., JENSEN, G. M. & BOISSONNAULT, W. G. 2018. Entrustable professional activities as a framework for continued professional competence: is now the time? *Physical Therapy,* 98, 3-7.
CHIRCULESCU, A. R. M., CHIRCULESCU, M. & MORRIS, J. F. 2007. Anatomical teaching for medical students from the perspective of European Union enlargement. *European Journal of Anatomy,* 11, 63-65.
CHO, K. K., MARJADI, B., LANGENDYK, V. & HU, W. 2017. The self-regulated learning of medical students in the clinical environment – a scoping review. *BMC Medical Education,* 17, 112.
CHRISTERSSON, C., BENGMARK, D., BENGTSSON, H., LINDH, C., ROHLIN, M., MALMÖ, U. & FACULTY OF, O. 2015. A predictive model for alternative admission to dental education. *European Journal of Dental Education,* 19, 251-258.

CIZEK, G. J. 2001. *Setting Performance Standards: Concepts, Methods, and Perspectives,* Mahwah, NJ, Lawrence Erlbaum Associates.

CLELAND, J., ARNOLD, R. & CHESSER, A. 2005. Failing finals is often a surprise for the student but not the teacher: identifying difficulties and supporting students with academic difficulties. *Medical Teacher,* 27**,** 504-508.

CLELAND, J., DOWELL, J., MCLACHLAN, J., NICHOLSON, S. & PATTERSON, F. 2012. Identifying best practice in the selection of medical students (literature review and interview survey). Available: https://www.gmc-uk.org/-/media/about/identifyingbestpracticeintheselectionofmedicalstudentspdf51119804.pdf?la=en&hash=7BF8F94D402EC8230728221C2598732D69D81851.

CLELAND, J. A., KNIGHT, L. V., REES, C. E., TRACEY, S. & BOND, C. M. 2008. Is it me or is it them? Factors that influence the passing of underperforming students. *Medical Education,* 42**,** 800-809.

COBB, K. A., BROWN, G., JAARSMA, D. A. & HAMMOND, R. A. 2013. The educational impact of assessment: a comparison of DOPS and MCQs. *Medical Teacher,* 35**,** e1598-e1607.

CODERRE, S. P., HARASYM, P., MANDIN, H. & FICK, G. 2004. The impact of two multiple-choice question formats on the problem-solving strategies used by novices and experts. *BMC Medical Education,* 4**,** 23.

COE, J. M., BRICKHOUSE, T. H., BHATTI, B. A. & BEST, A. M. 2018. Impact of Community-Based Clinical Training on Dental Students' Confidence in Treating Pediatric Patients. *Journal of Dental Education,* 82**,** 5-11.

COETZEE, K. & MONTEIRO, S. 2019. DRIFT happens, sometimes: Examining time based rater variance in a high-stakes OSCE. *Medical Teacher,* 41**,** 819-823.

COHEN, L., MANION, L. & MORRISON, K. 2017. Coding and content analysis. *Research Methods on Education.* London and New York: Routledge.

COHEN, L., MANION, L. & MORRISON, K. 2018. *Research methods in education,* New York, New Yok; London, England;, Routledge.

COHEN, R., REZNICK, R. K., TAYLOR, B. R., PROVAN, J. & ROTHMAN, A. 1990. Reliability and validity of the objective structured clinical examination in assessing surgical residents. *American Journal of Surgery,* 160**,** 302-305.

COHEN, S. N., FARRANT, P. B. & TAIBJEE, S. M. 2009. Assessing the assessments: U.K. dermatology trainees' views of the workplace assessment tools. *British Journal of Dermatology,* 161**,** 34-39.

COLLINS, J. 2006. Education techniques for lifelong learning: writing multiple-choice questions for continuing medical education activities and self-assessment modules. *Radiographics,* 26**,** 543-551.

COLTON, T. & PETERSON, O. L. 1967. An assay of medical students' abilities by oral examination. *Journal of Medical Education,* 42**,** 1005-1014.

CONSIDINE, J., BOTTI, M. & THOMAS, S. 2005. Design, format, validity and reliability of multiple choice questions for use in nursing research and education. *Collegian,* 12**,** 19-24.

COOK, D. A., DUPRAS, D. M., BECKMAN, T. J., THOMAS, K. G. & PANKRATZ, V. S. 2009. Effect of rater training on reliability and accuracy of mini-CEX scores: a randomized, controlled trial. *Journal of General Internal Medicine,* 24**,** 74-79.

COUGHLIN, P. A. & FEATHERSTONE, C. R. 2017. How to Write a High Quality Multiple Choice Question (MCQ): A Guide for Clinicians. *European Journal of Vascular and Endovascular Surgery,* 54**,** 654-658.

COX, K. R. 1982. How to improve oral examinations. *In:* COX, K. R. & EVANS, C. E. (eds.) *The Medical Teacher*. Edinburgh: Churchill Livingstone.
CRESWELL, J. 2003. *Research Design: qualitative, quantitative and mixed methods approaches,* USA, SAGE.
CRESWELL, J. W. & CRESWELL, J. D. 2018. *Research design: qualitative, quantitative, and mixed methods approaches.,* Los Angeles, SAGE.
CRESWELL, J. W. & PLANO CLARK, V. L. 2011. *Designing and Conducting Mixed Methods Research*
London, SAGE Publications Ltd.
CRONBACH, L. J. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika,* 16**,** 297-334.
CRONBACH, L. J. 1971. Test validation. *In:* THORNDIKE, R. L. (ed.) *Educational measurement*. 2nd ed. Washington, DC: American Council on Education.
CRONBACH, L. J. 1988. Five perspectives on validity argument. *In:* WAINER, H. & BRAUN, H. (eds.) *Test validity*. Hillsdale, NJ: Lawrence Erlbaum.
CROSSLEY, J., JOHNSON, G., BOOTH, J. & WADE, W. 2011. Good questions, good answers: construct alignment improves the performance of workplace-based assessment scales. *Medical Education,* 45**,** 560-569.
CROSSLEY, J. G. M., GROVES, J., CROKE, D. & BRENNAN, P. A. 2019. Examiner training: A study of examiners making sense of norm-referenced feedback. *Medical Teacher,* 41**,** 787-794.
CROTTY, M. 1998. *The foundations of social research,* Sydney, Allen and Unwin.
CUNNINGTON, J. P., NEVILLE, A. J. & NORMAN, G. R. 1996. The risks of thoroughness: Reliability and validity of global ratings and checklists in an OSCE. *Advances in Health Sciences Education Theory and Practice,* 1**,** 227-233.
DABHADKAR, S., WAGH, G., PANCHANADIKAR, T., MEHENDALE, S. & SAOJI, V. 2014. To evaluate Direct Observation of Procedural Skills in OBGY. *National Journal of Integrated Research in Medicine,* 5**,** 92-97.
DAELMANS, H. E., SCHERPBIER, A. J., VAN DER VLEUTEN, C. P. & DONKER, A. J. 2001. Reliability of clinical oral examinations re-examined. *Medical Teacher,* 23**,** 422-424.
DAMJANOV, I., FENDERSON, B. A., VELOSKI, J. J. & RUBIN, E. 1995. Testing of medical students with open-ended, uncued questions. *Human Pathology,* 26**,** 362-365.
DANIELS, V. J. & HARLEY, D. 2017. The effect on reliability and sensitivity to level of training of combining analytic and holistic rating scales for assessing communication skills in an internal medicine resident OSCE. *Patient Education and Counseling,* 100**,** 1382-1386.
DANNEFER, E. F., HENSON, L. C., BIERER, S. B., GRADY-WELIKY, T. A., MELDRUM, S., NOFZIGER, A. C., BARCLAY, C. & EPSTEIN, R. M. 2005. Peer assessment of professional competence. *Medical Education,* 39**,** 713-722.
DAVE, R. H. 1970. Psychomotor levels. *In:* ARMSTRONG, R. J. (ed.) *Developing and writing educational objectives.* Tucson AZ: Educational Innovators Press.
DAVEY, J., BRYANT, S. T. & DUMMER, P. M. H. 2015. The confidence of undergraduate dental students when performing root canal treatment and their perception of the quality of endodontic education. *European Journal of Dental Education,* 19**,** 229-234.
DAVIS, M. H. 2003. OSCE: the Dundee experience. *Medical Teacher,* 25**,** 255-261.
DAVIS, M. H. & KARUNATHILAKE, I. 2005. The place of the oral examination in today's assessment systems. *Medical Teacher,* 27**,** 294-297.

DAWSON, L. J., FOX, K., JELLICOE, M., ADDERTON, E., BISSELL, V. & YOUNGSON, C. C. 2021. Is the number of procedures completed a valid indicator of final year student competency in operative dentistry? *British dental journal,* 230, 663-670.

DAWSON, L. J., MASON, B. G., BISSELL, V. & YOUNGSON, C. 2017. Calling for a re-evaluation of the data required to credibly demonstrate a dental student is safe and ready to practice. *European Journal of Dental Education,* 21, 130–135.

DAY, S. C., NORCINI, J. J., DISERENS, D., CEBUL, R. D., SCHWARTZ, J. S., BECK, L. H., WEBSTER, G. D., SCHNABEL, T. G. & ELSTEIN, A. 1990. The validity of an essay test of clinical judgment. *Academic Medicine,* 65, S39-S40.

DE CHAMPLAIN, A. F. 2010. A primer on classical test theory and item response theory for assessments in medical education. *Medical Education,* 44, 109-117.

DEBREW, J. K. & LEWALLEN, L. P. 2014. To pass or to fail? Understanding the factors considered by faculty in the clinical evaluation of nursing students. *Nurse Education Today,* 34, 631-636.

DEELEY, S. J. & BOVILL, C. 2017. Staff student partnership in assessment: enhancing assessment literacy through democratic practices. *Assessment & Evaluation in Higher Education,* 42, 463-477.

DEKKER, M. C., FERDINAND, R. F., VAN LANG, N. D., BONGERS, I. L., VAN DER ENDE, J. & VERHULST, F. C. 2007. Developmental trajectories of depressive symptoms from early childhood to late adolescence: gender differences and adult outcome. *Journal of Child Psychology and Psychiatry,* 48, 657-666.

DELFINO, A. E., CHANDRATILAKE, M., ALTERMATT, F. R. & ECHEVARRIA, G. 2013. Validation and piloting of direct observation of practical skills tool to assess intubation in the Chilean context. *Medical Teacher,* 35, 231-236.

DELLINGES, M. A. & CURTIS, D. A. 2017. Will a Short Training Session Improve Multiple-Choice Item-Writing Quality by Dental School Faculty? A Pilot Study. *Journal of Dental Education,* 81, 948-955.

DENNEHY, P. C., SUSARLA, S. M. & KARIMBUX, N. Y. 2008. Relationship between dental students' performance on standardized multiple-choice examinations and OSCEs. *Journal of Dental Education,* 72, 585–592.

DENTAL SCHOOLS COUNCIL. 2021. *Entry requirements* [Online]. Available: https://www.dentalschoolscouncil.ac.uk/making-an-application/entry-requirements/ [Accessed April 2021].

DEPOY, E. & GITLIN, L. N. 2016. Collecting Data Through Measurement in Experimental-Type Research. *In:* DEPOY, E. & GITLIN, L. N. (eds.) *Introduction to Research.* 5th ed. St. Louis, Missouri: Elsevier.

DES MARCHAIS, J. E. & JEAN, P. 1993. Effects of examiner training on open-ended, higher taxonomic level questioning in oral certification examinations. *Teaching and Learning in Medicine,* 3, 24–28.

DESY, J., CODERRE, S., DAVIS, M., CUSANO, R. & MCLAUGHLIN, K. 2019. How can we reduce bias during an academic assessment reappraisal? *Medical Teacher,* 41, 1315-1318.

DEWEY, J. 1941. Propositions, Warranted Assertibility, and Truth. *The Journal of Philosophy,* 38, 169-186.

DHOLE, A. 2017. Assessment of Post Graduates (PG) for Extra Oral Radiograph Techniques by DOPS (Direct Observation of Procedural Skills) Method. *Journal of Research & Method in Education,* 7, 7-9.

DICKIE, J. 2017. The relationship between clinical longitudinal performance and measures of competence for undergraduate dental students. *In:* GLASGOW, U. O. (ed.) Master of Education dissertation ed.

DIGNATH, C. & BÜTTNER, G. 2008. Components of fostering self-regulated learning among students. A meta-analysis on intervention studies at primary and secondary school level. *Metacognition and learning,* 3**,** 231-264.

DING, C. S. & HERSHBERGER, S. L. 2002. Assessing Content Validity and Content Equivalence Using Structural Equation Modeling. *Structural equation modeling,* 9**,** 283-297.

DOWNING, S. M. 2002. Assessment of knowledge with written test formats. *In:* NORMAN, G., VAN DER VLEUTEN, C. & NEWBLE, D. (eds.) *International Handbook of Research in Medical Education.* Dordrecht: Kluwer.

DOWNING, S. M. 2003. Validity: on the meaningful interpretation of assessment data. *Medical Education,* 37**,** 830-837.

DOWNING, S. M. & HALADYNA, T. M. 2004. Validity threats: overcoming interference with proposed interpretations of assessment data. *Medical Education,* 38**,** 327-333.

DRIESSEN, E. W., VAN TARTWIJK, J., GOVAERTS, M., TEUNISSEN, P. & VAN DER VLEUTEN, C. P. 2012. The use of programmatic assessment in the clinical workplace: a Maastricht case report. *Medical Teacher,* 34**,** 226-231.

DUDEK, N. L., MARKS, M. B. & REGEHR, G. 2005. Failure to fail: the perspectives of clinical supervisors. *Academic Medicine,* 80**,** S84-S87.

DUFFY, K. 2006. *Weighing the balance: a grounded theory study of the factors that influence the decisions regarding the assessment of students' competence in practice.* Caledonian University.

DUFFY, K. & HARDICRE, J. 2007. Supporting failing students in practice. 1: Assessment. *Nursing Times,* 103**,** 28–29.

DUIJN, C., TEN CATE, O., KREMER, W. D. J. & BOK, H. G. J. 2019. The Development of Entrustable Professional Activities for Competency-Based Veterinary Education in Farm Animal Health. *Journal of Veterinary Medical Education,* 46**,** 218-224.

DUTHIE, S., HODGES, P., RAMSAY, I. & REID, W. 2006. EMQs: a new component of the MRCOG Part 2 exam. *The Obstetrician & Gynaecologist,* 8**,** 181-185.

EBERHARD, L., HASSEL, A., BAUMER, A., BECKER, F., BECK-MUBOTTER, J., BOMICKE, W., CORCODEL, N., COSGAREA, R., EIFFLER, C., GIANNAKOPOULOS, N. N., KRAUS, T., MAHABADI, J., RUES, S., SCHMITTER, M., WOLFF, D. & WEGE, K. C. 2011. Analysis of quality and feasibility of an objective structured clinical examination (OSCE) in preclinical dental education. *European Journal of Dental Education,* 15**,** 172-178.

EDWARDS, B. D. & ARTHUR, W., JR. 2007. An examination of factors contributing to a reduction in subgroup differences on a constructed-response paper-and-pencil test of scholastic achievement. *Journal of Applied Psychology,* 92**,** 794-801.

EGGLESTON, E. P., LAUB, J. H. & SAMPSON, R. J. 2004. Methodological Sensitivities to Latent Class Analysis of Long-Term Criminal Trajectories. *Journal of Quantitative Criminology,* 20**,** 1-26.

EPSTEIN, R. M. 2007. Assessment in medical education. *New England Journal of Medicine,* 356**,** 387-396.

ERFANI KHANGHAHI, M. & EBADI FARD AZAR, F. 2018. Direct observation of procedural skills (DOPS) evaluation method: Systematic review of evidence. *Medical Journal of The Islamic Republic of Iran,* 32**,** 45.

ESCUDIER, M. P., NEWTON, T. J., COX, M. J., REYNOLDS, P. A. & ODELL, E. W. 2011. University students' attainment and perceptions of computer delivered assessment; a comparison between computer-based and traditional tests in a 'high-stakes' examination. *Journal of Computer Assisted Learning,* 27**,** 440-447.

EUROPEAN BOARD OF MEDICAL ASSESSORS 2017. Guidelines for writing multiple-choice questions. Maastricht, The Netherlands: EMBA.

EUROPEAN PARLIAMENT AND COUNCIL OF EUROPEAN UNION 2016. EU General Data Protection Regulation (GDPR): Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). European Parliament and Council of European Union.

EVA, K. W., REITER, H. I., ROSENFELD, J., TRINH, K., WOOD, T. J. & NORMAN, G. R. 2012. Association Between a Medical School Admission Process Using the Multiple Mini-interview and National Licensing Examination Scores. *Journal of the American Medical Association,* 308**,** 2233.

EVA, K. W., REITER, H. I., TRINH, K., WASI, P., ROSENFELD, J. & NORMAN, G. R. 2009. Predictive validity of the multiple mini-interview for selecting medical trainees. *Medical Education,* 43**,** 767-775.

EVA, K. W., ROSENFELD, J., REITER, H. I. & NORMAN, G. R. 2004. An admissions OSCE: the multiple mini-interview. *Medical Education,* 38**,** 314-326.

EVANS, M., HASTINGS, N. & PEACOCK, B. 2000. "Bernoulli Distribution". *Statistical Distributions.* 3rd ed. New York: Wiley.

EXAMSOFT 2020. Turnitin Acquires ExamSoft, A Leading Assessment Platform. Examsoft.

FELD, L. S. & BRENNAN, R. L. 1989. *Educational Measurement,* New York, American Council on Education and Macmillan.

FELETTI, G. I. & SMITH, E. K. 1986. Modified essay questions: are they worth the effort? *Medical Education,* 20**,** 126–132.

FENDERSON, B. A., DAMJANOV, I., ROBESON, M. R., VELOSKI, J. J. & RUBIN, E. 1997. The virtues of extended matching and uncued tests as alternatives to multiple choice questions. *Human Pathology,* 28**,** 526-532.

FIELD, J. C., WALMSLEY, A. D., PAGANELLI, C., MCLOUGHLIN, J., SZEP, S., KAVADELLA, A., MANZANARES CESPEDES, M. C., DAVIES, J. R., DELAP, E., LEVY, G., GALLAGHER, J., ROGER-LEROI, V. & COWPE, J. G. 2017. The Graduating European Dentist: Contemporaneous Methods of Teaching, Learning and Assessment in Dental Undergraduate Education. *European Journal of Dental Education,* 21 Suppl 1**,** 28-35.

FINE, P., LEUNG, A., BENTALL, C. & LOUCA, C. 2019. The impact of confidence on clinical dental practice. *European Journal of Dental Education,* 23**,** 159-167.

FINK, A. 2010. Survey Research Methods. *In:* PETERSON, P., BAKER, E. & MCGAW, B. (eds.) *International Encyclopedia of Education.* 3rd ed.: Elsevier Science.

FISHLEDER, A. J., HENSON, L. C. & HULL, A. L. 2007. Cleveland Clinic Lerner College of Medicine: an innovative approach to medical education and the training of physician investigators. *Academic Medicine,* 82**,** 390-396.

FOLEY, J. I. & HIJAZI, K. 2013. The admissions process in a graduate-entry dental school: can we predict academic performance? *British dental journal,* 214**,** E4-E4.

FOSTER, J. T., ABRAHAMSON, S., LASS, S., GIRARD, R. & GARRIS, R. 1969. Analysis of an oral examination used in specialty board certification. *Journal of Medical Education,* 44**,** 951-954.

FOWELL, S. L. & BLIGH, J. G. 1998. Recent developments in assessing medical students. *Postgraduate Medical Journal,* 74**,** 18-24.

FRANKFURT, S., FRAZIER, P., SYED, M. & JUNG, K. 2016. Using Group-Based Trajectory and Growth Mixture Modeling to Identify Classes of Change Trajectories. *The Counseling Psychologist,* 44**,** 622-660.

GAUTHIER, G., ST-ONGE, C. & TAVARES, W. 2016. Rater cognition: review and integration of research findings. *Medical Education,* 50**,** 511-522.

GCC 2017. Education Standards. London: General Chiropractic Council.

GDC 2015a. Preparing for Practice (2015 revised edition). London: General Dental Council.

GDC 2015b. Standards for Education: Standards and requirements for providers. London: General Dental Council.

GDC 2019a. Quality assurance guidance for education providers. London: General Dental Council.

GDC 2019b. Shaping the direction of lifelong learning for dental professionals. London: General Dental Council.

GERHARD-SZEP, S., GÜNTSCH, A., POSPIECH, P., SÖHNEL, A., SCHEUTZEL, P., WASSMANN, T. & ZAHN, T. 2016. Assessment formats in dental medicine: An overview. *GMS Journal for Medical Education,* 33**,** Doc65.

GERROW, J. D., MURPHY, H. J., BOYD, M. A. & SCOTT, D. A. 2003. Concurrent Validity of Written and OSCE Components of the Canadian Dental Certification Examinations. *Journal of Dental Education,* 67**,** 896-901.

GHAZIVAKILI, Z., NOROUZI NIA, R., PANAHI, F., KARIMI, M., GHOLSORKHI, H. & AHMADI, Z. 2014. The role of critical thinking skills and learning styles of university students in their academic performance. *Journal of advances in medical education & professionalism,* 2**,** 95-102.

GIBBS, H., HABESHAW, S. & HABESHAW, T. 1988. *Interesting Ways to Teach: 53 Interesting Ways to Assess your Students,* Bristol, Technical and Educational Services.

GILMOUR, A. S. M., WELPLY, A., COWPE, J. G., BULLOCK, A. D. & JONES, R. J. 2016. The undergraduate preparation of dentists: Confidence levels of final year dental students at the School of Dentistry in Cardiff. *British dental journal,* 221**,** 349-354.

GINGERICH, A., REGEHR, G. & EVA, K. W. 2011. Rater-based assessments as social judgments: rethinking the etiology of rater errors. *Academic Medicine,* 86**,** S1-S7.

GIRARD, L. C., TREMBLAY, R. E., NAGIN, D. & COTE, S. M. 2019. Development of Aggression Subtypes from Childhood to Adolescence: a Group-Based Multi-Trajectory Modelling Perspective. *Journal of Abnormal Child Psychology,* 47**,** 825-838.

GMC 2014. Good medical practice. London: General Medical Council.

GMC 2018. Outcomes for graduates. London: General Medical Council.

GOC 2016. Core Competencies for Optometrists. London: General Optical Council.

GOOD, J. P., RAMOS, D. & D'AMORE, D. C. 2013. Learning style preferences and academic success of preclinical allied health students. *Journal of Allied Health,* 42**,** e81.

GOPEE, N. 2008. Assessing student nurses' clinical skills: the ethical competence of mentors. *International Journal of Therapy and Rehabilitation,* 15, 401-407.

GORMLEY, G. 2011. Summative OSCEs in undergraduate medical education. *Ulster medical journal,* 80, 127-132.

GOSC 2015. Guidance for Osteopathic Pre-registration Education. London: General Osteopathic Council.

GOVAERTS, M. J., VAN DER VLEUTEN, C. P. & SCHUWIRTH, L. W. 2002. Optimising the reproducibility of a performance-based assessment test in midwifery education. *Advances in Health Sciences Education Theory and Practice,* 7, 133-145.

GRAHAM, R., ZUBIAURRE, B. L. A. & ANDERSON, O. R. 2013. Reliability and Predictive Validity of a Comprehensive Preclinical OSCE in Dental Education. *Journal of Dental Education,* 77, 161-167.

GRAINGER, R., DAI, W., OSBORNE, E. & KENWRIGHT, D. 2018. Medical students create multiple-choice questions for learning in pathology education: a pilot study. *BMC Medical Education,* 18, 201.

GRANT, D. L. 1957. Studies in the reliability of the short-answer essay examination. *Journal of Educational Research,* 51, 109.

GRAVETTER, F. J. & WALLNAU, L. B. 2000. *Statistics for the behavioral sciences,* Belmont, CA, Wadsworth – Thomson Learning.

GRAY, S. A. & DEEM, L. P. 2002. Predicting Student Performance in Preclinical Technique Courses Using the Theory of Ability Determinants of Skilled Performance. *Journal of Dental Education,* 66, 721-727.

GREATER BRIGHTON METROPOLITAN COLLEGE. 2021. *Access to Higher Education - Diploma in Medicine and Dentistry - Level 3* [Online]. Brighton: Greater Brighton Metropolitan College. Available: https://www.gbmc.ac.uk/access-to-higher-education-diploma-medicine-level-3 [Accessed June 2021].

GREATRIX, R., NICHOLSON, S. & ANDERSON, S. 2021. Does the UKCAT predict performance in medical and dental school? A systematic review. *BMJ Open,* 11, e040128-e040128.

GREEN, J. & RASMUSSEN, K. 2018. Becoming a dentist: faculty perceptions of student experiences with threshold concepts in a Canadian dental program. *Canadian medical education journal,* 9, e102-e110.

GREENE, J. 2008. Is mixed methods social inquiry a distinctive methodology? *Journal of Mixed Methods Research,* 2, 7-22.

GREENE, J. C., VALERIE, J. & CARACELLI, G. W. F. 1989. Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis,* 11, 255-274.

GRIEVESON, B. 2002. Assessment in dental VT. can we do better? *British dental journal,* Suppl, 19-22.

GRUPPEN, L., GRUM, C., FINCHER, R., PARENTI, C., CLEARY, L., SWANEY, J., CASE, S., SWANSON, D. & WOOLLISCROFT, J. 1994. Multi-site reliability of a diagnostic pattern-recognition knowledge-assessment instrument. *Academic Medicine,* S65-S67.

GUNZLER, D., TANG, W., LU, N., WU, P. & TU, X. M. 2014. A class of distribution-free models for longitudinal mediation analysis. *Psychometrika,* 79, 543-568.

HALADYNA, T. M. 1999. *Developing and validating multiple-choice test items,* New Jersey, Lawrence Erlbaum.

HAMDY, H., PRASAD, K., ANDERSON, M., SCHERPBIER, A., WILLIAMS, R., ZWIERSTRA, R. & CUDDIHY, H. 2006. BEME systematic review: Predictive values of measurements obtained in medical schools and future performance in medical practice. *Medical Teacher,* 28**,** 103-116.

HAMDY, H., PRASAD, K., WILLIAMS, R. & SALIH, F. A. 2003. Reliability and validity of the direct observation clinical encounter examination (DOCEE). *Medical Education,* 37**,** 205-212.

HAMILTON, K. E., COATES, V., KELLY, B., BOORE, J. R., CUNDELL, J. H., GRACEY, J., MCFETRIDGE, B., MCGONIGLE, M. & SINCLAIR, M. 2007. Performance assessment in health care providers: a critical review of evidence and current practice. *Journal of Nursing Management,* 15**,** 773-791.

HANLEY, J. A. & MCNEIL, B. J. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology,* 143**,** 29.

HARDEN, R. M. 2016. Revisiting 'Assessment of clinical competence using an objective structured clinical examination (OSCE)'. *Medical Education,* 50**,** 376-379.

HARDEN, R. M. & GLEESON, F. A. 1979. Assessment of clinical competence using an objective structured clinical examination (OSCE). *Medical Education,* 13**,** 41-54.

HARDEN, R. M., STEVENSON, M., DOWNIE, W. W. & WILSON, G. M. 1975. Assessment of clinical competence using objective structured examination. *British Medical Journal,* 1 447–451.

HARRIS, P., BHANJI, F., TOPPS, M., ROSS, S., LIEBERMAN, S., FRANK, J. R., SNELL, L., SHERBINO, J. & ICBME COLLABORATORS 2017. Evolving concepts of assessment in a competency-based world. *Medical Teacher,* 39**,** 603-608.

HASSEL, S. & RIDOUT, N. 2017. An Investigation of First-Year Students' and Lecturers' Expectations of University Education. *Frontiers in psychology,* 8**,** 2218-2218.

HATTAR, S., ALHADIDI, A., ALTARAWNEH, S., HAMDAN, A. A. S., SHAINI, F. J. & WAHAB, F. K. 2021. Dental students' experience and perceived confidence level in different restorative procedures. *European Journal of Dental Education,* 25**,** 207-214.

HATTIE, J. & TIMPERLEY, H. 2007. The power of feedback. *Review of Educational Research,* 77**,** 81-112.

HEA 2012. A Marked Improvement. York: The Higher Education Academy.

HECKER, K. & VIOLATO, C. 2009. Validity, Reliability, and Defensibility of Assessments in Veterinary Education. *Journal of Veterinary Medical Education,* 36**,** 271-275.

HECKLER, N. C., RICE, M. & HOBSON, B. C. 2013. Turnitin systems: a deterrent to plagiarism in college classrooms. *Journal of Research on Technology in Education,* 45**,** 229–248.

HEINICKE, M. R., ZUCKERMAN, C. K. & CRAVALHO, D. A. 2017. An Evaluation of Readiness Assessment Tests in a College Classroom: Exam Performance, Attendance, and Participation. *Behavior analysis (Washington, D.C.),* 17**,** 129-141.

HEJRI, S. M., JALILI, M., MUIJTJENS, A. M. M. & VAN DER VLEUTEN, C. P. M. 2013. Assessing the reliability of the borderline regression method as a standard setting procedure for objective structured clinical examination. *Journal of research in medical sciences,* 18**,** 887-891.

HENGAMEH, H., AFSANEH, R., MORTEZA, K., HOSEIN, M., MARJAN, S. M. & ABBAS, E. 2015. The Effect of Applying Direct Observation of Procedural Skills (DOPS) on Nursing Students' Clinical Skills: A Randomized Clinical Trial. *Global Journal of Health Science,* 7**,** 17-21.

HERIOT-WATT UNIVERSITY 2019. Student Learning Experience: Threshold Criteria. Edinburgh: Heriot-Watt University,.

HIFT, R. J. 2014. Should essays and other "open-ended"-type questions retain a place in written summative assessment in clinical medicine? *BMC Medical Education,* 14**,** 249.

HOANG, N. S. & LAU, J. N. 2018. A Call for Mixed Methods in Competency-Based Medical Education: How We Can Prevent the Overfitting of Curriculum and Assessment. *Academic Medicine,* 93**,** 996-1001.

HODGES, B. 2003. Validity and the OSCE. *Medical Teacher,* 25**,** 250-254.

HODGES, B. 2013. Assessment in the post-psychometric era: Learning to love the subjective and collective. *Medical Teacher,* 35**,** 564-568.

HODGES, B., REGEHR, G., HANSON, M. & MCNAUGHTON, N. 1998. Validation of an objective structured clinical examination in psychiatry. *Academic Medicine,* 73**,** 910-912.

HODGES, B., REGEHR, G. & MARTIN, D. 2001. Difficulties in recognizing one's own incompetence: novice physicians who are unskilled and unaware of it. *Academic Medicine,* 76**,** S87-S89.

HOLMBOE, E. S., HAWKINS, R. E. & HUOT, S. J. 2004. Effects of training in direct observation of medical residents' clinical competence: a randomized trial. *Annals of Internal Medicine,* 140**,** 874-881.

HOMER, M. & PELL, G. 2009. The impact of the inclusion of simulated patient ratings on the reliability of OSCE assessments under the borderline regression method. *Medical Teacher,* 31**,** 420-425.

HOMMES, J., RIENTIES, B., DE GRAVE, W., BOS, G., SCHUWIRTH, L. & SCHERPBIER, A. 2012. Visualising the invisible: a network approach to reveal the informal social side of student learning. *Advances in Health Sciences Education: Theory and Practice,* 17**,** 743-757.

HOPKINS, K. D. 1998. *Educational and Psychological Measurement and Evaluation,* Needham Heights, MA, Allynd & Bacon.

HOX, J. J. 2010. *Multilevel analysis: Techniques and applications,* New York, NY, Routledge.

HUNTER, M. L., OLIVER, R. & LEWIS, R. 2007. The effect of a community dental service outreach programme on the confidence of undergraduate students to treat children: a pilot study. *European Journal of Dental Education,* 11**,** 10-13.

HUSBANDS, A. & DOWELL, J. 2013. Predictive validity of the Dundee multiple mini-interview. *Medical Education,* 47**,** 717-725.

ICO 2018. Guide to the General Data Protection Regulation (GDPR). Information Commissioner's Office.

İLÇIN, N., TOMRUK, M., YEŞILYAPRAK, S. S., KARADIBAK, D. & SAVCı, S. 2018. The relationship between learning styles and academic performance in TURKISH physiotherapy students. *BMC Medical Education,* 18**,** 291-291.

ILGEN, J. S., MA, I. W., HATALA, R. & COOK, D. A. 2015. A systematic review of validity evidence for checklists versus global rating scales in simulation-based assessment. *Medical Education,* 49**,** 161-173.

INUWA, I. M., AL RAWAHY, M., TARANIKANTI, V. & HABBAL, O. 2011. Anatomy "steeplechase" online: necessity sometimes is the catalyst for innovation. *Anatomical Sciences Education,* 4**,** 115-118.

JANZ, N. K., ZIMMERMAN, M. A., WREN, P. A., ISRAEL, B. A., FREUDENBERG, N. & CARTER, R. J. 1996. Evaluation of 37 AIDS Prevention Projects: Successful Approaches and Barriers to Program Effectiveness. *Health education quarterly,* 23**,** 80-97.

JAVAEED, A. 2018. Assessment of Higher Ordered Thinking in Medical Education: Multiple Choice Questions and Modified Essay Questions. *MedEdPublish,* 7.

JAVAEED, A., KHAN, A. S., KHAN, S. H. & GHAURI, S. K. 2019. Perceptions of plagiarism among undergraduate medical students in Rawalpindi, Pakistan. *Pakistan Journal of Medical Sciences,* 35**,** 532-536.

JERVIS, A. & TILKI, M. 2011. Why are nurse mentors failing to fail student nurses who do not meet clinical performance standards? *British Journal of Nursing,* 20**,** 582.

JESTER, J. M., NIGG, J. T., BUU, A., PUTTLER, L. I., GLASS, J. M., HEITZEG, M. M., FITZGERALD, H. E. & ZUCKER, R. A. 2008. Trajectories of childhood aggression and inattention/hyperactivity: differential effects on substance abuse in adolescence. *Journal of the American Academy of Child and Adolescent Psychiatry,* 47**,** 1158-1165.

JOHNSON, R. B. & ONWUEGBUZIE, A. J. 2004. Mixed Methods Research: A Research Paradigm Whose Time Has Come. *Educational Researcher,* 33**,** 14-26.

JOHNSON, R. B., ONWUEGBUZIE, A. J. & TURNER, L. A. 2007. Toward a Definition of Mixed Methods Research. *Journal of Mixed Methods Research,* 1**,** 112-133.

JOINT COMMISSION ON NATIONAL DENTAL EXAMINATIONS 2020. The Joint Commission on National Dental Examinations (JCNDE) Test Item Development Guide. Joint Commission on National Dental Examinations.

JOLLY, B. 2014. Written assessment. *In:* SWANWICK, T. (ed.) *Understanding Medical Education: Evidence, Theory and Practice.* 2nd ed.: John Wiley & Sons, Ltd.

JOLLY, B. & DALTON, M. J. 2019. Written assessment. *In:* SWANWICK, T., FORREST, K. & O'BRIEN, B. C. (eds.) *Understanding Medical Education: Evidence, theory, and practice.* 3rd ed. Hoboken, NJ: Wiley-Blackwell.

JOLLY, B. & GRANT, J. 1997. *The Good Assessment Guide - A Practical Guide to Assessment and Appraisal for Higher Specialist Training,* London, Joint Centre for Education in Medicine.

JONES, B. L. & NAGIN, D. S. 2012. *A Stata Plugin for Estimating Group-Based Trajectory Models* [Online]. Available: https://ssrc.indiana.edu/doc/wimdocs/2013-03-29_nagin_trajectory_stata-plugin-info.pdf [Accessed 27th November 2020].

JONES, B. L. & NAGIN, D. S. 2013. A Note on a Stata Plugin for Estimating Group-based Trajectory Models. *Sociological Methods & Research,* 42**,** 608-613.

JONGE, L. D., TIMMERMAN, A. A., GOVAERTS, M. J., MURIS, J. W., MUIJTJENS, A. M., KRAMER, A. W. M. & VLEUTEN, C. P. M. V. D. 2017. Stakeholder perspectives on workplace-based performance assessment: towards a better understanding of assessor behaviour. *Advances in Health Sciences Education: Theory and Practice,* 22**,** 1213-1243.

JORDAN, S. & MITCHELL, T. 2009. E-assessment for learning? The potential of short free-text questions with tailored feedback. *British Journal of Educational Technology,* 40**,** 371–385.

JUNG, T. & WICKRAMA, K. A. S. 2008. An introduction to latent class growth analysis and growth mixture modeling. *Social and Personality Psychology Compass,* 2, 302–317.
KANE, M. 2006. Content-Related Validity Evidence in Test Development. *In:* DOWNING, S., M. & HALADYNA, T. M. (eds.) *Handbook of test development*. Lawrence Erlbaum Associates Publishers.
KANE, M. T. 1992. The assessment of professional competence. *Eval Health Prof,* 15, 163-182.
KANE, M. T. 2013. Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement,* 50, 1-73.
KARRAS, D. J. 1997. Statistical methodology: II. reliability and variability assessment in study design, part A. *Academic Emergency Medicine : Official Journal of the Society for Academic Emergency Medicine,* 4, 64-71.
KASS, R. E. & RAFTERY, A. E. 1995. Bayes Factors. *Journal of the American Statistical Association,* 90, 377-395.
KEARNEY, R. A., PUCHALSKI, S. A., YANG, H. Y. & SKAKUN, E. N. 2002. The inter-rater and intra-rater reliability of a new Canadian oral examination format in anesthesia is fair to good. *Canadian Journal of Anesthesia,* 49, 232-236.
KEMP, J. E., MORRISON, G. R. & ROSS, S. M. 1994. Developing evaluation instruments. *Designing effective instruction*. New York, NY: MacMillan College Publishing.
KEYNAN, A., FRIEDMAN, M. & BENBASSAT, J. 1987. Reliability of global rating scales in the assessment of clinical competence of medical students. *Medical Education,* 21, 477-481.
KINCHIN, I. M., CABOT, L. B., KOBUS, M. & WOOLFORD, M. 2011. Threshold concepts in dental education. *European Journal of Dental Education,* 15, 210-215.
KITZINGER, J. 1994. The methodology of focus groups: the importance of interaction between research participants. *Sociology of Health & Illness,* 16, 103-121.
KLIJN, S. L., WEIJENBERG, M. P., LEMMENS, P., VAN DEN BRANDT, P. A. & LIMA PASSOS, V. 2017. Introducing the fit-criteria assessment plot – A visualisation tool to assist class enumeration in group-based trajectory modelling. *Statistical Methods in Medical Research,* 26, 2424-2436.
KLINE, P. 2000. *A Pyschometrics Primer,* London, Free Association Books.
KNIVETON, B. H. 1996. A correlational analysis of multiplechoice and essay assessment measures. *Research in Education,* 56, 73-84.
KOGAN, J. R., CONFORTI, L. N., BERNABEO, E., IOBST, W. & HOLMBOE, E. 2015. How faculty members experience workplace-based assessment rater training: a qualitative study. *Medical Education,* 49, 692-708.
KOGAN, J. R., HOLMBOE, E. S. & HAUER, K. E. 2009. Tools for Direct Observation and Assessment of Clinical Skills of Medical Trainees: A Systematic Review. *JAMA : the journal of the American Medical Association,* 302, 1316-1326.
KORNELL, N. & BJORK, R. A. 2007. The promise and perils of self-regulated study. *Psychonomic Bulletin & Review,* 14, 219-224.
KORNELL, N. & BJORK, R. A. 2008. Optimising self-regulated study: The benefits-and costs-of dropping flashcards. *Memory (Hove),* 16, 125-136.
KOTHE, C., HAMPE, W. & HISSBACH, J. 2013. Das Hamburger Auswahlverfahren in der Zahnmedizin – Einführung des HAM-Nat als fachspezifischer

Studierfähigkeitstest [The Hamburg Selection Procedure for Dental Students – Introduction of the HAM-Nat as subject-specific test for study aptitude]. *GMS Zeitschrift für Medizinische Ausbildung,* 30**,** Doc46.

KOTHE, C., HISSBACH, J. & HAMPE, W. 2014. Prediction of practical performance in preclinical laboratory courses - the return of wire bending for admission of dental students in Hamburg. *GMS Zeitschrift für Medizinische Ausbildung,* 31**,** Doc22-Doc22.

KOVACIC, D. 2018. Using the Content Validity Index to Determine Content Validity of an Instrument Assessing Health Care Providers' General Knowledge of Human Trafficking. *Journal of Human Trafficking,* 4**,** 327-335.

KRAMER, A., MUIJTJENS, A., JANSEN, K., DÜSMAN, H., TAN, L. & VAN DER VLEUTEN, C. 2003. Comparison of a rational and an empirical standard setting procedure for an OSCE. *Medical Education,* 37**,** 132-139.

KRAMER, G. A., ALBINO, J. E., ANDRIEU, S. C., HENDRICSON, W. D., HENSON, L., HORN, B. D., NEUMANN, L. M. & YOUNG, S. K. 2009. Dental student assessment toolbox. *Journal of Dental Education,* 73**,** 12-35.

KRATHWOHL, D. R. 1993. *Methods of Educational and Social Science Research: An Integrated Approach,* New York, Longman/Addison Wesley Longman.

KREITER, C. D., FERGUSON, K. & GRUPPEN, L. D. 1999. Evaluating the usefulness of computerized adaptive testing for medical in-course assessment. *Academic Medicine,* 74**,** 1125-1128.

KRUEGER, R. A. 1998. *Developing questions for focus groups,* Thousand Oaks, Calif;London;, SAGE Publications.

KRUGER, J. & DUNNING, D. 1999. Unskilled and Unaware of It: How Difficulties in Recognizing One's Own Incompetence Lead to Inflated Self-Assessments. *Journal of personality and social psychology,* 77**,** 1121-1134.

KUDER, G. F. & RICHARDSON, M. W. 1937. The theory of the estimation of test reliability. *Psychometrika,* 2**,** 151-160.

KUHPAYEHZADE, J., HEMMATI, A., BARADARAN, H., MIRHOSSEINI, F. & SARVIEH, M. 2014. Validity and Reliability of Direct Observation of Procedural Skills in Evaluating Clinical Skills of midwifery Students of Kashan Nursing and Midwifery School. *Journal of Sabzevar University of Medical Sciences,* 21**,** 145-154.

KUMWENDA, B., CLELAND, J. A., WALKER, K., LEE, A. J. & GREATRIX, R. 2017. The relationship between school type and academic performance at medical school: a national, multi-cohort study. *BMJ Open,* 7**,** e016291-e016291.

LALA, R., WOOD, D. & BAKER, S. 2013. Validity of the UKCAT in Applicant Selection and Predicting Exam Performance in UK Dental Students. *Journal of Dental Education,* 77**,** 1159-1170.

LAMBE, P., KAY, E. & BRISTOW, D. 2018. Exploring uses of the UK Clinical Aptitude Test-situational judgement test in a dental student selection process. *European Journal of Dental Education,* 22**,** 23-29.

LAND, R., MEYER, J. & SMITH, J. 2008. *Threshold concepts within the disciplines,* Rotterdam, Sense Publishers.

LAROCQUE, S. & LUHANGA, F. L. 2013. Exploring the issue of failure to fail in a nursing program. *International journal of nursing education scholarship,* 10.

LARSEN, D. P., BUTLER, A. C. & ROEDIGER, H. L., 3RD 2008. Test-enhanced learning in medical education. *Medical Education,* 42**,** 959-966.

LARSON, J. L., WILLIAMS, R. G., KETCHUM, J., BOEHLER, M. L. & DUNNINGTON, G. L. 2005. Feasibility, reliability and validity of an operative performance rating system for evaluating surgery residents. *Surgery,* 138, 640-647; discussion 647-649.

LAU, S. T., ANG, E., SAMARASEKERA, D. D. & SHOREY, S. 2020. Development of undergraduate nursing entrustable professional activities to enhance clinical care and practice. *Nurse Education Today,* 87, 104347.

LEACOCK, C. & CHODOROW, M. 2003. C-rater: Automated Scoring of Short-Answer Questions. *Computers and the Humanities,* 37, 389–405.

LEADBEATTER, D., MADDEN, J., ROSS, B. & RUSSELL, E. 2020. Transition to dental practice: Newly graduated dentists' views of being successful in dental practice. *European Journal of Dental Education,* 24, 753-762.

LEE, M. & WIMMERS, P. F. 2011. Clinical competence understood through the construct validity of three clerkship assessments. *Medical Education,* 45, 849-857.

LEECH, N. L. & ONWUEGBUZIE, A. J. 2009. A typology of mixed methods research designs. *Quality & Quantity: International Journal of Methodology,* 43, 265–275.

LEPPINK, J. & PÉREZ-FUSTER, P. 2017. We need more replication research – A case for test-retest reliability. *Perspectives on Medical Education,* 6, 158-164.

LICARI, F. W. & CHAMBERS, D. W. 2008. Some Paradoxes in Competency-Based Dental Education. *Journal of Dental Education,* 72, 8-18.

LIN, C., CHANG, J. Z., HSU, T., LIU, Y., YU, S., TSAI, S., LAI, E. H. & LIN, C. 2013. Correlation of rater training and reliability in performance assessment: Experience in a school of dentistry. *Journal of dental sciences,* 8, 256-260.

LIN, G., TU, J. X., ZHANG, H., WANG, H., HE, H. & GUNZLER, D. 2016. Modern methods for longitudinal data analysis, capabilities, caveats and cautions. *Shanghai archives of psychiatry,* 28, 293-300.

LINDLEY, D. V. 1958. Fiducial Distributions and Bayes' Theorem. *Journal of the Royal Statistical Society. Series B, Methodological,* 20, 102-107.

LINN, R. L., KLEIN, S. P. & HART, F. M. 1972. The nature and correlates of law school essay grades. *Educational and Psychological Measurement,* 32, 267–279.

LOCKYER, J., CARRACCIO, C., CHAN, M., HART, D., SMEE, S., TOUCHIE, C., HOLMBOE, E. S., FRANK, J. R. & COLLABORATORS, I. 2017. Core principles of assessment in competency-based medical education. *Medical Teacher,* 39, 609-616.

LÜFTENEGGER, M., FINSTERWALD, M., KLUG, J., BERGSMANN, E., VAN DE SCHOOT, R., SCHOBER, B. & WAGNER, P. 2016. Fostering pupils' lifelong learning competencies in the classroom: evaluation of a training programme using a multivariate multilevel growth curve approach. *European journal of developmental psychology,* 13, 719-736.

LURIE, S. J., MOONEY, C. J. & LYNESS, J. M. 2009. Measurement of the General Competencies of the Accreditation Council for Graduate Medical Education: A Systematic Review. *Academic Medicine,* 84, 301-309.

LYNCH, C. D., ASH, P. J., CHADWICK, B. L. & HANNIGAN, A. 2010. Effect of Community-Based Clinical Teaching Programs on Student Confidence: A View from the United Kingdom. *Journal of Dental Education,* 74, 510-516.

LYNCH, J., EVERETT, B., RAMJAN, L. M., CALLINS, R., GLEW, P. & SALAMONSON, Y. 2017. Plagiarism in nursing education: an integrative review. *Journal of Clinical Nursing,* 26, 2845-2864.

MACCORQUODALE, K. & MEEHL, P. E. 1948. On the distinction between hypothetical constructs and intervening variables. *Psychological Review,* 55, 95-107.

MACKENZIE, N. & KNIPE, S. 2006. Research dilemmas: Paradigms, methods and methodology. *Issues in Educational Research,* 16, 193-205.

MACLUSKEY, M., DURHAM, J., BALMER, C., BELL, A., COWPE, J., DAWSON, L., FREEMAN, C., HANSON, C., MCDONAGH, A., JONES, J., MILLSOPP, L. & OLIVER, R. 2011. Dental student suturing skills: a multicentre trial of a checklist-based assessment. *European Journal of Dental Education,* 15, 244-249.

MAGUIRE, M. & DELAHUNT, B. 2017. Doing a Thematic Analysis: A Practical, Step-by-Step Guide for Learning and Teaching Scholars. *Journal of Teaching and Learning in Higher Education,* 9.

MARKHAM, L. R. 1976. Influences of handwriting quality on teacher evaluation of written work. *American Educational Research Journal,* 13, 277-283.

MARRIOTT, J., PURDIE, H., CROSSLEY, J. & BEARD, J. D. 2011. Evaluation of procedure-based assessment for assessing trainees' skills in the operating theatre. *British Journal of Surgery,* 98, 450-457.

MASTERS, J. C., HULSMEYER, B. S., PIKE, M. E., LEICHTY, K., MILLER, M. T. & VERST, A. L. 2001. Assessment of multiple-choice questions in selected test banks accompanying text books used in nursing education. *Journal of Nursing Education,* 40, 25-32.

MAYS, K. A. & BRANCH-MAYS, G. L. 2016. A Systematic Review of the Use of Self-Assessment in Preclinical and Clinical Dental Education. *Journal of Dental Education,* 80, 902-913.

MCCONNELL, M. M., ST-ONGE, C. & YOUNG, M. E. 2015. The benefits of testing for learning on later performance. *Advances in Health Sciences Education: Theory and Practice,* 20, 305-320.

MCCOUBRIE, P. 2004. Improving the fairness of multiple-choice questions: a literature review. *Medical Teacher,* 26, 709-712.

MCDANIEL, M. A., ROEDIGER, H. L., 3RD & MCDERMOTT, K. B. 2007. Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review,* 14, 200-206.

MCKERLIE, R. Accessed 2021. Bachelor of Dental Surgery Professional Examinations - Years 1-4: Calculation of Grade Boundaries and Secondary Banding. Glasgow: Univerity of Glasgow Dental School.

MCLACHLAN, J. C. 2006. The relationship between assessment and learning. *Medical Education,* 40, 716-717.

MCLEOD, R., MIRES, G. & KER, J. 2012. Direct observed procedural skills assessment in the undergraduate setting. *Clinical Teacher,* 9, 228-232.

MCMANUS, I. C. 2012. The misinterpretation of the standard error of measurement in medical education: A primer on the problems, pitfalls and peculiarities of the three different standard errors of measurement. *Medical Teacher,* 34, 569-576.

MCMANUS, I. C., THOMPSON, M. & MOLLON, J. 2006. Assessment of examiner leniency and stringency ('hawk-dove effect') in the MRCP(UK) clinical examination (PACES) using multi-facet Rasch modelling. *BMC Medical Education,* 6, 42-42.

MCMILLAN, W. 2013. Transition to university: the role played by emotion. *European Journal of Dental Education,* 17**,** 169-176.
MEDICAL COUNCIL OF CANADA 2010. Guidelines for the Development of Multiple-Choice Questions. Medical Council of Canada.
MEDICAL SCHOOLS COUNCIL & DENTAL SCHOOLS COUNCIL. Accessed 2021. *Access to medicine and dentistry courses* [Online]. London: Medical Schools Council and Dental Schools Council. Available: https://www.bsms.ac.uk/_pdf/undergraduate/access-to-medicine-msc-guidelines.pdf [Accessed June 2021].
MELHUISH, E., SYLVA, K., SAMMONS, P., SIRAJ-BLATCHFORD, I. & TAGGART, B. 2011. Effective Pre-School, Primary & Secondary Education Project: An Investigation of Children's Learning Trajectories From 3 to 11 Years of Age in Literacy and Numeracy. London, UK: Institute of Education, University of London.
MEMON, M. A., JOUGHIN, G. R. & MEMON, B. 2010. Oral assessment and postgraduate medical examinations: establishing conditions for validity, reliability and fairness. *Advances in Health Science Education: Theory and Practice,* 15**,** 277-289.
MERTENS, D. M. 2003. Mixed methods and the politics of human research: The transformative-emancipatory perspective. *In:* TASHAKKORI, A. & TEDDLIE, C. (eds.) *Handbook of mixed methods in social and behavioral research.* Thousand Oaks, CA: SAGE.
MERTENS, D. M. 2005. *Research methods in education and psychology: integrating diversity with quantitative and qualitative approaches,* Thousand Oaks, SAGE.
MESSICK, S. 1989. Validity. *In:* LINN, R. L. (ed.) *Educational measurement* New York, NY: American Council on Education and Macmillan.
MILLER, G. E. 1990. The assessment of clinical skills/competence/performance. *Academic Medicine,* 65**,** S63-67.
MITTON, C., PEACOCK, S., STORCH, J., SMITH, N. & CORNELISSEN, E. 2010. Moral Distress among Healthcare Managers: Conditions, Consequences and Potential Responses. *Healthcare policy,* 6**,** 99-112.
MOINEAU, G., POWER, B., PION, A. M., WOOD, T. J. & HUMPHREY-MURTO, S. 2011. Comparison of student examiner to faculty examiner scoring and feedback in an OSCE. *Medical Education,* 45**,** 183-191.
MONTI, M., KLOCKNER-CRONAUER, C., HAUTZ, S. C., SCHNABEL, K. P., BRECKWOLDT, J., JUNOD-PERRON, N., FELLER, S., BONVIN, R. & HUWENDIEK, S. 2020. Improving the assessment of communication competencies in a national licensing OSCE: lessons learned from an experts' symposium. *BMC Medical Education,* 20**,** 171.
MORA, P. A., BENNETT, I. M., ELO, I. T., MATHEW, L., COYNE, J. C. & CULHANE, J. F. 2009. Distinct trajectories of perinatal depressive symptomatology: evidence from growth mixture modeling. *American Journal of Epidemiology,* 169**,** 24-32.
MORGAN, D. L. 2007. Paradigms lost and pragmatism regained: Methodological implications of combining qualitative and quantitative methods. *Journal of Mixed Methods Research,* 1**,** 48-76.
MORGAN, D. L. 2014. Pragmatism as a paradigm for social research. *Qualitative Inquiry,* 20**,** 1045–1053.
MORRELL, D. 1984. The Assessment of Clinical Competence in General Family Practice. *Journal of the Royal Society of Medicine,* 77**,** 87.

MORRISON, S. & FREE, K. W. 2001. Writing multiple-choice test items that promote and measure critical thinking. *Journal of Nursing Education,* 40**,** 17-24.

MORSE, J. M. 1991. Approaches to qualitative-quantitative methodological triangulation. *Nursing Research,* 40**,** 120-123.

MOSS, P. 2007. Reconstructing validity. *Educational Researcher,* 36.

MUELLER, J. 2005. The Authentic Assessment Toolbox: Enhancing Student Learning through Online Faculty Development 1.

MUIJS, D. 2010. *Doing Quantitative Research in Education with SPSS,* London, SAGE Publications.

MULDER, H., TEN CATE, O., DAALDER, R. & BERKVENS, J. 2010. Building a competency-based workplace curriculum around entrustable professional activities: The case of physician assistant training. *Medical Teacher,* 32**,** e453-459.

MUTHÉN, B. & SHEDDEN, K. 1999. Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics,* 55**,** 463–469.

MUTHÉN, B. O. 2001. Latent variable mixture modeling. *In:* MARCOULIDES, G. A. & SCHUMACKER, R. E. (eds.) *New developments and techniques in structural equation modeling*. Mahwah, NJ: Lawrence Erlbaum Associates.

MUZZIN, L. J. 1995. Oral examination. *In:* SHANNON, S. & NORMAN, G. (eds.) *Evaluation methods: a resource handbook*. Hamilton: McMaster University.

NAEEM, N. 2013. Validity, reliability, feasibility, acceptability and educational impact of direct observation of procedural skills (DOPS). *Journal of the College of Physicians and Surgeons Pakistan,* 23**,** 77-82.

NAGIN, D. 2005. *Group-Based Modeling of Development,* Cambridge, MA, Harvard University Press.

NAGIN, D. & PIQUERO, A. 2010. Using the Group-Based Trajectory Model to Study Crime Over the Life Course. *Journal of Criminal Justice Education,* 21**,** 105-116.

NAGIN, D. S. & ODGERS, C. L. 2010. Group-based trajectory modeling in clinical research. *Annual Review of Clinical Psychology,* 6**,** 109-138.

NAGIN, D. S. & TREMBLAY, R. E. 2001. Analyzing developmental trajectories of distinct but related behaviors: a group-based method. *Psychol Methods,* 6**,** 18-34.

NAGIN, D. S. & TREMBLAY, R. E. 2005. Developmental trajectory groups: Fact or a useful statistical fiction? *Criminology (Beverly Hills),* 43**,** 873-904.

NEARY, M. 2000. Responsive assessment of clinical competence: Part 1. *Nursing Standard,* 15**,** 34-36.

NEATH, A. A. & CAVANAUGH, J. E. 2012. The Bayesian information criterion: Background, derivation, and applications. *Wiley Interdisciplinary Reviews: Computational Statistics,* 4**,** 199-203.

NENDAZ, M. R. & TEKIAN, A. 1999. Assessment in problem-based learning medical schools: A literature review. *Teaching and Learning in Medicine,* 11**,** 232-243.

NERI, M. T. & KROLL, T. 2003. Understanding the consequences of access barriers to health care: experiences of adults with disabilities. *Disability and Rehabilitation,* 25**,** 85–96.

NEVE, H., LLOYD, H. & COLLETT, T. 2017. Understanding students' experiences of professionalism learning: a 'threshold' approach. *Teaching in higher education,* 22**,** 92-108.

NEWBLE, D. 2004. Techniques for measuring clinical competence: objective structured clinical examinations. *Medical Education,* 38**,** 199-203.
NEWBLE, D. I., HOARE, J. & SHELDRAKE, P. F. 1980. The selection and training of examiners for clinical examinations. *Medical Education,* 14**,** 345-349.
NEWBLE, D. I. & SWANSON, D. B. 1988. Psychometric characteristics of the objective structured clinical examination. *Medical Education,* 22**,** 325-334.
NEWSTEAD, S. & DENNIS, I. 1994. The reliability of exam marking in psychology: examiners
examined. *Psychologist,* 7**,** 216–219.
NHS EDUCATION FOR SCOTLAND Accessed June 2021. How to assess your Vocational Dental Practitioner (VDP) using the Longitudinal Evaluation of Performance (LEP). Glasgow: NHS Education for Scotland.
NICHOLSON, N. 1984. A Theory of Work Role Transitions. *Administrative science quarterly,* 29**,** 172-191.
NICKBAKHT, M., AMIRI, M. & LATIFI, S. M. 2013. Study of the reliability and validity of objective structured clinical examination (OSCE) in the assessment of clinical skills of audiology students. *Global Journal of Health Science,* 5**,** 64-68.
NORCINI, J., ANDERSON, B., BOLLELA, V., BURCH, V., COSTA, M. J., DUVIVIER, R., GALBRAITH, R., HAYS, R., KENT, A., PERROTT, V. & ROBERTS, T. 2011. Criteria for good assessment: Consensus statement and recommendations from the Ottawa 2010 Conference. *Medical Teacher,* 33**,** 206-214.
NORCINI, J. J., SWANSON, D. B., GROSSO, L. J. & WEBSTER, G. D. 1985. Reliability, validity and efficiency of multiple choice question and patient management problem item formats in assessment of clinical competence. *Medical Education,* 19**,** 238-247.
NORCINI, J. J. & ZAIDI, Z. 2019. Workplace assessment. *In:* SWANWICK, T., FORREST, K. & O'BRIEN, B. C. (eds.) *Understanding Medical Education: Evidence, Theory, and Practice*. Hoboken, NJ: Wiley-Blackwell.
NORMAN, G. 2000. Examining the examination: Canadian versus US certification exam. *Canadian Association of Radiologists Journal,* 51**,** 208-209; author reply 211.
NORMAN, G., BORDAGE, G., PAGE, G. & KEANE, D. 2006. How specific is case specificity? *Medical Education,* 40**,** 618-623.
NUNNALLY, J. C. & BERNSTEIN, I. H. 1994. The Assessment of Reliability. *Psychometric Theory,* 3**,** 248-292.
NUNNALLY, J. O. 1978. *Psychometric Theory,* New York, McGraw Hill.
ORCUTT, H. K., ERICKSON, D. J. & WOLFE, J. 2004. The course of PTSD symptoms among Gulf War veterans: a growth mixture modeling approach. *Journal of Traumatic Stress,* 17**,** 195-202.
PALMER, D. & RIDEOUT, E. 1995. Essays. *In:* SHANNON, S. & NORMAN, G. (eds.) *Evaluation methods: a resource handbook*. Hamilton, Canada: McMaster University.
PANCZYK, M. & GOTLIB, J. 2015. Assessment of reliability, sensitivity, objectivity and validity of MCQ Pharmacology Exams as a potential output variable for predictive analysis. *Indian Journal of Pharmaceutical Education and Research,* 49**,** 1-9.
PARK, R. S., CHIBNALL, J. T., BLASKIEWICZ, R. J., FURMAN, G. E., POWELL, J. K. & MOHR, C. J. 2004. Construct validity of an objective structured clinical examination (OSCE) in psychiatry: associations with the clinical skills examination and other indicators. *Academic Psychiatry,* 28**,** 122-128.

PARK, S. E., ANDERSON, N. K. & KARIMBUX, N. Y. 2016. OSCE and Case Presentations As Active Assessments of Dental Student Performance. *Journal of Dental Education,* 80**,** 334-338.

PASCUAL RAMOS, V., MEDRANO RAMÍREZ, G., SOLÍS VALLEJO, E., BERNARD MEDINA, A. G., FLORES ALVARADO, D. E., PORTELA HERNÁNDEZ, M., ANDRADE ORTEGA, L., VERA LASTRA, O., ESPINOSA MORALES, R., MIRANDA LIMÓN, J. M., MALDONADO VELÁZQUEZ MDEL, R., JARA QUEZADA, L. J., AMEZCUA GUERRA, L. M., LÓPEZ ZEPEDA, J., SAAVEDRA SALINAS, M. Á. & ARCE SALINAS, C. A. 2015. Performance of an objective structured clinical examination in a national certification process of trainees in rheumatology. *Reumatologia Clinica,* 11**,** 215-220.

PATEL, U. S., TONNI, I., GADBURY-AMYOT, C., VAN DER VLEUTEN, C. P. M. & ESCUDIER, M. 2018. Assessment in a global context: An international perspective on dental education. *European Journal of Dental Education,* 22 Suppl 1**,** 21-27.

PATTON, M. Q. 1990. *Qualitative evaluation and research methods,* Newbury Park, CA, SAGE Publications.

PETRUSA, E. R. 2002. Clinical Performance Assessment. *In:* NORMAN, G. R., VAN DER VLEUTEN, C.P.M., NEWBLE, D.I. (ed.) *International Handbook on Research in Medical Education.*

PITTENGER, A. L., CHAPMAN, S. A., FRAIL, C. K., MOON, J. Y., UNDEBERG, M. R. & ORZOFF, J. H. 2016. Entrustable Professional Activities for Specialty Pharmacy Practice. *American Journal of Pharmaceutical Education,* 80**,** 178.

POLIT, D. & HUNGLER, B. 1999. *Nursing Research: Principle and Method,* Philadelphia, Lippincott Company.

POLIT, D. F., BECK, C. T. & OWEN, S. V. 2007. Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Research in Nursing & Health,* 30**,** 459-467.

PRESCOTT-CLEMENTS, L., HURST, Y. & RENNIE, J. S. 2003. Satisfactory completion of dental vocational training in Scotland: a system of assessment. *British dental journal,* Suppl**,** 17-21.

PRESCOTT-CLEMENTS, L., VAN DER VLEUTEN, C. P., SCHUWIRTH, L. W., HURST, Y. & RENNIE, J. S. 2008. Evidence for validity within workplace assessment: the Longitudinal Evaluation of Performance (LEP). *Medical Education,* 42**,** 488-495.

PRESCOTT, L. E., MCKINLAY, P. & RENNIE, J. S. 2001. The development of an assessment system for dental vocational training and general professional training: a Scottish approach. *British dental journal,* 190**,** 41-44.

PRESCOTT, L. E., NORCINI, J. J., MCKINLAY, P. & RENNIE, J. S. 2002. Facing the challenges of competency-based assessment of postgraduate dental training: Longitudinal Evaluation of Performance (LEP). *Medical Education,* 36**,** 92-97.

PRICE, M., RUST, C., O'DONOVAN, B., HANDLEY, K. & BRYANT, R. 2012. *Assessment Literacy: The Foundation for Improving Student Learning,* Wheatley, Oxford Brookes University.

PRINCE, M. 2012. Epidemiology - Core Psychiatry. *In:* WRIGHT, P., STERN, J. & PHELAN, M. (eds.) *Core psychiatry.* Saunders Elsevier.

PUTHIAPARAMPIL, T. & RAHMAN, M. M. 2020. Very short answer questions: a viable alternative to multiple choice questions. *BMC Medical Education,* 20**,** 141.

QAA 2002. Subject Benchmark Statement: Dentistry. The Quality Assurance Agency for Higher Education.
QAA 2018. UK Quality Code for Higher Education. Gloucester: The Quality Assurance Agency for Higher Education.
QUIRK, M. E. 2006. *Intuition and metacognition in medical education: keys to developing expertise,* New York, NY, Springer Pub. Co.
RADEMAKERS, J., TEN CATE, T. J. & BAR, P. R. 2005. Progress testing with short answer questions. *Medical Teacher,* 27**,** 578-582.
RAFTERY, A. E. 1995. Bayesian Model Selection in Social Research. *Sociological Methodology,* 25**,** 111-163.
RAHAYU, G. R., SUHOYO, Y., NURHIDAYAH, R., HASDIANDA, M. A., DEWI, S. P., CHANIAGO, Y., WIKANINGRUM, R., HARIYANTO, T., WONODIREKSO, S. & ACHMAD, T. 2016. Large-scale multi-site OSCEs for national competency examination of medical doctors in Indonesia. *Medical Teacher,* 38**,** 801-807.
RAJAN, S., CHEN, H. Y., CHEN, J. J., CHIN-YOU, S., CHEE, S., CHRUN, R., BYUN, J. & ABUZAR, M. 2020. Final year dental students' self-assessed confidence in general dentistry. *European Journal of Dental Education,* 24**,** 233-242.
RAUF, L. 2021. Case-Based Discussion in United Kingdom General Practice Training: A Critical Analysis. *Curēus,* 13**,** e13166-e13166.
REGAN, S., WONG, C., LASCHINGER, H. K., CUMMINGS, G., LEITER, M., MACPHEE, M., RHÉAUME, A., RITCHIE, J. A., WOLFF, A. C., JEFFS, L., YOUNG-RITCHIE, C., GRINSPUN, D., GURNHAM, M. E., FOSTER, B., HUCKSTEP, S., RUFFOLO, M., SHAMIAN, J., BURKOSKI, V., WOOD, K. & READ, E. 2017. Starting Out: qualitative perspectives of new graduate nurses and nurse leaders on transition to practice. *Journal of Nursing Management,* 25**,** 246-255.
REGEHR, G., MACRAE, H., REZNICK, R. K. & SZALAY, D. 1998. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Academic Medicine,* 73**,** 993-997.
REZNICK, R., REGEHR, G., MACRAE, H., MARTIN, J. & MCCULLOCH, W. 1997. Testing technical skill via an innovative 'bench station' examination. *American Journal of Surgery,* 173**,** 226–230.
ROBERTS, C., NEWBLE, D., JOLLY, B., REED, M. & HAMPTON, K. 2006. Assuring the quality of high-stakes undergraduate assessments of clinical competence. *Medical Teacher,* 28**,** 535-543.
ROBSON, C. & MCCARTAN, K. 2016. *Real World Research*, Wiley.
RODRIGUEZ, M. C. 2003. Construct Equivalence of Multiple-Choice and Constructed-Response Items: A Random Effects Synthesis of Correlations. *Journal of Educational Measurement,* 40**,** 163-184.
ROEDER, K., LYNCH, K. G. & NAGIN, D. S. 1999. Modeling Uncertainty in Latent Class Membership: A Case Study in Criminology. *Journal of the American Statistical Association,* 94**,** 766-776.
ROSS, M., CARROLL, G., KNIGHT, J., CHAMBERLAIN, M., FOTHERGILL-BOURBONNAIS, F. & LINTON, J. 1988. Using the OSCE to measure clinical skills performance in nursing. *Journal of Advanced Nursing,* 13**,** 45-56.
ROSSMAN, G. B. & WILSON, B. L. 1985. Numbers and words: Combining quantitative and qualitative methods in a single large-scale evaluation study. *Evaluation Review,* 9.

ROUDSARI, R. V. 2017. *Assessment of Competence in Dentistry: The Expectations, Perceptions, and Predictions*. Doctor of Philosophy in Clinical Dentistry, University of Manchester.

ROWNTREE, D. 1987. *Assessing Students: How Shall We Know Them,* London, Kogan Page.

ROYAL COLLEGE OF NURSING 2017. RCN Guidance for Mentors of Nursing and Midwifery Students. London: Royal College of Nursing.

ROYAL COLLEGE OF PHYSICIANS AND SURGEONS OF CANADA. 2019. *Information About the Royal College Short Answer Question (SAQ) Format* [Online]. Royal College of Physicians and Surgeons of Canada. Available: https://www.royalcollege.ca/rcsite/documents/credential-exams/short-answer-question-format-information-e [Accessed January 2021].

ROYAL, K. D. 2017. Four tenets of modern validity theory for medical education assessment and evaluation. *Advances in medical education and practice,* 8, 567-570.

RUSHFORTH, H. E. 2007. Objective structured clinical examination (OSCE): review of literature and implications for nursing education. *Nurse Education Today,* 27, 481-490.

RYDING, H. A. & MURPHY, H. J. 1999. Employing oral examinations (viva voce) in assessing dental students' clinical reasoning skills. *Journal of Dental Education,* 63, 682-687.

SAHEBALZAMANI, M. F. H. & JAHANTIGH, M. 2012. Validity and reliability of direct observation of procedural skills in evaluating the clinical skills of nursing students of Zahedan nursing and midwifery school. *Zahedan Journal of Research in Medical Sciences,* 14, 76-81.

SALDANA, K. 2013. *The coding manual for qualitative researchers,* Los Angeles, SAGE.

SALWEN, H. 2021. Threshold concepts, obstacles or scientific dead ends? *Teaching in higher education,* 26, 36-49.

SAM, A. H., FIELD, S. M., COLLARES, C. F., VAN DER VLEUTEN, C. P. M., WASS, V. J., MELVILLE, C., HARRIS, J. & MEERAN, K. 2018. Very-short-answer questions: reliability, discrimination and acceptability. *Medical Education,* 52, 447-455.

SAM, A. H., HAMEED, S., HARRIS, J. & MEERAN, K. 2016. Validity of very short answer versus single best answer questions for undergraduate assessment. *BMC Medical Education,* 16, 266.

SAM, A. H., WESTACOTT, R., GURNELL, M., WILSON, R., MEERAN, K. & BROWN, C. 2019. Comparing single-best-answer and very-short-answer questions for the assessment of applied medical knowledge in 20 UK medical schools: Cross-sectional study. *BMJ Open,* 9, e032550.

SAMUELS, A. 2006. Extended Matching Questions and the Royal Australian and New Zealand College of Psychiatrists written examination: an overview. *Australas Psychiatry,* 14, 63-66.

SCHIFFERDECKER, K. E. & REED, V. A. 2009. Using mixed methods research in medical education: basic guidelines for researchers. *Medical Education,* 43, 637-644.

SCHOONHEIM-KLEIN, M., MUIJTJENS, A., HABETS, L., MANOGUE, M., VAN DER VLEUTEN, C., HOOGSTRATEN, J. & VAN DER VELDEN, U. 2008. On the reliability of a dental OSCE, using SEM: effect of different days. *European Journal of Dental Education,* 12, 131-137.

SCHOONHEIM-KLEIN, M., MUIJTJENS, A., HABETS, L., MANOGUE, M., VAN DER VLEUTEN, C. & VAN DER VELDEN, U. 2009. Who will pass the dental OSCE?

Comparison of the Angoff and the borderline regression standard setting methods. *European Journal of Dental Education,* 13**,** 162-171.
SCHRAW, G., CRIPPEN, K. J. & HARTLEY, K. 2006. Promoting Self-Regulation in Science Education: Metacognition as Part of a Broader Perspective on Learning. *Research in science education (Australasian Science Education Research Association),* 36**,** 111-139.
SCHUBERT, A., TETZLAFF, J. E., TAN, M., RYCKMAN, J. V. & MASCHA, E. 1999. Consistency, inter-rater reliability, and validity of 441 consecutive mock oral examinations in anesthesiology: implications for use as a tool for assessment of residents. *Anesthesiology,* 91**,** 288-298.
SCHUWIRTH, L. & VAN DER VLEUTEN, C. 2004a. Merging views on assessment. *Medical Education,* 38**,** 1208–1210.
SCHUWIRTH, L., VAN DER VLEUTEN, C. & DURNING, S. J. 2017. What programmatic assessment in medical education can learn from healthcare. *Perspectives in Medical Education,* 6**,** 211–215.
SCHUWIRTH, L. W. & VAN DER VLEUTEN, C. P. 2003. ABC of learning and teaching in medicine: Written assessment. *British Medical Journal,* 326**,** 643-645.
SCHUWIRTH, L. W. & VAN DER VLEUTEN, C. P. 2004b. Different written assessment methods: what can be said about their strengths and weaknesses? *Medical Education,* 38**,** 974-979.
SCHUWIRTH, L. W. T. & VAN DER VLEUTEN, C. P. M. 2014. How to design a useful test: The principles of assessment. *In:* SWANWICK, T. (ed.) *Understanding Medical Education: Evidence, Theory and*
Practice. 2nd ed.: John Wiley & Sons, Ltd.
SCHUWIRTH, L. W. T. & VAN DER VLEUTEN, C. P. M. 2019. How to Design a Useful Test: The Principles of Assessment. *In:* SWANWICK, T., FORREST, K. & O'BRIEN, B. C. (eds.) *Understanding Medical Education: Evidence, Theory, and Practice*. 3rd ed. Hoboken, NJ: Wiley-Blackwell.
SCHUWIRTH, L. W. T., VERHEGGEN, M. M., VAN DER VLEUTEN, C. P. M., BOSHUIZEN, H. P. A. & DINANT, G. J. 2001. Do short cases elicit different thinking processes than factual knowledge questions do? *Medical Education,* 35.
SCHWARZ, G. 1978. Estimating the Dimension of a Model. *The Annals of Statistics,* 6**,** 461-464.
SCOTT, B. J. J., EVANS, D. J. P., DRUMMOND, J. R., MOSSEY, P. A. & STIRRUPS, D. R. 2001. An investigation into the use of a structured clinical operative test for the assessment of a clinical skill. *European Journal of Dental Education,* 5**,** 31-37.
SCOTTISH CHIEF DENTAL OFFICER 2020. NHS Dental Services. *In:* DIVISION, P. H. D.-C. D. O. D. (ed.). St Andrew's House, Edinburgh: Scottish Government.
SCOTTISH GOVERNMENT 2020. Coronavirus (COVID-19) update: First Minister's speech 24 March 2020. Scottish Parliament, Edinburgh: Scottish Government.
SCOULLER, K. 1998. The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay. *Higher Education,* 35**,** 453-472.
SCULLY, D. 2017. Constructing multiple-choice items to measure higher-order thinking. *Practical Assessment, Research & Evaluation,* 22**,** 1-13.
SEAL, A. 2016. Thematic analysis. *In:* GILBERT, G. N. & STONEMAN, P. (eds.) *Researching social life*. London: SAGE.

SETYONUGROHO, W., KENNEDY, K. M. & KROPMANS, T. J. 2015. Reliability and validity of OSCE checklists used to assess the communication skills of undergraduate medical students: A systematic review. *Patient Education and Counseling*, 1482-1491.

SHARMA, B. 2016. A focus on reliability in developmental research through Cronbach's Alpha among medical, dental and paramedical professionals. *Asian Pacific Journal of Health Sciences,* 3, 271-278.

SHEARER, D. 2016. *Longitudinal associations between periodontitis and glycated haemoglobin*. Doctor of Philosophy, University of Otago.

SHEPARD, L. 1993. Evaluating test validity. *In:* DARLING-HAMMOND, L. (ed.) *Review of research in education* Washington, DC: American Educational Research Association.

SHUMWAY, J. M. & HARDEN, R. M. 2003. AMEE Guide No. 25: The assessment of learning outcomes for the competent and reflective physician. *Medical Teacher,* 25, 569-584.

SILBER, C. G., NASCA, T. J., PASKIN, D. L., EIGER, G., ROBESON, M. & VELOSKI, J. J. 2004. Do global rating forms enable program directors to assess the ACGME competencies? *Academic Medicine,* 79, 549-556.

SIRECI, S. & FAULKNER-BOND, M. 2014. Validity evidence based on test content. *Psicothema,* 26, 100-107.

SKAKUN, E., MAGUIRE, T. & COOK, D. 1994. Strategy choices in multiple choice items. *Academic Medicine,* 69, S7–S9.

SMITH, C. D., WORSFOLD, K., DAVIES, L., FISHER, R. & MCPHAIL, R. 2013. Assessment literacy and student learning: the case for explicitly developing students 'assessment literacy'. *Assessment & Evaluation in Higher Education,* 38, 44-60.

SMITH, C. F. & MCMANUS, B. 2015. The integrated anatomy practical paper: A robust assessment method for anatomy education today. *Anatomical Sciences Education,* 8, 63-73.

STANFORD ENCYCLOPEDIA OF PHILOSOPHY. 2003. *Bayes' Theorem* [Online]. Stanford Encyclopedia of Philosophy. Available: https://plato.stanford.edu/entries/bayes-theorem/ [Accessed].

STOKES, E. 2011. *Rehabilitation outcome measures*, Churchill Livingstone.

STOPFORD, R. 2020. Threshold concepts and certainty: a critical analysis of 'troublesomeness'. *Higher Education,* 82, 163-179.

STUART, C. C. 2007. *Assessment, supervision and support in clinical practice: A guide for nurses, midwives and other health professionals,* Philadelphia, Churchill Livingstone Elsevier.

SULLIVAN, G. M. 2011. A primer on the validity of assessment instruments. *Journal of Graduate Medical Education,* 3, 119-120.

SURRY, L. T., TORRE, D. & DURNING, S. J. 2017. Exploring examinee behaviours as validity evidence for multiple-choice question examinations. *Medical Education,* 51, 1075-1085.

SUTCLIFFE, A. G., GARDINER, J. & MELHUISH, E. 2017. Educational Progress of Looked-After Children in England: A Study Using Group Trajectory Analysis. *Pediatrics,* 140.

SUTO, I., NADAS, R. & BELL, J. 2011. Who should mark what? A study of factors affecting marking accuracy in a biology examination. *Research Papers in Education,* 26, 21–52.

SUTO, W. M. I. & NADAS, R. 2008. What determines GCSE marking accuracy? An exploration of expertise among maths and physics markers. *Research Papers in Education,* 23, 477–497.

SUTTON, J. & AUSTIN, Z. 2015. Qualitative Research: Data Collection, Analysis, and Management. *The Canadian journal of hospital pharmacy,* 68**,** 226-231.

SWANSON, D. B., HOLTZMAN, K. Z. & ALLBEE, K. 2008. Measurement characteristics of content-parallel single-best-answer and extended-matching questions in relation to number and source of options. *Academic Medicine,* 83**,** S21-S24.

SWANSON, D. B., HOLTZMAN, K. Z., CLAUSER, B. E. & SAWHILL, A. J. 2005. Psychometric characteristics and response times for one-best-answer questions in relation to number and source of options. *Academic Medicine,* 80**,** S93-6.

SWANSON, D. B., NORMAN, G.R., LINN, R.L. 1995. Performance-based assessment: lessons from the health professions. *Educational Research,* 24 5–11.

TAGHVA, A., PANAGHI, L., RASOULIAN, M., BOLHARI, J., ZARGHAMI, M. & ESFAHANI, M. N. 2010. Evaluation of reliability and validity of the Psychiatry OSCE in Iran. *Academic Psychiatry,* 34**,** 154-157.

TANG, W., HE, H. & TU, X. M. 2012. *Applied Categorical and Count Data Analysis,* Boca Raton, FL, Chapman & Hall/CRC.

TARRANT, M., WARE, J. & MOHAMMED, A. M. 2009. An assessment of functioning and non-functioning distractors in multiple-choice questions: a descriptive analysis. *BMC Medical Education,* 9**,** 40.

TARRANT, M. & WARE, J. A. 2012. Framework for improving the quality of multiple-choice Assessments. *Nurse Educator,* 37.

TASHAKKORI, A. Growing pains? Agreements, disagreements, and new directions in conceptualizing mixed methods. 2nd annual Mixed Methods Conference, 2006 Humerton School of Health Studies, Cambridge, UK.

TASHAKKORI, A. & TEDDLIE, C. 1998. *Mixed Methodology: Combining Qualitative and Quantitative Approaches (Vol. 46),* Thousand Oaks, CA, SAGE Publications.

TASHAKKORI, A. & TEDDLIE, C. 2010. *SAGE Handbook of Mixed Methods in Social & Behavioral Research,* Thousand Oaks, California, SAGE Publications, Inc.

TAVAKOL, M. & DENNICK, R. 2011. Making sense of Cronbach's alpha. *International journal of medical education,* 2**,** 53-55.

TEKKOL, I. A. & DEMIREL, M. 2018. An Investigation of Self-Directed Learning Skills of Undergraduate Students. *Frontiers in psychology,* 9**,** 2324-2324.

TEN CATE, O. 2013. Competency-based education, entrustable professional activities, and the power of language. *Journal of Graduate Medical Education,* 5**,** 6-7.

TEN CATE, O. & TAYLOR, D. R. 2020. The recommended description of an entrustable professional activity: AMEE Guide No. 140. *Medical Teacher***,** 1-9.

TENNANT, M. & SCRIVA, J. 2000. Clinical assessment in dental education: a new method. *Australian Dental Journal,* 45**,** 125-130.

TENZIN, K., DORJI, T. & TENZIN, T. 2017. Construction of Multiple Choice Questions Before and After An Educational Intervention. *Journal of Nepal Medical Association,* 56**,** 112-116.

TENZIN, K., GYAMTSHO, S., WANGDON, T., BUTTIA, P. C., CHANDAN, L. & REGE, N. 2019. Effect of use of direct observation of procedural skills for assessment for learning in Obstetrics and Gynaecology postgraduate students at Medical University, Bhutan: a prospective study. *Bhutan Health Journal,* 5**,** 9-14.

TERRY, R., HING, W., ORR, R. & MILNE, N. 2017. Do coursework summative assessments predict clinical performance? A systematic review. *BMC Medical Education,* 17**,** 40-40.

THOMAS, R. M. 2003. *Blending qualitative and quantitative research methods in theses and dissertations,* Thousand Oaks, CA, Corwin.

THOMSON, W. M., SHEARER, D. M., BROADBENT, J. M., FOSTER PAGE, L. A. & POULTON, R. 2013. The natural history of periodontal attachment loss during the third and fourth decades of life. *Journal of Clinical Periodontology,* 40**,** 672-680.

TIRPUDE, A. P., GAIKWAD, M., TIRPUDE, P. A., JAIN, M. & BORA, S. 2019. Retrospective analysis of prevalent anatomy spotter's examination: an educational audit. *Korean Journal of Medical Education,* 31**,** 115-124.

TOBIN, J. 1958. Estimation of relationships for limited dependent variables. *Econometrica,* 26**,** 24–36.

TONNI, I., GADBURY-AMYOT, C. C., GOVAERTS, M., TEN CATE, O., DAVIS, J., GARCIA, L. T. & VALACHOVIC, R. W. 2020. ADEA-ADEE Shaping the Future of Dental Education III: Assessment in competency-based dental education: Ways forward. *Journal of Dental Education,* 84**,** 97-104.

TREJO-MEJIA, J. A., SANCHEZ-MENDIOLA, M., MENDEZ-RAMIREZ, I. & MARTINEZ-GONZALEZ, A. 2016. Reliability analysis of the objective structured clinical examination using generalizability theory. *Medical Education Online,* 21**,** 31650.

TSUI, K., LIU, C., LUI, J., LEE, S., TAN, R. & CHANG, P. 2013. Direct observation of procedural skills to improve validity of students' measurement of prostate volume in predicting treatment outcomes. *Urological Science,* 24**,** 84-88.

TURAN, S., DEMIREL, Ö. & SAYEK, S. 2009. Metacognitive awareness and self-regulated learning skills of medical students in different medical curricula. *Medical Teacher,* 31**,** e477-e483.

TURNBULL, J., DANOFF, D. & NORMAN, G. 1996. Content specificity and oral certification exams. *Medical Education,* 30**,** 56-59.

TURNITIN 2020. Turnitin Acquires ExamSoft, A Leading Assessment Platform. Turnitin.

UK HEALTH AND SAFETY EXECUTIVE. Accessed 2021. *Who regulates health and social care* [Online]. UK Health and Safety Executive,. [Accessed June 2021].

UMA, E., ISMAIL RASHID, A. H., ABAS, A. L., NETTEM, S., NAGRAJ, S. K. & MASTURA, N. 2017. Hybrid Tool for Assessment of Professionalism among Dental Undergraduate Students. *International journal of applied and basic medical research,* 7**,** S8-S14.

UNIVERSITY OF GLASGOW 2020. Guide to the Code of Assessment – 2: Grading student performance Glasgow: University of Glasgow.

UNIVERSITY OF GLASGOW. 2021a. *BDS Entry Requirements* [Online]. Glasgow: University of Glasgow,. Available: https://www.gla.ac.uk/schools/dental/undergraduate/academicentryrequirements/# [Accessed June 2021].

UNIVERSITY OF GLASGOW. 2021b. *Undergradute 2021 - Dentistry BDS* [Online]. University of Glasgow. Available: https://www.gla.ac.uk/undergraduate/degrees/dentistry/ [Accessed Jan 2021].

UNIVERSITY OF GLASGOW. 2021c. *Widening Participation* [Online]. Glasgow: University of Glasgow. Available: https://www.gla.ac.uk/study/wp/ [Accessed June 2021].
UNIVERSITY OF GLASGOW. Accessed 2021. *Final BDS Examination Regulations* [Online]. Glasgow: University of Glasgow. [Accessed].
UTKIN, L. V. 2006. A method for processing the unreliable expert judgments about parameters of probability distributions. *European journal of operational research,* 175, 385-398.
VAN BRUGGEN, L., MANRIQUE-VAN WOUDENBERGH, M., SPIERENBURG, E. & VOS, J. 2012. Preferred question types for computer-based assessment of clinical reasoning: a literature study. *Perspectives on Medical Education,* 1, 162-171.
VAN DER VLEUTEN, C. & NEWBLE, D. I. 1994. Methods of assessment in certification. *In:* NEWBLE, D. I., JOLLY, B. C. & WAKEFORD, R. E. (eds.) *The Certification and Recertification of Doctors: Issues in the Assessment of Clinical Competence*. Cambridge: Cambridge University Press.
VAN DER VLEUTEN, C. & VERHOEVEN, B. 2013. In-training assessment developments in postgraduate education in Europe. *ANZ Journal of Surgery,* 83, 454–459.
VAN DER VLEUTEN, C. P. & SCHUWIRTH, L. W. 2005. Assessing professional competence: from methods to programmes. *Medical Education,* 39, 309-317.
VAN DER VLEUTEN, C. P., SCHUWIRTH, L. W., DRIESSEN, E. W., DIJKSTRA, J., TIGELAAR, D., BAARTMAN, L. K. & VAN TARTWIJK, J. 2012. A model for programmatic assessment fit for purpose. *Medical Teacher,* 34, 205-214.
VAN DER VLEUTEN, C. P., SCHUWIRTH, L. W., SCHEELE, F., DRIESSEN, E. W. & HODGES, B. 2010. The assessment of professional competence: building blocks for theory development. *Best Practice & Research: Clinical Obstetrics & Gynaecology,* 24, 703-719.
VAN DER VLEUTEN, C. P. M. 1996. The assessment of professional competence: Developments, research and practical implications. *Advances in Health Sciences Education,* 1, 41-67.
VAN DER VLEUTEN, C. P. M. 2016. Revisiting 'Assessing professional competence: from methods to programmes'. *Medical Education,* 50, 885–888.
VAN HERPEN, S. G. A., MEEUWISSE, M., HOFMAN, W. H. A. & SEVERIENS, S. E. 2020. A head start in higher education: the effect of a transition intervention on interaction, sense of belonging, and academic performance. *Studies in Higher Education,* 45, 862-877.
VANDERBILT, A. A., FELDMAN, M. & WOOD, I. K. 2013. Assessment in undergraduate medical education: a review of course exams. *Medical Education Online,* 18, 1-5.
VARKEY, P., NATT, N., LESNICK, T., DOWNING, S. & YUDKOWSKY, R. 2008. Validity evidence for an OSCE to assess competency in systems-based practice and practice-based learning and improvement: a preliminary investigation. *Academic Medicine,* 83, 775-780.
VELOSKI, J. J., RABINOWITZ, H. K., ROBESON, M. R. & YOUNG, P. R. 1999. Patients don't present with five choices: an alternative to multiple-choice tests in assessing physicians' competence. *Academic Medicine,* 74, 539-546.
VERMA, M., CHATWAL, J. & SINGH, T. 1997. Reliability of essay type questions – effect of structuring. *Assessment in Education,* 4, 265–270.

VIOLATO, C. 1991. Item difficulty and discrimination as a function of stem completeness. *Psychological Reports,* 69, 739-743.
WAGENMAKERS, E. J. 2007. A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review,* 14, 779-804.
WAINER, H. & THISSEN, D. 1993. Combining multiple-choice and constructed-response test scores: toward a Marxist theory of test construction. *Applied Measurement in Education,* 6, 103-118.
WAKEFORD, R., SOUTHGATE, L. & WASS, V. 1995. Improving oral examinations: selecting, training, and monitoring examiners for the MRCGP. Royal College of General Practitioners. *British Medical Journal,* 311, 931-935.
WALKER, A., EARL, C., COSTA, B. & CUDDIHY, L. 2013. Graduate nurses' transition and integration into the workplace: A qualitative comparison of graduate nurses' and Nurse Unit Managers' perspectives. *Nurse Education Today,* 33, 291-296.
WALLERSTEDT, S., ERICKSON, G. & WALLERSTEDT, S. M. 2012. Short Answer Questions or Modified Essay Questions — More than a Technical Issue. *International Journal of Clinical Medicine,* 3, 28-30.
WALTERS, K., OSBORN, D. & RAVEN, P. 2005. The development, validity and reliability of a multimodality objective structured clinical examination in psychiatry. *Medical Education,* 39, 292-298.
WASS, V., MCGIBBON, D. & VAN DER VLEUTEN, C. 2001. Composite undergraduate clinical examinations: how should the components be combined to maximize reliability? *Medical Education,* 35, 326-330.
WASS, V., WAKEFORD, R., NEIGHBOUR, R., VAN DER VLEUTEN, C. & PRACTITIONERS, R. C. O. G. 2003. Achieving acceptable reliability in oral examinations: an analysis of the Royal College of General Practitioners membership examination's oral component. *Medical Education,* 37, 126-131.
WATLING, C. J., KENYON, C. F., SCHULZ, V., GOLDSZMIDT, M. A., ZIBROWSKI, E. & LINGARD, L. 2010. An Exploration of Faculty Perspectives on the In-Training Evaluation of Residents. *Academic Medicine,* 85, 1157-1162.
WATSON, R., STIMPSON, A., TOPPING, A. & POROCK, D. 2002. Clinical competence assessment in nursing: a systematic review of the literature. *Journal of Advanced Nursing,* 39, 421-431.
WEBB, N. L. Assessment literacy in a standards-based urban education setting. Annual Meeting of the American Educational Research Association, 2002 New Orleans, LA.
WEINSTEIN, S. E. & WU, S. 2009. Readiness assessment tests versus frequent quizzes: Student preferences. *International Journal on Teaching and Learning in Higher Education,* 21, 181–186.
WEISBURD, D., BUSHWAY, S., LUM, C. & YANG, S. M. 2004. Trajectories of crime at places: a longitudinal study of street segments in the city of Seattle. *Criminology,* 42.
WEITZ, G., VINZENTIUS, C., TWESTEN, C., LEHNERT, H., BONNEMEIER, H. & KÖNIG, I. R. 2014. Effects of a rater training on rating accuracy in a physical examination skills assessment. *GMS Zeitschrift für Medizinische Ausbildung,* 31, Doc41-Doc41.
WILBY, K. J., DOLMANS, D. H. J. M., AUSTIN, Z. & GOVAERTS, M. J. B. 2019. Assessors' interpretations of narrative data on communication skills in a summative OSCE. *Medical Education,* 53, 1003-1012.
WILES, C. M., DAWSON, K., HUGHES, T. A., LLEWELYN, J. G., MORRIS, H. R., PICKERSGILL, T. P., ROBERTSON, N. P. & SMITH, P. E. 2007. Clinical skills

evaluation of trainees in a neurology department. *Clinical Medicine,* 7**,** 365-369.

WILKINSON, J. R., CROSSLEY, J. G., WRAGG, A., MILLS, P., COWAN, G. & WADE, W. 2008. Implementing workplace-based assessment across the medical specialties in the United Kingdom. *Medical Education,* 42**,** 364-373.

WILKINSON, T. J. 2007. Assessment of clinical performance: gathering evidence. *Internal medicine journal,* 37**,** 631-636.

WILKINSON, T. J. & FRAMPTON, C. M. 2004. Comprehensive undergraduate medical assessments improve prediction of clinical performance. *Medical Education,* 38**,** 1111-1116.

WILLIAMS, J. C., BAILLIE, S., RHIND, S. M., WARMAN, S., SANDY, J. & IRELAND, A. 2015. A Guide to Assessment in Dental Education. University of Bristol.

WILLIAMS, R. G., KLAMEN, D. A. & MCGAGHIE, W. C. 2003. Cognitive, social and environmental sources of bias in clinical performance ratings. *Teaching and Learning in Medicine,* 15**,** 270-292.

WINCKEL, C. P., REZNICK, R. K., COHEN, R. & TAYLOR, B. 1994. Reliability and construct validity of a structured technical skills assessment form. *American Journal of Surgery,* 167**,** 423-427.

WIPULANUSAT, W., PANUWATWANICH, K., STEWART, R. A. & SUNKPHO, J. 2020. Applying Mixed Methods Sequential Explanatory Design to Innovation Management. *In:* PANUWATWANICH, K. & KO, C. H. (eds.) *The 10th International Conference on Engineering, Project, and Production Management*. Singapore: Springer.

WONG, W. Y. A., ROBERTS, C. & THISTLETHWAITE, J. 2020. Impact of Structured Feedback on Examiner Judgements in Objective Structured Clinical Examinations (OSCEs) Using Generalisability Theory. *Health professions education,* 6**,** 271-281.

WOOD, T. 2009. Assessment not only drives learning, it may also help learning. *Medical Education,* 43**,** 5-6.

WORMALD, B. W., SCHOEMAN, S., SOMASUNDERAM, A. & PENN, M. 2009. Assessment drives learning: an unavoidable truth? *Anatomical Sciences Education,* 2**,** 199-204.

WRAGG, A., WADE, W., FULLER, G., COWAN, G. & MILLS, P. 2003. Assessing the performance of specialist registrars. *Clinical Medicine,* 3**,** 131-134.

YANG, J. C. & LAUBE, D. W. 1983. Improvement of reliability of an oral examination by a structured evaluation instrument. *Journal of Medical Education,* 58**,** 864-872.

YAQINUDDIN, A., ZAFAR, M., IKRAM, M. F. & GANGULY, P. 2013. What is an objective structured practical examination in anatomy? *Anatomical Sciences Education,* 6**,** 125-133.

YEPES-RIOS, M., DUDEK, N., DUBOYCE, R., CURTIS, J., ALLARD, R. J. & VARPIO, L. 2016. The failure to fail underperforming trainees in health professions education: A BEME systematic review: BEME Guide No. 42. *Medical Teacher,* 38**,** 1092-1099.

YIU, C. K. Y., MCGRATH, C., BRIDGES, S., CORBET, E. F., BOTELHO, M. G., DYSON, J. E. & CHAN, L. K. 2012. Self-perceived preparedness for dental practice amongst graduates of The University of Hong Kong's integrated PBL dental curriculum. *European Journal of Dental Education,* 16**,** e96-e105.

YORKE, M. & LONGDEN, B. 2004. *Retention and student success in higher education,* Buckingham, Society for Research into Higher Education & Open University Press.

ZAFAR, M., YAQINUDDIN, A., IKRAM, F. & GANGULY, P. 2013. Practical Examinations - OSPE, OSCE and Spot. *In:* GANGULY, P. (ed.) *Education in Anatomical Sciences*. New York: Nova Publishers.

ZAMANZADEH, V., GHAHRAMANIAN, A., RASSOULI, M., ABBASZADEH, A., ALAVI-MAJD, H. & NIKANFAR, A. R. 2015. Design and Implementation Content Validity Study: Development of an instrument for measuring Patient-Centered Communication. *Journal of Caring Sciences,* 4**,** 165-178.

ZIJLSTRA-SHAW, S., ROBERTS, T. & ROBINSON, P. G. 2017. Evaluation of an assessment system for professionalism amongst dental students. *European Journal of Dental Education,* 21**,** e89-e100.

ZIMMERMAN, B. J. 1986. Becoming a self-regulated learner: Which are the key subprocesses? *Contemporary educational psychology,* 11**,** 307-313.

Appendix 1 – Assessment dataset variables

LIFTUPP© – list of variables available in pseudonymised data sets (part 1). Key variables marked with *.

Variable name	Description	Data type	Required for:
STUDENT*	Unique numerical code assigned to participants	Numerical	Data processing
GENDER*	Participant's gender/sex	Categorical	Analysis
YEAR_OF_STUDY*	Delineates which year student graduated.	Numerical	Analysis
STUDENT_GROUP	Code identifying the clinical group to which student belongs.	String	NA
STUDENT_YEAR_OF_STUDY*	BDS year in which data was recorded (BDS3-5)	Numerical	Data processing and analysis
CLINIC_DATE*	Date LIFTUPP© assessment was recorded	Interval	Data processing and analysis
CLINIC_START	Clinic start time	Interval	NA
CLINIC_END	Clinic end time	Interval	NA
TIMETABLE_STAFF*	Unique numerical code depicting faculty who conducted LIFTUPP© assessment	Numerical	Data processing and analysis
PRESENT	Binary code (0 or 1) indicating student attendance at clinic	Categorical	NA
FORM_TYPE*	Dental subject area being assessed (e.g., oral surgery)	Categorical	Data processing and analysis
FORM_ID*	Numerical ID code for FORM_TYPE	Numerical	Data processing
SECTION_TYPE*	Procedure/skill/attribute assessed	Categorical	Data processing and analysis
SECTION_ID*	Numerical ID code for SECTION_TYPE	Numerical	Data processing
SECTION_QUAD_AND_TOOTH*	Numerical code of tooth on which procedure being assessed was performed.	Numerical	Data processing
SECTION_MAXIMUM_DIFFICULTY	Procedural difficulty rating	Numerical	NA
QUESTION_TEXT*	Stage of procedure/skill being assessed	Categorical	Data processing
QUESTION_ID*	Numerical code for QUESTION_TEXT	Numerical	Data processing

QUESTION_ANALYSIS_ID	Numerical ID code for QUESTION_ANALYSIS_SHORT_NAME	Numerical	NA
QUESTION_ANALYSIS_SHORT_NAME	Shortened version of QUESTION_TEXT*	Categorical	NA
QUESTION_ANALYSIS DOMAIN*	GDC domain of clinical competence to which procedure/skill/attribute being assessed belongs	Categorical	Data processing and analysis

LIFTUPP© – list of variables available in pseudonymised data sets (part 2). Key variables marked with *.

Variable name	Description	Data type	Required for:
FORM_ANALYSIS_PROCEDURE_NAME*	Additional detail for procedure/skill/attribute assessed	Categorical	Data processing
FORM_ANALYSIS_PROCEDURE_CODE*	Numerical code for FORM_ANALYSIS_ PROCEDURE_CODE	Numerical	Data processing
TREATMENT_AREA_ID	Numerical ID code for TREATMENT_AREA	Numerical	NA
TREATMENT_AREA	Location where treatment was provided in patient's mouth.	Categorical	NA
TREATMENT_STAGE_ID	Numerical ID code for TREATMENT_STAGE_ID	Numerical	NA
TREATMENT_STAGE	Unknown – no data recorded under this variable.	Unknown	NA
TREATMENT_MATERIAL_ID	Numerical ID code for MATERIAL	Numerical	NA
MATERIAL*	Material used in the treatment item being assessed	Categorical	Data processing
DIAGNOSIS_GROUP_ID	Numerical ID code for DIAGNOSIS_GROUP_ID	Numerical	NA
DIAGNOSIS_GROUP	Categorise of diseases/conditions under which diagnosis fits.	Categorical	NA
DIAGNOSIS_ID	Numerical ID code for DIAGNOSIS	Numerical	NA
DIAGNOSIS	The disease or condition requiring treatment	Categorical	NA
QUESTION_RATING*	LIFTUPP© score awarded (1 to 6)	Numerical	Data processing and analysis
QUESTION_FEEDBACK	Written feedback given to student by assessor	String	NA
FEEDBACK_AGREED	Binary code (0 or 1) indicating if student agreed to feedback given by assessor	Numerical	NA
ANONYMISED_PATIENT_ID*	Unique alphanumeric code assigned to individual patients	String	Data processing
PATIENT_SIMULATED*	Binary code (0 or 1) indicating if procedure was performed on a simulated/mannequin patient	Numerical	Data processing
PATIENT_PEADS	Binary code (0 or 1) indicating if procedure was performed on a child patient	Numerical	NA

LIFTUPP© - Derived variables

Variable	Description	Data type	Required for:
TRANSFORMED_CLINIC_DATE	Numerical representation of date LIFTUPP© assessment was recorded. Presented as number of days since 1st January 1960.	Numerical	Data analysis
THRESHOLD_4_MET	Binary number (1 or 0) which depicts whether student achieved a performance indicator ≥4.	Numerical	Data analysis
THRESHOLD_5_MET	Binary number (1 or 0) which depicts whether student achieved a performance indicator ≥5.	Numerical	Data analysis

Undergraduate examinations - list of variables available in pseudonymised data sets. All undergraduate examination variables were regarded as key. MCQ = Multiple-choice Question. MSA = Multiple-short answer. OSCE = Objective structured clinical examination.

Variable name	Description	Data type	Required for:
STUDENT	Unique numerical code assigned to participants	Numerical	Data processing
GENDER*	Participant's gender/sex	Categorical	Analysis
SUMMAT1	Grade for BDS1 Summative Essay	Categorical	Analysis
WRITTEN1_PCT	Collective percentage score for 2 x BDS1 MCQs	Numerical	Data processing and analysis
WRITTEN1	Collective grade awarded for BDS1 written examinations (2 x MCQs)	Categorical	Analysis
OSCE1_PCT	Percentage score for BDS1 OSCE	Numerical	Data processing and analysis
OSCE1	Grade awarded for BDS1 OSCE	Categorical	Analysis
MCQ2_PCT	Percentage score for BDS2 MCQ	Numerical	Data processing and analysis
MSA2_PCT	Percentage score for BDS2 MSA	Numerical	Data processing and analysis
WRITTEN2	Collective grade awarded for BDS2 written exams (MCQ + MSA)	Categorical	Analysis
OSCE2_PCT	Percentage score for BDS2 OSCE	Numerical	Data processing and analysis
OSCE2	Grade awarded for BDS2 OSCE	Categorical	Analysis
ANAT3_PCT	Percentage score for BDS3 anatomy examination	Numerical	Data processing and analysis
MCQ3_PCT	Percentage score for BDS3 MCQ	Numerical	Data processing and analysis
MSA3_PCT	Percentage score for BDS2 MSA	Numerical	Data processing and analysis
WRITTEN3	Collective grade awarded for BDS3 written exams (anatomy + MCQ + MSA)	Categorical	Analysis
OSCE3_PCT	Percentage score for BDS3 OSCE	Numerical	Data processing and analysis
OSCE3	Grade awarded for BDS3 OSCE	Categorical	Analysis
MSA4_PCT	Percentage score for BDS4 final written examination (MSA)	Numerical	Data processing and analysis
MSA4	Grade awarded for BDS4 final written examination (MSA)	Categorical	Analysis
OSCE5_PCT	Percentage score for BDS5 final OSCE	Numerical	Data processing and analysis
OSCE5	Grade awarded for BDS5 final OSCE	Categorical	Analysis

NOTE: This table does not include variables for resit examination results as the need for these additional assessments varied between BDS years and cohorts.

Longitudinal evaluations of performance (LEPs) – list of variables available in pseudonymised data sets. Key variables marked with *.

Variable name	Description	Data type	Required for:
STUDENT*	Unique numerical code assigned to participants	Numerical	Data processing
GENDER*	Participant's gender/sex	Categorical	Analysis
LEP_ORDER*	Sequential number depicting order assessments were carried out in.	Numerical	Data processing and analysis
CODE_OF_POST*	Block of VDT year (1-3) in which assessed was conducted	Numerical	Data processing
EVALUATOR_STATUS*	Role of the recording assessor (VDT trainer or external assessor)	Categorical	Analysis
LEP_DATE*	Date LEP assessment was recorded	Interval	Data processing and analysis
LEP_TITLE*	Dental subject area or procedure being assessed*	Categorical	Data processing and analysis
DETAILS_OF_ENCOUNTER*	Description of procedure(s) being assessed (written by assessor)	String	Data processing
EXAMINATION_AND_CONSULTATION_SKILL	LEP score awarded for assessment of examination and consultant skills (1- 9).	Numerical	Analysis
CLINICAL_JUDGEMENT_AND_DIAGNOSIS	LEP score awarded for assessment of clinical judgement and consultant skills (1- 9).	Numerical	Analysis
TECHNICAL_ABILITY_AND_MANUAL_DEX*	LEP score awarded for assessment of hands-on clinical skills (1- 9).	Numerical	Analysis
COMMUNICATION_SKILLS	LEP score awarded for assessment of communication skills (1- 9).	Numerical	Analysis
PROFESSIONALISM	LEP score awarded for assessment of professionalism (1- 9).	Numerical	Analysis
KNOWLEDGE__LEVEL_AND_APPLICATION	LEP score awarded for assessment of knowledge and its clinical application (1- 9).	Numerical	Analysis
ORGANISATION	LEP score awarded for assessment of organisational skills (1- 9).	Numerical	Analysis
TRAINEE_S_INSIGHT_INTO_PERFORMANCE	LEP score awarded for assessment of trainee's ability to critique their own clinical performance (1- 9).	Numerical	Analysis

Longitudinal evaluations of performance (LEPs) - LIFTUPP© - Derived variables

Variable	Description	Data type	Required for:
TRANSFORMED_CLINIC_DATE	Numerical representation of date LEP assessment was recorded. Presented as number of days since 1st January 1960.	Numerical	Data analysis
THRESHOLD_4_MET	Binary number (1 or 0) which depicts whether VDP achieved a score ≥4.	Numerical	Data analysis
THRESHOLD_5_MET	Binary number (1 or 0) which depicts whether VDP achieved a score ≥5.	Numerical	Data analysis
THRESHOLD_6_MET	Binary number (1 or 0) which depicts whether VDP achieved a score ≥6.	Numerical	Data analysis
THRESHOLD_7_MET	Binary number (1 or 0) which depicts whether VDP achieved a score ≥7.	Numerical	Data analysis

Appendix 2 – LIFTUPP©: Key procedural stages

Dental subject	Procedure	Key stage(s)
Restorative dentistry	Direct restorations	"Appropriate restoration of tooth contour and anatomy"
	Endodontics	1. "Ability to gain appropriate access" 2. "Ability to identify canals" 3. "Ability to negotiate canals" 4. "Biomechanical prep" 5. "Obturation"
	Fissure sealant	"Appropriate restoration of tooth contour and anatomy"
	Indirect restorations	1. "Appropriate tooth reduction" 2. "Impression taking" 3. "Fit of indirect restoration"
	Occlusal splint	"Impression taking" "Fit"
	Periodontics	"Supra gingival debridement (Hand)" OR "Supra gingival debridement (Ultrasonic)"
	Prevention	Any prevention entry (e.g., oral hygiene instruction).
	Prosthodontics	1. "Impression taking (1st imps)" 2. "Impression taking (2nd imps)" 3. "Ability to perform a Registration" 4. "Try in teeth" 5. "Fit"
	Root surface debridement	"RSD"
Oral Surgery	Biopsy	"Appropriate incision"
	Extraction	"Appropriate tooth movement"
	Minor oral surgery	"Incision and raising flap"
	Suture	"Appropriate wound edge apposition"
Paediatric dentistry	Direct restorations	"Appropriate restoration of tooth contour and anatomy"
	Extraction	"Appropriate tooth movement"
	Fissure sealant	"Appropriate restoration of tooth contour and anatomy"
	Preformed crowns	"Crown fit"
	Prevention	Any prevention entry (e.g., oral hygiene instruction).
Radiology	Other views	"Radiology technique"
	Periapical views	"Radiology technique"

Appendix 3 – Group-based trajectory model variations simulated (LIFTUPP© and longitudinal evaluations of performance)

Single-group models	
0	2
1	3

Two-group models	
0 0	2 0
0 1	2 1
0 2	2 2
0 3	2 3
1 0	3 0
1 1	3 1
1 2	3 2
1 3	3 3

Three-group models	
0 0 0	2 0 0
0 0 1	2 0 1
0 0 2	2 0 2
0 0 3	2 0 3
0 1 0	2 1 0
0 1 1	2 1 1
0 1 2	2 1 2
0 1 3	2 1 3
0 2 0	2 2 0
0 2 1	2 2 1
0 2 2	2 2 2
0 2 3	2 2 3
0 3 0	2 3 0
0 3 1	2 3 1
0 3 2	2 3 2
0 3 3	2 3 3
1 0 0	3 0 0
1 0 1	3 0 1
1 0 2	3 0 2
1 0 3	3 0 3
1 1 0	3 1 0
1 1 1	3 1 1
1 1 2	3 1 2
1 1 3	3 1 3
1 2 0	3 2 0
1 2 1	3 2 1
1 2 2	3 2 2
1 2 3	3 2 3
1 3 0	3 3 0
1 3 1	3 3 1
1 3 2	3 3 2
1 3 3	3 3 3

Four-group models					
0000	0223	1112	2001	**2230**	3112
0001	0230	1113	2002	**2231**	3113
0002	0231	1120	2003	**2232**	3120
0003	0232	1121	2010	**2233**	3121
0010	0233	1122	2011	**2300**	3122
0011	0300	1123	2012	**2301**	3123
0012	0301	1130	2113	**2302**	3130
0113	0302	1131	2020	**2303**	3131
0020	0303	1132	2021	**2310**	3132
0021	0310	1133	2022	**2311**	3133
0022	0311	1200	2023	**2312**	3200
0023	0312	1201	2030	**2313**	3201
0030	0313	1202	2031	**2320**	3202
0031	0320	1203	2032	**2321**	3203
0032	0321	1210	2033	**2322**	3210
0033	0322	1211	2100	**2323**	3211
0100	0323	1212	2101	**2330**	3212
0101	0330	1213	2102	**2331**	3213
0102	0331	1220	2103	**2332**	3220
0103	0332	1221	2110	**2333**	3221
0110	0333	1222	2111	**3000**	3222
0111	1000	1223	2112	**3001**	3223
0112	1001	1230	2113	**3002**	3230
0113	1002	1231	2120	**3003**	3231
0120	1003	1232	2121	**3010**	3232
0121	1010	1233	2122	**3011**	3233
0122	1011	1300	2123	**3012**	3300
0123	1012	1301	2130	**3113**	3301
0130	1013	1302	2131	**3020**	3302
0131	1020	1303	2132	**3021**	3303
0132	1021	1310	2133	**3022**	3310
0133	1022	1311	2200	**3023**	3311
0200	1023	1312	2201	**3030**	3312
0201	1030	1313	2202	**3031**	3313
0202	1031	1320	2203	**3032**	3320
0203	1032	1321	2210	**3033**	3321
0210	1033	1322	2211	**3100**	3322
0211	1100	1323	2212	**3101**	3323
0212	1101	1330	2213	**3102**	3330
0213	1102	1331	2220	**3103**	3331
0220	1103	1332	2221	**3110**	3332
0221	1110	1333	2222	**3111**	3333
0222	1111	2000	2223		

Appendix 4 – Focus group participation leaflets, privacy notice and written consent forms

Participant information leaflet: Undergraduate students, vocational dental practitioners and recently qualified dentists

University of Glasgow | College of Medical, Veterinary & Life Sciences

Dental Education Study Information (Plain Language Statement) for

Undergraduate Dental Students/Vocational Dental Practitioners/Recently qualified dentists

1. **Study title**

Longitudinal assessment of dental students

2. **Invitation**

You are being invited to take part in an online focus group as part of a research study. Before you decide it is important for you to understand why the research is being done and what it will involve. Please take time to read the following information carefully and discuss it with others if you wish. Ask us if there is anything that is not clear or if you would like more information. Take time to decide whether you wish to take part.

3. **What is the purpose of the study?**

Over the past few years, the University of Glasgow Dental School has made significant investments in implementing a longitudinal assessment system, known as LIFTUPP©, into their undergraduate programme. Now they would like to know the extent to which LIFTUPP© can contribute to the assessment of clinical competence.

To start gaining an insight into this, LIFTUPP© data will need to be compared against methods currently used to assess competence performance at undergraduate level as well as the clinical performance of recently qualified dentists (who will have certified as competent practitioners using these current assessment methods).

4. **Why have I been chosen?**

If you have been given this information sheet, it is because you have been identified as a key stakeholder regarding LIFTUPP©, i.e., you are influenced by and/or contribute to LIFTUPP© data. We would like to know a bit more about your thoughts of LIFTUPP© when it is compared against other dental education assessment methods, so you have been invited to take part in a focus group to discuss your opinions with us, dental students and recently qualified vocational dental practitioners.

5. **Do I have to take part?**

Taking part is entirely voluntary. If you do decide to take part, you will be given this information sheet to keep and be asked to sign a consent form and privacy notice. If you decide to take part, you are still free to withdraw without giving a reason up until your

focus group discussion has been transcribed and anonymised (see below). After this point, we will no longer be able to identify what was said by individual participants and, therefore, we won't be able remove your responses. Your decision to take part or not will not affect your grades or academic records in any way.

6. **What will happen to me if I take part?**

If you take part, you will be invited to attend an online focus group on [date] using Microsoft Teams video conferencing. You will be shown the results of some comparisons between LIFTUPP© data, traditional undergraduate assessments (such as Objective Structured Clinical Examinations (OSCEs), Multiple Short Answer (MSA) papers and Multiple-Choice Question (MCQ) papers) and postgraduate longitudinal clinical assessment data (known as LEPs). You will then be involved in a group discussion (focus group) about your thoughts and opinions on how the results from these comparisons could be used to improve educational assessment within dentistry. This discussion will be led by the projects lead researcher (Jamie Dickie).

Audio from the resulting conversations will recorded using Microsoft Team's "record" feature - which can only be activated by the projects lead researcher. A backup copy of the audio will be recorded on a Dictaphone placed next to the speakers of the lead researcher's laptop computer.

A video recording of the meeting will also be made. However, you do not have to be filmed as part of the meeting should you not wish to. You will be able to join the meeting with audio input only and can activate video input at your own discretion.

The audio and video recordings made via Microsoft Teams then be automatically uploaded into the lead researcher's Microsoft Streams account –a cloud data storage platform which is part of the PhD researcher's password protected University of Glasgow Office 365 account. Using the transcription feature within Microsoft Streams, the audio from the focus group will be anonymously transcribed and anonymized codes will be used to refer to participants thereafter rather than personal information.

The backup audio from the Dictaphone will be immediately uploaded to a password protected folder on the University of Glasgow secure networked "J-Drive". Once the transcriptions of the discussion have been completed and checked. All audio and video recording files will be deleted.

Your responses are for research purposes only and will not contribute to your academic records in any way.

Each focus group will last for approximately one-and-a-half hours (90 minutes).

7. **What do I have to do?**

You will have to take part in a group discussion about comparisons between LIFTUPP©, traditional undergraduate assessments and postgraduate assessment with your peers (as detailed above).

8. **What will I need to take part?**

You will need a unique meeting ID and password for gaining entry to the meeting. This will be emailed to you by the lead researcher if you consent to participate.

You will also need:

- A computer, laptop, tablet or mobile phone with a stable internet connection;
- A microphone which permit audio input into your computer/laptop/tablet/mobile;
- Microsoft Teams video conferencing software.

A camera/webcam is optional should you wish to contribute to the video recording of the focus group.

9. **What if I don't have or have never used Microsoft Teams?**
Microsoft Teams software is free to download from https://www.microsoft.com/en-gb/microsoft-365/microsoft-teams/free

If you are a student at the University of Glasgow you can access Microsoft Teams by logging into Office 365 using your University email address and GUID password and choosing the "Teams" tile.
https://www.gla.ac.uk/myglasgow/anywhere/office365/teams/

Please consult Microsoft's privacy policies (https://www.microsoft.com/en-gb/trust-center/privacy?rtc=1) before deciding if you are happy to create an account with them.

Please also ensure you download the **free-to-use** version of the software as participation in this study should not result in any financial cost to you.

Details on using Microsoft Teams can be found at:
https://gla.sharepoint.com/sites/m365learningpathways/SitePages/Get-started-with-Microsoft-Teams.aspx

If you would like to get used to using the software or wish to test your internet connection and/or hardware (microphone and/or camera), you are welcome to email the lead researcher to arrange a practice meeting in advance.

10. **What are the possible disadvantages and risks of taking part?**
Other than the time the focus groups will take, there are no significant risks anticipated with taking part in this study.

11. **What are the possible benefits of taking part?**
There are no direct benefits to students for participating in the study. Taking part will only help contribute towards improving future assessment within dental education, as the information gathered as part of this study will help inform dental assessors and regulators on the use of longitudinal data as a method of assessment.

12. **Will my taking part in this study be kept confidential?**
All information which is collected about you, or responses that you provide, during the research will be protected by the research team and quotations used in subsequent publications will be anonymised. While the research team will make every effort to ensure that your participation is anonymised, it is important to note that due to the

nature of focus groups, we cannot make this assurance on behalf of the other participants.

You are encouraged to join the focus group from a quiet, secluded space to reduce any possible distractions and ensure privacy. If you wish to be visible in the video recording of the meeting, but are only able to broadcast from a space you consider to be not private, you can make use of Microsoft Team's "apply background" feature to block out what is behind you on camera. Details on how to use virtual backgrounds are provided in the Microsoft Team's user guide (https://support.microsoft.com/en-us/office/change-your-background-for-a-teams-meeting-f77a2381-443a-499d-825e-509a140f4780).

Please note that assurances on confidentiality will be strictly adhered to unless evidence of serious harm, or risk of serious harm, is uncovered. In such cases the University may be obliged to contact relevant statutory bodies/agencies.

13. **What will happen to the results of the research study?**

The results of the research study will be written up as a PhD thesis. They will also be published as part of academic journal articles and may be discussed at academic conferences and knowledge exchange events by the study team. No identifying personal data will be published.

Upon request, the researchers will be happy to share draft manuscripts of academic journal articles prior to submission.

14. **Who is organising and funding the research?**

This is a PhD research project for which the registration fees have been funded through the Dorothy Geddes Studentship awarded by the University of Glasgow Dental School.

NHS Education for Scotland (NES) have also provided funding to allow the researcher to conduct the study.

The evaluation is being led by researchers from the University of Glasgow (Jamie Dickie, Kurt Naudi, Andrea Sherriff and Michael McEwan).

15. **Who has reviewed the study?**

This study has been reviewed by the MVLS Ethics Committee at the University of Glasgow.

16. **Contact for Further Information**

For further information, please contact Jamie Dickie at Jamie.Dickie@glasgow.ac.uk or Kurt Naudi at Kurt.Naudi@glasgow.ac.uk.

Thank you for taking the time to take part in this study!

Participant information leaflet: Faculty, vocational trainer and NHS Education for Scotland representative(s))

University of Glasgow | College of Medical, Veterinary & Life Sciences

Dental Education Study Information (Plain Language Statement) for

Faculty/Vocational trainers/NHS Education for Scotland representative(s)

1. **Study title**

Longitudinal assessment of dental students

2. **Invitation**

You are being invited to take part in an online focus group as part of a research study. Before you decide it is important for you to understand why the research is being done and what it will involve. Please take time to read the following information carefully and discuss it with others if you wish. Ask us if there is anything that is not clear or if you would like more information. Take time to decide whether you wish to take part.

3. **What is the purpose of the study?**

Over the past few years, the University of Glasgow Dental School has made significant investments in implementing a longitudinal assessment system, known as LIFTUPP©, into their undergraduate programme. Now they would like to know the extent to which LIFTUPP© can contribute to the assessment of clinical competence.

To start gaining an insight into this, LIFTUPP© data will need to be compared against methods currently used to assess competence performance at undergraduate level as well as the clinical performance of recently qualified dentists (who will have certified as competent practitioners using these current assessment methods).

4. **Why have I been chosen?**

If you have been given this information sheet, it is because you have been identified as a key stakeholder regarding LIFTUPP©, i.e., you are influenced by and/or contribute to LIFTUPP© data. We would like to know a bit more about your thoughts of LIFTUPP© when it is compared against other dental education assessment methods, so you have been invited to take part in an online focus group to discuss your opinions with University of Glasgow dental faculty members, Dental Vocational Trainers and representations from NHS Education for Scotland (NES).

5. **Do I have to take part?**

Taking part is entirely voluntary. If you do decide to take part, you will be given this information sheet to keep and be asked to sign a consent form. If you decide to take part, you are still free to withdraw without giving a reason up until your online focus group discussion has been transcribed and anonymised (see below). After this point, we will no

longer be able to identify what was said by individual participants and, therefore, we won't be able remove your responses.

6. **What will happen to me if I take part?**

If you take part, you will be invited to attend an online focus group on [date] using Microsoft Teams video conferencing. You will be shown the results of some comparisons between LIFTUPP© data, traditional undergraduate assessments (such as Objective Structured Clinical Examinations (OSCEs), Multiple Short Answer (MSA) papers and Multiple-Choice Question (MCQ) papers) and postgraduate longitudinal clinical assessment data (known as LEPs). You will then be involved in a group discussion about your thoughts and opinions on how the results from these comparisons could be used to improve educational assessment within dentistry. This discussion will be led by the projects lead researcher (Jamie Dickie).

Audio from the resulting conversations will recorded using Microsoft Team's "record" feature - which can only be activated by the projects lead researcher. A backup copy of the audio will be recorded on a Dictaphone placed next to the speakers of the lead researcher's laptop computer.

A video recording of the meeting will also be made. However, you do not have to be filmed as part of the meeting should you not wish to. You will be able to join the meeting with audio input only and can activate video input at your own discretion.

The audio and video recordings made via Microsoft Teams then be automatically uploaded into the lead researcher's Microsoft Streams account –a cloud data storage platform which is part of the PhD researcher's password protected University of Glasgow Office 365 account. Using the transcription feature within Microsoft Streams, the audio from the focus group will be anonymously transcribed and anonymized codes will be used to refer to participants thereafter rather than personal information.

The backup audio from the Dictaphone will be immediately uploaded to a password protected folder on the University of Glasgow secure networked "J-Drive". Once the transcriptions of the discussion have been completed and checked. All audio and video recording files will be deleted.

Your responses will not affect undergraduate dental student grades or academic records in any way.

The focus group will last for approximately one-and-a-half hours (90 minutes).

7. **What do I have to do?**

You will have to take part in a group discussion about comparisons between LIFTUPP©, traditional undergraduate assessments and postgraduate assessment with your peers (as detailed above).

8. **What will I need to take part?**

You will need a unique meeting ID and password for gaining entry to the meeting. This will be emailed to you by the lead researcher if you consent to participate.

You will also need:

- A computer, laptop, tablet or mobile phone with a stable internet connection;
- A microphone which permit audio input into your computer/laptop/tablet/mobile;
- Microsoft Teams video conferencing software.

A camera/webcam is optional should you wish to contribute to the video recording of the focus group.

9. **What if I don't have or have never used Microsoft Teams?**
Microsoft Teams software is free to download from https://www.microsoft.com/en-gb/microsoft-365/microsoft-teams/free

If you are a student at the University of Glasgow you can access Microsoft Teams by logging into Office 365 using your University email address and GUID password and choosing the "Teams" tile.
https://www.gla.ac.uk/myglasgow/anywhere/office365/teams/

Please consult Microsoft's privacy policies (https://www.microsoft.com/en-gb/trust-center/privacy?rtc=1) before deciding if you are happy to create an account with them.

Please also ensure you download the **free-to-use** version of the software as participation in this study should not result in any financial cost to you.

Details on using Microsoft Teams can be found at:
https://gla.sharepoint.com/sites/m365learningpathways/SitePages/Get-started-with-Microsoft-Teams.aspx

If you would like to get used to using the software or wish to test your internet connection and/or hardware (microphone and/or camera), you are welcome to email the lead researcher to arrange a practice meeting in advance.

10. **What are the possible disadvantages and risks of taking part?**
Other than the time the focus groups will take, there are no significant risks anticipated with taking part in this study.

11. **What are the possible benefits of taking part?**
The information gathered as part of this study will help inform dental assessors and regulators on the use of longitudinal data as a method of assessment.

12. **Will my taking part in this study be kept confidential?**
All information which is collected about you, or responses that you provide, during the research will be protected by the research team and quotations used in subsequent publications will be anonymised. While the research team will make every effort to ensure that your participation is anonymised, it is important to note that due to the nature of focus groups, we cannot make this assurance on behalf of the other participants.

You are encouraged to join the focus group from a quiet, secluded space to reduce any possible distractions and ensure privacy. If you wish to be visible in the video recording of

the meeting but are only able to broadcast from a space you consider to be not private, you can make use of Microsoft Team's "apply background" feature to block out what is behind you on camera. Details on how to use virtual backgrounds are provided in the Microsoft Team's user guide (https://support.microsoft.com/en-us/office/change-your-background-for-a-teams-meeting-f77a2381-443a-499d-825e-509a140f4780).

Please note that assurances on confidentiality will be strictly adhered to unless evidence of serious harm, or risk of serious harm, is uncovered. In such cases the University may be obliged to contact relevant statutory bodies/agencies.

13. **What will happen to the results of the research study?**

The results of the research study will be written up as part of a PhD thesis. They will also be published as part of academic journal articles and may be discussed at academic conferences and knowledge exchange events by the study team. No identifying personal data will be published.

Upon request, the researchers will be happy to share draft manuscripts of academic journal articles prior to submission.

14. **Who is organising and funding the research?**

This is a PhD research project for which the registration fees have been funded through the Dorothy Geddes Studentship awarded by the University of Glasgow Dental School.

NHS Education for Scotland (NES) have also provided funding to allow the researcher to conduct the study.

The evaluation is being led by researchers from the University of Glasgow (Jamie Dickie, Kurt Naudi, Andrea Sherriff and Michael McEwan).

15. **Who has reviewed the study?**

This study has been reviewed by the MVLS Ethics Committee at the University of Glasgow.

16. **Contact for Further Information**

For further information, please contact Jamie Dickie at Jamie.Dickie@glasgow.ac.uk. or Kurt Naudi at Kurt.Naudi@glasgow.ac.uk.

Thank you for taking the time to take part in this study!

iii) Privacy notice

Privacy Notice for participation in PhD research project – Longitudinal assessment: Building a validity argument

Your Personal Data

The University of Glasgow *will be what's known as the 'Data Controller' of your personal data processed in relation to participation in a PhD research project. This privacy notice will explain how The University of Glasgow will process your personal data.*

Why we need it

We are inviting you to take part in an online focus group to discuss how competence assessment within dental education could be enhanced. We are collecting some basic personal data including name, email address and (where applicable) job title or year of study in the Bachelor of Dental Surgery (BDS) curriculum. Your name and email address will not be shared with any other party. Your job title or year or year of study will help provide a demographic overview of the sample.

We are also collecting your voice and image through audio and video recording as part of a focus group discussion. The focus group will be conducted online using Microsoft Teams video conferencing software. . Your use of a webcam, and therefore our recording of the video will be entirely optional, however your participation in the focus group will imply your consent to audio recording. Therefore, your voice will be collected during the recording and, if you choose to participate in the video recording, your appearance will be recorded as well.

A backup audio recording of the focus group will be made via an audio recording device placed next to the speakers of the focus group moderator/PhD researcher's computer. This additional recording will be made in case there are any technical difficulties with Microsoft Teams.

We will only collect data that we need in order to provide and oversee this service to you.

Legal basis for processing your data

We must have a legal basis for processing all personal data. In this instance, the legal basis is consent.

A consent clause is provided at the end of this privacy notice.

In addition, if special categories/sensitive personal information are being processed, an additional basis needs to be specified – please contact the DP&FOI Office to discuss.

What we do with it and who we share it with

- *All the personal data you submit is processed by staff at the University of Glasgow in the United Kingdom.*
- *Video and audio recordings of the focus group will be obtained via Microsoft Teams and will be automatically uploaded to Microsoft Stream – an online cloud data storage space. Both Microsoft Teams and Streams are institutional approved by the University of Glasgow.*
- *Within Microsoft Stream, the video and audio recording of the focus group will be used to produce a transcript of the discussions that take place. During the transcription process, personal data (such as name) will not be used. Instead, participants will only be referred to by a numerical code. This will anonymise you within the transcript.*
- *Upon completion and verification of the transcript, the video and audio recordings will be deleted from Microsoft Stream.*
- *While the research team will make every effort to ensure that your participation is anonymised, it is important to note that due to the nature of focus groups, we cannot make this assurance on behalf of the other participants.*
- *The transcripts produced will be used by the PhD researcher (Jamie Dickie) to identify key themes/topics that were discussed in the focus group, which will form part of the results of the research project. These results will form part of a thesis which will be submitted to the University of Glasgow.*

How long do we keep it for

*The transcription containing your data will be retained by the University for **10-years**. After this time, the transcription will be securely deleted.*

What are your rights?*

You can request access to the information we process about you at any time. If at any point you believe that the information we process relating to you is incorrect, you can request to see this information and may in some instances request to have it restricted, corrected or, erased. You may also have the right to object to the processing of video and audio data and the right to data portability up until your focus group discussion has been transcribed and anonymised. After this point, we will no longer be able to identify what was said by individual participants and, therefore, we won't be able remove your responses.

Where we have relied upon your consent to process your data, you also have the right to withdraw your consent up until your focus group discussion has been transcribed and anonymised.

If you wish to exercise any of these rights, please submit your request via the webform or contact dp@gla.ac.uk.

*Please note that the ability to exercise these rights will vary and depend on the legal basis on which the processing is being carried out.

Complaints

If you wish to raise a complaint on how we have handled your personal data, you can contact the University Data Protection Officer who will investigate the matter.

Our Data Protection Officer can be contacted at dataprotectionofficer@glasgow.ac.uk

If you are not satisfied with our response or believe we are not processing your personal data in accordance with the law, you can complain to the Information Commissioner's Office (ICO) https://ico.org.uk/

☐ I consent to the University processing my personal data for the purposes detailed above.

I have read and understand how my personal data will be used.

Signed:
……

Date:
……

iv) Written consent form - Undergraduate students, vocational dental practitioners and recently qualified dentists

University of Glasgow | Dental School

Consent Form

Undergraduate Student/Vocational Dental Practitioner/Recently qualified dentist

Title of Project: Longitudinal assessment of dental students

Name of Researchers: Jamie Dickie, Kurt Naudi, Andrea Sherriff, Michael McEwan

Please initial box

1. I confirm that I have read and understand the Plain Language Statement for the above study and have had the opportunity to ask questions. ☐
2. I understand that my participation is voluntary and that I am free to withdraw at any time, without giving any reason. ☐
3. I consent to participating an *online focus group*, the audio of which will be recorded and transcribed. ☐
4. I understand the *focus group* transcription will be anonymised and analysed by the researchers. ☐
5. I understand that participation in video recording of the *online focus group* is entirely optional and within my own control. ☐
6. I understand the outcomes of *online focus group* discussions will not contribute to my academic attainment or personal records. ☐
7. I agree to the use of anonymised quotations in publications. ☐
8. I understand that participation or non-participation in the research will have no effect on academic or personal records. ☐

By signing below, I agree to take part in the above study.

Name of Participant	Date	Signature
Name of Person taking consent (if different from researcher)	Date	Signature
Researcher	Date	Signature

v) Written consent form - Faculty, vocational trainer and NHS Education for Scotland representative(s)

University of Glasgow | Dental School

Consent Form

Faculty/Vocational Trainer/NHS Education for Scotland Representative

Title of Project: Longitudinal assessment of dental students

Name of Researchers: Jamie Dickie, Kurt Naudi, Andrea Sherriff, Michael McEwan

Please initial box

1. I confirm that I have read and understand the Plain Language Statement for the above study and have had the opportunity to ask questions. ☐
2. I understand that my participation is voluntary and that I am free to withdraw at any time, without giving any reason. ☐
3. I consent to participating in an *online focus group*, the audio of which will be recorded and transcribed. ☐
4. I understand the *focus group* transcription will be anonymised and analysed by the researchers. ☐
5. I understand that participation in video recording of the *online focus group* is entirely optional and within my own control. ☐
6. I understand that discussions arising during the *focus group* will have no influence on undergraduate student attainment or personal records. ☐
7. I agree to the use of anonymised quotations in publications. ☐

By signing below I agree to take part in the above study.

Name of Participant	Date	Signature
Name of Person taking consent (if different from researcher)	Date	Signature
Researcher	Date	Signature

Appendix 5 – Focus group topic guide

Focus group topic guide

Research question

According to key stakeholders in dental education, how might the findings of research questions 2a, 2b, 2c and 2d be used to enhance competence assessment within dentistry?

Sampling

The sample will be drawn from key stakeholders within dental assessment.

Recruited participants will be arranged into two focus groups.

Key stakeholders who will be invited for participation will include:

Focus group 1

i. BDS year representatives (BDS years 1, 2, 3, 4 and 5).

ii. Undergraduate dental students (across all BDS years - 1, 2, 3, 4 and 5).

iii. Recent and current vocational dental practitioners (VDPs) (i.e., recently graduated dentists who completed vocational training in either 2017-18, 2018-19, 2019-20 or are currently undertaking vocational training).

Focus group 2

1. Director and assistant director of Dental Education.
2. Head of School.
3. Glasgow Dental School's year-group coordinators.
4. Clinical teaching staff/faculty at Glasgow Dental School (including Teaching Leads for specific dental subjects (e.g., Restorative Dentistry, Paediatric Dentistry, Oral Surgery and Oral Medicine) and Lead for Clinical Dentistry).
5. NHS Education for Scotland staff - a) Dean of Postgraduate Dental Education, b) Associate Dean of Postgraduate Dental Education and c) Associate Dean of Vocational Training.
6. Dental vocational training trainers.

Equipment and accoutrements (face to face)

- Audio recording equipment x 2
- Place cards with names
- Note pads and pens
- Snack foods and drinks

Equipment and accoutrements (online) – may be required instead of face to face focus groups due to COVID-19 pandemic.

Moderator/PhD researcher:

- Laptop with webcam
- Microphone
- Microsoft Teams online video conferencing software
- Microsoft Streams software
- Audio recording equipment x 1
- Note pad and pens

Participants

- PC/laptop/tablet/mobile phone
- Microphone
- Webcam (optional)
- Microsoft Teams online video conferencing software

NOTE: Participants will be informed on the equipment they require to join the focus groups using information leaflets.

Moderation

- Facilitator: Jamie Dickie (PhD researcher)
- Level of moderation: medium

Recording (if online focus groups used)

- Audio and video recordings made via Microsoft Teams online video conferencing software (part of the PhD researcher's University of Glasgow Office365 account).
- Audio and video files will be automatically saved to Microsoft Stream (also part of the PhD researcher's University of Glasgow Office365 account).

OUTLINE OF FOCUS GROUPS

Introduction (5 mins)

- Provide background details on why the study being conducted - refer to General Dental Council's (GDC's) domains of clinical competence and their definition of the "safe beginner"
- State the purpose of the focus group within the study
- Provide details on format of focus groups (presentation of results followed by discussions)
- Ground rules: One person talking at a time, no side conversations, everyone participating/no domination of the discussions, free to make notes before responses.
- Reiterate that discussions will be recorded and that transcriptions will be anonymised
- Recheck consent

Opening questions (2-3 mins)

Participant introductions (including explanation of roles) – Tell us who you are, what your current role within dental education is?

Introduction question (5 mins)

- What methods of assessment within dental education and training are you familiar with and have you participated in any of them either as an assessor, candidate or both?

Presentation 1: ***Longitudinal assessment data (5 mins)***

- Powerpoint - Briefly discuss the method(s) used
- Have hard copies of charts and tables available/display charts and tables via Screenshare

Topic 1 - Longitudinal assessment data (15 mins)

Transition question

What do you think these results show in terms of using longitudinal data as a method of assessment for undergraduate dental students?

Refocus questions

- What can be determined about student development from this form of assessment?
- Is this information useful? How is useful/how is it not useful?
- What are the apparent advantages of this assessment method?
- Are there any apparent drawbacks to this assessment method?
- Does longitudinal assessment gather too much data? Should it gather more data?

Presentation 2: ***Longitudinal assessment data vs UG exams (5 mins)***

- Powerpoint - Briefly discuss the method(s) used
- Have hard copies of charts and tables available/display charts and tables via Screenshare

Topic 2 – Longitudinal assessment data vs UG exams (10 mins)

Transition question

What do you think these results show in terms of how students' longitudinal clinical performance relates to their performance in the professional undergraduate examinations?

Refocus questions

- Are there any relationships between data from each of these assessment methods?
- Is there a strong relationship? A weak one? None?
- Is this information valuable? What does it tell us? What doesn't it tell us?

Presentation 3: ***Longitudinal assessment data vs postgraduate performance (5 mins)***

- Powerpoint - Briefly discuss the method(s) used
- Have hard copies of charts and tables available/display charts and tables via Screenshare

Topic 3 – Longitudinal assessment data vs postgraduate performance (10 mins)

Transition question

What do you think these results show in terms of how students perform in undergraduate clinics and how they perform in postgraduate vocational training?

Refocus questions

- Are there any relationships between data from each of these assessment methods?
- Is there a strong relationship? A weak one? None?
- Is this information valuable? What does it tell us? What doesn't it tell us?

Presentation 4: ***UG exams vs postgraduate performance (5 mins)***

- Powerpoint - Briefly discuss the method(s) used
- Have additional copies of charts and tables available

Topic 4 – UG exams vs postgraduate performance (10 mins)

Transition question

What do you think these results show in terms of how students perform in their undergraduate professional examinations and how they perform clinically in postgraduate vocational training?

Refocus questions

- Are there any relationships between data from each of these assessment methods?
- Is there a strong relationship? A weak one? None?
- Is this information valuable? What does it tell us? What doesn't it tell us?

Topic 5 – Enhancing assessment in dental education (30 – 40 mins)

Key questions

One of the things we are especially interested in is if longitudinal assessment could be used to assess competence - what are your thoughts on this having seen the results of this study?

How could the results of this study be used to enhance assessment within dentistry?

Refocus questions

- Is competence assessed adequately within dental education?
- Are there any apparent advantages and disadvantages for each assessment method?
- Do some assessment methods appear to be better/more valuable than others? If so, in what way?
- Should some methods no longer be used or modified? If so, which ones? Which ones should be used? Should they be formative or summative assessments? Should/could they all be used together?
- Which assessment method(s) could/should be used to assess undergraduate dental student clinical competence?
- Can the results of comparisons between each of these assessment methods guide us in improving assessment of dental undergraduates?
- Do participants think any of the assessment methods used (either individually or combined) for undergraduates prove which dental students are ready to graduate and practice on the public?
- Are the findings of this study relevant?
- Are the findings of this study valuable?
- What should be done with the information generated by this study?
- Should any changes be made on how dental student competence is assessed?
- How could the results of this study be used to enhance assessment within dentistry?
    - -If so – what should we use them for? What action(s) should we take?
    - -If not – what should out next step/approach be to enhance assessment?

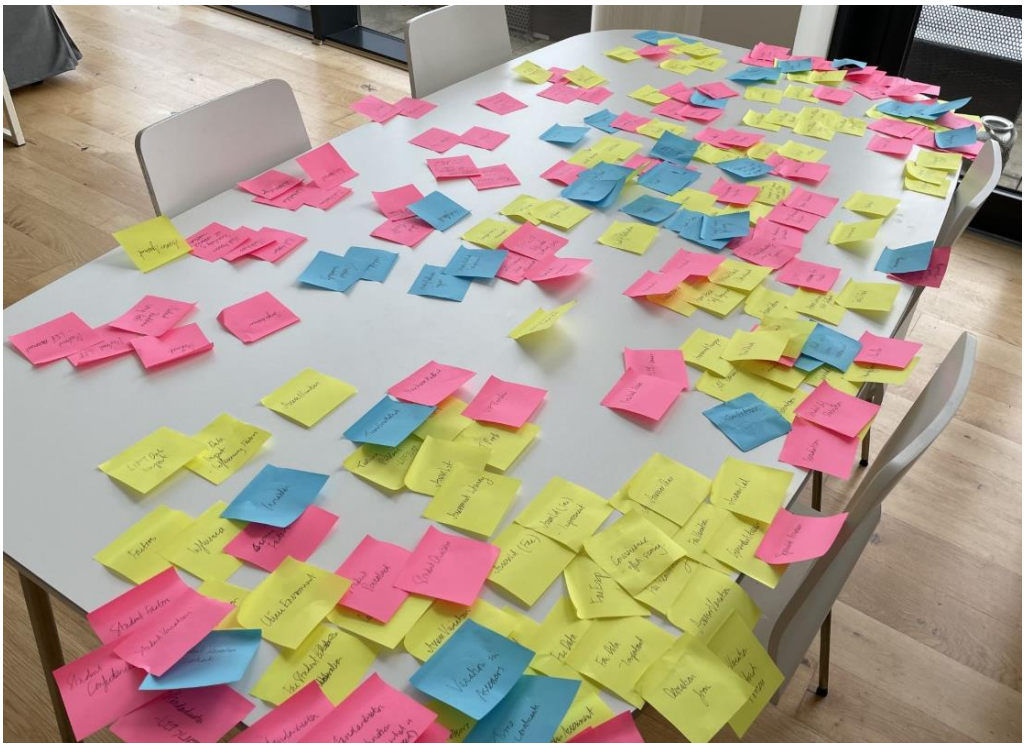
Closure (10mins)

Summarise discussions

Final question: Is there anything that we should have talked about but didn't?

Thank participants

Appendix 6 – Focus group transcripts thematic analysis: Identification of categories and themes



Appendix 7 – Approvals and data management

i. Ethical approval application – University of Glasgow College of Medical, Veterinary & Life Sciences (Reference number: 200170146)

University of Glasgow | College of Medical, Veterinary & Life Sciences

College of Medical, Veterinary & Life Sciences Ethics Committee for Non-Clinical Research Involving Human Subjects

APPLICATION FORM FOR ETHICAL APPROVAL

NOTES:
THIS APPLICATION FORM SHOULD BE TYPED NOT HAND WRITTEN.

ALL QUESTIONS MUST BE ANSWERED. "NOT APPLICABLE" IS A SATISFACTORY ANSWER WHERE APPROPRIATE.

Project Title: Can longitudinal performance data from the BDS programme determine performance in vocational training?

Has this application been previously submitted to this or any other ethics committee? No

If 'Yes', please state the title and reference number.

Is this project from a commercial source, or funded by a research grant of any kind?
No

If 'Yes', has it been referred to Research Support Office? NA
Has it been allocated a project Number? NA
Give details, and ensure that this is stated on the Informed Consent Form.

Insurance Coverage and Restrictions:
****Please Note: The Insurance restrictions set out below relate to research of a clinical nature. Non clinical research is not subject to restriction and no additional insurance is required****

The University insurance cover is restricted under specific circumstances, including, but not limited to the following -

- **work conducted outside of the European Union.**
- **work involving the use of research subjects outside Great Britain, Northern Ireland, the Channel Islands or the Isle of Man.**
- **the use of hazardous materials.**
- **number of participants in excess of 5000.**
- **work involving research subjects known to be pregnant at the time of the project.**

All such projects must be referred to Research Support Office and coverage confirmed before ethical approval is sought. Please contact Dr Debra Stuart in the University's Research Governance Office: debra.stuart@glasgow.ac.uk

Please tick here if this project has been referred to Research Support Office to confirm adequate insurance coverage. ☐

Date of submission: ***25th April 2018***

Name of all person(s) submitting research proposal:
Jamie Dickie, Kurt Naudi, Andrea Sherriff, Michael McEwan

Position(s) held:
Jamie Dickie - Clinical Lecturer/Honorary Specialty Registrar – Restorative Dentistry
Kurt Naudi - **Clinical Senior University Teacher/Honorary Consultant – Oral Surgery**
Andrea Sherriff – Senior Lecturer – Statistics
Michael McEwan - Senior Academic and Digital Development Adviser (University Lecturer) - Academic and Digital Development, Academic Services

School/Group/Institute/Centre:
University of Glasgow Dental School

Address for correspondence relating to this submission:
Room L7, level 9, Glasgow Dental Hospital and School, 378 Sauchiehall Street, Glasgow, G2 3JZ

Email address:
Jamie.Dickie@glasgow.ac.uk;

Name of Principal Researcher (if different from above, e.g., Student's Supervisor):

Position held:

Undergraduate student project:
No If 'Yes', please state degree being undertaken:

Postgraduate student project:
Yes If 'Yes', please state degree being undertaken:

Doctor of Philosophy (PhD)

1. Describe the purposes of the research proposed. Please include the background and scientific justification for the research. Why is this an area of importance?

In December 2014, the Scottish Oral Health Research Collaboration (SOHRC) completed a priority setting exercise to establish a basis for a focussed dental education research strategy within and between the Scottish Dental Schools and NHS Boards (Ajjawi et al., 2017). The results of the exercise were presented at the first SOHRC conference in February 2015, where the top three priorities were identified as:

1. The role of assessments in identifying competence;
2. Ensuring that the undergraduate curriculum prepares for practice;
3. Promotion of teamwork within the dental team.

In accordance with the first of these priorities, the lead researcher conducted a dissertation study that was submitted to the University of Glasgow in part fulfilment of the requirements for the degree of Master of Education (MEd). This study aimed to investigate if ***longitudinal data*** on undergraduate dental students' clinical performance could be used to measure their competence in performing an operative procedure. Dawson et al. (2017) and Albino et al. (2008) have previously argued that longitudinal performance data is one of the strongest means of assessing students in medical-based subjects. They both suggested that longitudinal evaluation across multiple sessions of patient care may provide a richer data source for evaluation of professionalism, personal attributes and clinical application compared to single encounter evaluations (e.g., clinical competency examinations), which may only record "*best day*" or "*worst day*" performances not representative of a student's true capabilities.

Longitudinal clinical performance data were obtained from an electronic development and assessment system, known as LIFTUPP©, which has recently been adopted by a number of UK Dental Schools. Thirteen dental students' LIFTUPP© data were analysed and formatted to establish and illustrate patterns of development and then compared to outcomes obtained from a standalone competence test and faculty subjective opinion. Qualitative and quantitative evaluations were made to determine if there was any association between LIFTUPP© data patterns and the results obtained from the more traditional assessments.

Overall, the study was unable to demonstrate a statistical association between the data sets. However, LIFTUPP© appeared to offer a richer collection of data on student development compared to the standalone competence test and faculty subjective opinion, both of which yielded a number of inconsistencies in terms of assessment. It may be that LIFTUPP© data could significantly contribute to summative decisions on student progression by a panel of academic staff, but the validity of using LIFTUPP© for this purpose could not be determined in this study due to scale and timeframe constraints associated with a MEd dissertation.

Furthermore, since the more traditional assessment methods were inconsistent, it could be argued that they are not particularly valid assessments as they were unable to definitively

determine which students could competently place a direct restoration. As a result, no meaningful comparisons could be made regarding the LIFTUPP© system's validity even if the sample size were to be greatly increased. Therefore, it was suggested that an investigation on the determination of dental student competence using longitudinal data from LIFTUPP© would require a number of alternative approaches.

Firstly, LIFTUPP©'s ability to conveniently measure a range of skills and procedures that coincide the GDC's four domains of competent practice (Clinical, Communication, Professionalism and Management & Leadership) will need to be established. These domains are outlined in the GDC's Preparing for Practice (PfP) document (2015). The MEd study was only able to focus on the assessment of one particular clinical procedure due to time and study size constraints.

Secondly, since there are currently no robust evaluations of longitudinal assessments using objective outcome measures, their association with current undergraduate and postgraduate assessment methods should be explored. Comparison with undergraduate modes of assessment, such as Objective Structured Clinical Examinations (OSCEs) and written examinations (multiple short answer (MSA) and multiple choice question (MCQ) papers), would investigate if longitudinal assessment of dental students is beneficial, whereas comparison with postgraduate assessment would be of value since the aim of any assessment at the point of graduation must be to predict future performance in practice. Any associations between current undergraduate and postgraduate assessment methods should also be investigated as they will serve as a baseline against which new assessment methods (in this case longitudinal assessment) can be compared.

Comparing longitudinal undergraduate and postgraduate performance data will also serve as an opportunity to compare LIFTUPP© against an assessment method that has had its validity previously investigated. Studies by Prescott-Clements et al. (2001; 2001; 2003; 2008) have argued that there is a degree of face validity associated with Longitudinal Evaluations of Performance (LEP) – a continuous assessment method used in the postgraduate vocational training (VT) schemes by the Scottish dental deaneries. LEPs are completed by Vocational Dental Practitioners (VDPs') assigned VT Trainers to assess clinical judgement, technical ability, management and leadership skills, professionalism and communication skills (which correlate with the GDC's four domains of clinical competence). Each VDP's LEPs are then submitted to a National Review Panel to determine if they have satisfactory completed VT and are safe and prepared to enter the NHS dental workforce as independent practitioners.

These initial lines of enquiry have been suggested by the lead researcher. However, to ensure that this project provides meaningful contribution to the assessment of competence, key stakeholders within dental assessment are to be consulted in the early stages to refine the proposed research questions (see below). Once the initial study findings are known, further discussion with key stakeholders will help determine if longitudinal assessment of dental students is beneficial and generate future lines of enquiry.

With these considerations, the proposed study research questions are as follows:

Research questions (RQs)

Within the GDC's framework for competent clinical practice:

1. What are the main patterns of longitudinal assessment over time within a year and across years?
2. What is the association between undergraduate longitudinal methods of assessment and standalone methods?
3. What is the association between undergraduate longitudinal methods of assessment and postgraduate assessment?
4. What is the association between the undergraduate standalone methods of assessment and postgraduate assessment?

According to key stakeholders within dental assessment:

5. Can the findings of RQs 1, 2 and 3 be used to improve assessment?

2. Describe the design of the study and methods to be used. Include sample size and the calculation used to determine this. Statistical advice should be obtained if in doubt.

This is a mixed methods study with both quantitative and qualitative methods applied as appropriate.

Quantitative data will be sourced from the LIFTUPP© and LEP systems, which both continually collect information on undergraduate student and VDP performance over the course of undergraduate training and VT respectively. Additional undergraduate assessment data from OSCE, MSA and MCQ examinations will also be collected.

Qualitative data will be generated from discussions with key stakeholders in dental assessment. Key stakeholders will also have a role in refining the research questions during the early stages of the study (see below).

Initial consultation with key stakeholders

The University of Glasgow Dental School has made (and continues to make) significant investments in implementing the LIFTUPP© system into their undergraduate programme. However, since LIFTUPP© has only been in use since the 2014/15 academic term, the extent to which these data provide any meaningful contribution to the assessment off competence has yet to be investigated. Key stakeholders will be involved in the early stages of refining the research questions through focus groups and/or semi-structured interviews to ensure their needs are being met by the project.

The list of key stakeholders will include undergraduate dental students, as well as Glasgow Dental School's year-group co-ordinators, Director of Dental Education, Head of School, Teaching Leads for specific dental subjects (e.g., Restorative Dentistry, Paediatric Dentistry, Oral Surgery and Oral Medicine) and Lead for Clinical Dentistry. Once key stakeholders have been recruited, focus groups and/or semi-structured interviews will be used to determine if research questions 1, 2, 3 and 4 are to be modified or if additional questions are required for the study.

Once the research questions have been agreed, any required modification to methodology will be considered. Assuming that no changes to methodology are necessary in light of modified research questions, the following study design will be adopted.

Methods for Research Questions 1, 2, 3 and 4 (Quantitative)

Design and setting

Three longitudinal cohorts from existing assessment data will be created, each comprising of individuals who graduated Bachelor of Dental Surgery (BDS) from the University of Glasgow between June 2017 and June 2019 and subsequently complete their postgraduate VT in a Scottish training scheme. Data on individual clinical performance at both undergraduate and postgraduate level will be collected as well as results from undergraduate professional examinations throughout the five-year curriculum.

Sampling

The sample for the quantitative aspect of the study will be drawn from University of Glasgow undergraduate dental students and VDPs. No power calculation has been used to define a required sample size, as the number of participants will be fixed by those students who graduate from Glasgow from 2017 onwards and complete their VT in Scotland.

Details for each cohort are as follows:

Cohort 1

- *Includes those who graduated in 2017 and will complete VT in 2018.*
- *n= 79*

Cohort 2

- *Includes those who will graduate in 2017 and complete VT in 2018.*
- *n= 94*

Cohort 3

- *Includes those who graduated in 2019 and complete VT in 2020.*
- *n= 72*

*Estimated total participants – 245**

*This total may decrease should some students fail to successfully complete the BDS course or opt not to enlist in a Scottish VT post.

Data sources

Data will be collected from a number of sources, which are detailed below:

1. LIFTUPP© - An electronic longitudinal assessment tool which forms a database on clinical activity and performance throughout undergraduate dental training. Data controller – The University of Glasgow.
2. LEPs – Longitudinal work-based assessments used in Scottish VT schemes. Data controller – NHS Education for Scotland (NES).
3. Undergraduate examination results – The results of professional assessments undertaken by students and recorded as part of their academic records. This will include data from the following examination types over the 5-year curriculum:
    - OSCEs – 4 exams (used for assessment at the end of 1st, 2nd, 3rd and 5th year);
    - Multiple Short Answer (MSA) – 4 exams (used for assessment at the end of 1st, 2nd, 3rd and 4th year);
    - Multiple Choice Question (MCQ) – 3 exams (used for assessment at the end of 1st, 2nd and 3rd year).

    Data controller – The University of Glasgow

Participant demographics will also be available from each of these three data sources.

Data Linkage **(see section 15)**
OSCE, MSA, MCQ, LIFTUPP© and LEP data are to be linked and pseudonymised (using name, gender and date of birth) by a third party (member of University academic staff not involved in the research) and stored on a secure networked drive at the University of Glasgow prior to analyses by the lead researcher. A data sharing agreement between the University of Glasgow and NES is currently being arranged to facilitate the transfer of LEP data via a secure networked NHS.NET email account (see section 9).

Statistical modelling and analysis
All analyses will be undertaken on the secure networked drive (University of Glasgow) on pseudonymised data. Only the student and supervisors will have access to these data. No-one from the study team will have access to the directory where the keyed data are stored. All data tables and output will be checked for potentially disclosive information (e.g., small cell sizes), and only aggregated data will be presented. A detailed Statistical Analysis Plan (SAP) will be developed by the PhD student under supervision of supervisor as part of their research training.

General analysis
To ensure integrity, all data will be cleansed and subjected to quality assurance. This process will involve producing frequency tabulations and histograms to check ranges and observe any unusual and typographical errors. Cross tabulations and scatter plots will be used to detect logic errors and variables (such as gender and date of birth) will be used to check for linkage errors. Summary statistics will be used to describe all variables – i.e., means, standard deviations for continuous (symmetrically distributed) data, medians, modes, Q1, Q3, minimum and maximum for continuous (non-symmetrically distributed) or discrete/data, frequencies/proportions for categorical data. Confidence intervals will be calculated around all estimates, using bootstrap methods if/where appropriate.
NOTE: The variables and analyses for each research question are to be detailed following further investigation and processing of data source samples.
Early discussions have focused on using methods for modelling longitudinal data to determine and analyse trajectories of assessment data, such as Latent Class Models. Some examples of methods that are likely to be used include general linear models and logistic regression. Stata (Statcorp) will be used for all analyses.

Methods for Research Question 5 (Qualitative)

Design and setting

Focus groups and/or semi-structured interviews will be used to record the perceptions and opinions of previously identified key stakeholders. Protocols will be drawn up in advance to ensure that discussions are structured and generate data that determine how the findings of research questions 1, 2 and 3 can be used to inform dental education and assessment. Discussions will be audibly recorded and transcribed (see section 15).

Amendment (October 2020): Due to the COVID-19 outbreak, it is currently not possible to stage focus groups in person. As a contingency, online focus groups will be conducting using Microsoft Teams video conferencing software.

Sampling

The sample for the qualitative aspect of the study will be drawn from key stakeholders within dental assessment. Key stakeholders who will be invited for participation will include undergraduate dental students as well as Glasgow Dental School's Director of Dental Education, Head of School, Year-Group Co-Ordinators, Teaching Leads for specific dental subjects (e.g., Restorative Dentistry, Paediatric Dentistry, Oral Surgery and Oral Medicine) and Lead for Clinical Dentistry.

NOTE: The list of key stakeholders was previously expanded to include BDS year representatives, recent and current vocational dental practitioners, clinical teaching staff/faculty at Glasgow Dental School, dental vocational training trainers and NHS Education for Scotland staff (such as the Dean of Postgraduate Dental Education, the Associate Dean of Postgraduate Dental Education and/or the Associate Dean of Vocational Training).

Amendment (October 2020): Equivalent faculty from other UK dental institutions will also be invited to participate.

Data sources

Audio recordings and transcripts of focus group/semi-structured interview conversations with key stakeholders in dental assessment.

Amendment (April 2020): Audio and video recordings and transcripts will be produced using Microsoft Teams video conferencing software. These recordings will be automatically uploaded to Microsoft Streams to allow transcriptions of the audio. This will also be produced using Microsoft Stream's "transcribe" feature.

Both Microsoft Teams and Microsoft Stream have been institutional approved by the University of Glasgow and are part of the university's Office 365 package.

Recruitment and consent

Key stakeholders who have been identified as potential participants will be contacted via email by the PhD student (Jamie Dickie). They will be provided with a participant information leaflet outlining the project and invited to voluntarily participate in the study (see section 4).

Further details on the participant information leaflet and the consent process are provided in sections 10 and 13 respectively.

Amendment (October 2020): Different participant information leaflets and consent forms will be distributed since online focus groups are required. Copies of these participant information leaflets and consent forms are provided alongside this document.

Participants will be required to have (or obtain) access to Microsoft Teams, as well as hardware supported by Microsoft (computer, laptop, tablet or mobile phone) and a microphone. Those who wish to contribute to the video recording of the meeting will also require a camera/webcam.

Many key stakeholders identified for participation (such as students and faculty members) will already have access to Microsoft Teams via the University of Glasgow. Other volunteers may need to download the software prior to the focus groups.

If participants need to acquire the software, they will be invited to sign up for and download the "free to use" version of Microsoft Teams to avoid any financial cost (https://www.microsoft.com/en-gb/microsoft-365/microsoft-teams/free).

Those unfamiliar with using Microsoft Teams will be able to arrange practice meetings with the PhD student prior to the focus groups if they believe it would help them become accustomed. This will also allow them to test their connection and microphone (+/- camera) in advance.

Participants will be made aware of the opportunity to familiarise themselves with using Microsoft Teams (if necessary) in the participant information leaflets – copies of which provided alongside this document.

Analysis

Framework analysis will be used in NVivo (QSR International) to analyse key themes. All data will be anonymised and stored on the University of Glasgow's secure networked J-Drive.

3. Describe the research procedures as they affect the research subject and any other parties involved. It should be clear exactly (i) what will happen to the research participant, (ii) how many times and (iii) in what order.

The quantitative research procedures should not have any further effect on the participants. LIFTUPP© and LEP data are already being collected as part of ongoing formative assessment in the Glasgow undergraduate BDS curriculum and for satisfactory completion of the VT scheme respectively.
The University of Glasgow owns and stores undergraduate student LIFTUPP© data which gradually builds as they progress through dental school. LIFTUPP© data are currently used for formative assessment but are also being considered for future summative assessment.

LEP data will need to be transferred by NES - the data controller for LEPs. A data sharing agreement is currently being arranged between the University of Glasgow and NES to facilitate this transfer (see section 9). In principle, this agreement will certify that LEP data for the research participants will be made available to the researchers at the end of three VT periods, i.e., August 2018 for the 2017/18 cohort, August 2019 for the 2019/20 cohort and August 2020/21 cohort.

For the qualitative aspect of the study, those who wish to participate will be invited to a focus group/semi-structured interview where they will be encouraged to engage in conversation on the results from research questions 1, 2, 3 and 4 and how they may be used to improve dental assessment. The audio of their answers will be recorded on a Dictaphone which will then be transcribed. Each key stakeholder participant will only need to attend one focus group session. Undergraduate students will participate in separate focus groups from staff (see section 10).

Overall, there is no additional assessment load imposed on undergraduate student and VDP participants by the quantitative component of this project. Instead, it will make further use of existing and prospectively generated data. Those who choose to participate in the qualitative component will need to volunteer to commit a few hours of their own time.

Amendment (October 2020): Audio and video recordings of online focus group discussions will be producing using the "record" function within Microsoft Teams video conferencing software. An additional audio recording will also be made using a Dictaphone placed next to the speakers of the PhD student's laptop computer. This additional recording will serve as a backup should any technical difficulties occur with Microsoft Teams. The PhD student will moderate the focus groups in non-shared private accommodation to ensure privacy and confidentiality - i.e., no-one other than the PhD student will be able to listen to the audio output needed to make this backup recording.

The audio and video recordings made via Microsoft Teams will be stored on Microsoft Steams – which is part of the PhD student's password protected University of Glasgow Office 365 account. Upon completion of each focus group, the backup audio from the Dictaphone will be immediately be uploaded into a folder

on the University of Glasgow's secure networked J-Drive the access to which will be password protected. The audio on the Dictaphone will then be deleted.

4. How will potential participants in the study be (i) identified, (ii) approached and (iii) recruited? Give details for cases and controls separately if appropriate.

Please note for the following section that all quantitative data is secondary data gathered for existing undergraduate and postgraduate assessment and will be pseudonymised.

Participants for the quantitative component will be identified through a list of inclusion criteria. Those who are eligible to be included in the study will need to have had all undergraduate clinical activity recorded and assessed via the LIFTUPP© system. In Glasgow, LIFTUPP© has been progressively "rolled out" since the 2014/15 academic term, meaning the graduating class of 2017 was the first student cohort to meet this criterion.

Another requirement is that, once progressing to VT, participants must have enlisted in a Scottish training scheme. Whilst other UK countries offer VT posts, they do not use the LEP system as part of their postgraduate assessment process. England and Wales use "A Dental Evaluation of Performance Tool" (ADEPT), whereas Northern Ireland use a series of work-based assessments known as Direct Observation of Procedural Skills (DOPS) and Case-based Discussions (CBD). The validity of these assessment methods within dental education has yet to be published, whereas LEPs have been scrutinised in a previous study (Prescott-Clements et al.; 2008) and were suggested to have a degree of validity.

Undergraduate and postgraduate assessment data (i.e., MSCE/MSA/MCQ, LIFTUPP© and LEP) from those who met these criteria will be pseudonymised after linkage by a third party and will be treated in accordance with the Data Security Protocol (COH) (see section 15).

As discussed in section 2, key stakeholders who will be invited for participation in the qualitative component will include undergraduate dental students as well as Glasgow Dental School's Director of Dental Education, Head of School, Year-Group Co-Ordinators, Teaching Leads for specific dental subjects (e.g., Restorative Dentistry, Paediatric Dentistry, Oral Surgery and Oral Medicine) and Lead for Clinical Dentistry. This list may have further additions made if the researchers identify groups/individuals who are influenced by and/or contribute to longitudinal assessment data whose input would be worth including in the study.

Key stakeholders will be contacted via email by the PhD student (Jamie Dickie). They will be provided with a participant information leaflet outlining the project and invited to voluntarily participate in the study. Those who wish to participate will be asked to reply to the PhD student's email. They will then be provided with a written consent form. Completed forms are to be returned to and counter signed by the PhD student (see section 13 for further details on the consent process). In the event of a low response rate, email reminders will be sent every 2-weeks for a period of 6-weeks (i.e., 3 reminders in total).

Recruitment of key stakeholders will commence after the quantitative analysis component of the study has been completed.

NOTE: A copy of the invitation email that will be sent to recruit key stakeholders has been provided with this application.

5. What are the ethical considerations involved in this proposal? You may wish, for example, to comment on issues to do with consent, confidentiality, risk to subjects, etc.

The main ethical considerations in this proposal are i) confidentiality and ii) two of the researchers' dual roles as assessors and researchers.

i) Participant OSCE, MCQ, MSA, LIFTUPP© and LEP data will need to be linked for comparisons to be made. It will also allow narratives on the development of competent dental practice to be established. For the datasets to be linked correctly, personal identifying factors will need to be made available. However, in order to preserve confidentiality, the researchers will not have access to the personal identifying data. Instead a third party who has no links to the study will perform the linkage and provide the study team with a pseudonymised linked dataset. OSCE, MCQ, MSA, LIFTUPP© and LEP data are to be linked and then pseudonymised by a third party before being transferred to the PhD student and his immediate supervisory team (only), where data will be stored on the University of Glasgow's secure networked J-Drive. This process will not affect undergraduate and VDP participants since the data is retrospective. The details of this process are described further in section 15. NOTE: LIFTUPP© and LEP data contain no patient details.

ii) Working with pseudonymised data will significantly reduce the possibility of the researchers who are also assessors of undergraduate students being able to identify individuals' assessment data. LEP data will not be identifiable since the researchers have no role in postgraduate assessment. However, there is a small chance that two of the researcher's (Jamie Dickie and Kurt Naudi) may recognise LIFTUPP© and examination performance data patterns as their academic responsibilities involve analysing these data sets in an unanonymised format in order to contribute to decisions on student progress. The size of the pseudonymised data set for this research project will mitigate this problem as it is likely that only data extremes (i.e., exceptionally good or poor performance compared to peers) will be attributable to certain individuals.

Despite this, student progress outcomes cannot be influenced since the assessment data will be analysed retrospectively. Furthermore, any influence from the two researchers who also have roles as assessors is diluted since they sit on a panel of 10 academic faculty that discusses student clinical experience and performance. Decisions on student progress are made collectively by the panel, i.e., the two researchers involved in this project are unable to cast sole judgement(s) on whether students proceed through the BDS course and/or graduate.

Amendment (October 2020): Since online focus groups are required, additional ethical considerations need to be considered regarding privacy and confidentiality.
Microsoft Teams and Microsoft Stream have been institutional approved by the University of Glasgow. Both platforms have been integrated into the University of Glasgow's Office365, which means they have

enterprise-grade security and the compliance required (https://docs.microsoft.com/en-us/microsoftteams/security-compliance-overview).

Under the EU's GDPR, Microsoft act as data "processors", meaning they will store, delete or disclose data as per the direction of the "customer". Therefore, the "customer" are the data controllers. In the case of this study, the PhD student is the "customer".

Personal data entered by customers joining a meeting via Microsoft Teams may include forenames and surnames. This research project will use these data to help participants identify each other during the focus group discussions – essentially serving as ID cards/place names. Participant names will be removed from transcriptions to ensure confidentiality.

Participants will be asked to refer to Microsoft's data privacy policies as part of the consent process (https://www.microsoft.com/en-gb/trust-center/privacy?rtc=1).

Some participants may join the online focus groups from their own homes and therefore their privacy may be compromised if they are using video communication. Those who wish to participate will be asked to ensure they join meetings from a private space. Alternatively, participants could also use the "apply background" feature of Microsoft Teams which can increase privacy by blocking out the background of the room from which they are broadcasting. Upon entry to the focus groups, all participants will be reminded of the need for confidentiality and that other participants have that right even if individuals have no issues with it themselves. Therefore, the need for discretion of all participants is required and if participants join from a non-private space, they will not be permitted to take part in, listen to or watch the discussions.

The host of the meeting (i.e., the PhD student) will use Microsoft Team's settings to ensure no video footage can be recorded until after participants have entered the online meeting room and then choose to activate their video footage themselves. This also gives participants the option on whether they wish to appear on the video recording of the meeting or not.

Participants who are external to the University of Glasgow (i.e., faculty from other UK dental institutions) will be provided with "guess access" – a feature available via Microsoft Teams. This will allow these participants to attend the meeting and see the content shared by the PhD researcher. Details on how "guess access" is granted will be provided to participants external to the University of Glasgow following the consenting process (see section 13).

Participants who have internal access to Microsoft Teams (i.e., University of Glasgow students and faculty) will be provided with details on how they can join the meeting following the consenting process.

6. Outline the reasons why the possible benefits to be gained from the project justify any risks or discomforts involved.

The project will inform assessor and regulators of the utility of a new assessment method and the overall risk of this project is low since it is secondary analysis of pseudonymised data.

At present, no robust evaluations of longitudinal assessments using objective outcome measures within dental education have been published. This study will aim to begin addressing this issue and investigate if continued investment in longitudinal assessment systems is merited within dental schools. It will also establish early foundations for determining the validity of longitudinal assessment in measuring dental student clinical competence and contribute to the SOHRC's and the University of Glasgow's dental education research strategies, as it coincides with their priorities of investigating the role of assessments in identifying competence and ensuring that the undergraduate curriculum prepares students for practice.

The study will be one of the first PhD projects to be completed among the dental educational research groups across Scotland and may suggest additional lines of enquiry as part of the SOHRC's remit. These studies could form an evidence-base for future developments on dental student teaching, assessment and curricular development.

Furthermore, the study will question if Glasgow's current competence assessment methods are still sufficient or potentially inferior to longitudinal assessment in establishing if dental students have been appropriately trained and assessed. Identifying individuals who are not (or not yet) fit to gain entry onto the professional register will ultimately safeguard and improve patient care, therefore assessment methods relating to student progression should be the most valid available. This corresponds with the new General Data Protection Regulation (GDPR) processing conditions on the use of data that are necessary for performance of a task in the public interest.

7. Who are the investigators (including assistants) who will conduct the research? What are their qualifications and experience?

Jamie Dickie

Jamie Dickie graduated Bachelor of Dental Surgery (BDS) from the University of Glasgow in 2009. He subsequently completed vocational training in general practice and dental foundation posts in hospital and community health services. During this period, he passed his Member of the Faculty of Dental Surgery (MFDS) examinations for the Royal College of Physicians and Surgeons of Glasgow.
In 2014, he was appointed Clinical Lecturer/Honorary Specialty Registrar in Restorative Dentistry at the University of Glasgow Dental School. He obtained a Master of Education (MEd) in Learning and Teaching in Higher Education in 2017.

Kurt Naudi

Dr Kurt Naudi graduated Bachelor of Dental Surgery (BChD) from the University of Malta in 2001. He went on complete postgraduate dental training positions in both Malta and the UK, during which time he completed his membership examinations for the Royal College of Physicians and Surgeons of Glasgow.
In 2011, he completed a Doctorate in Dental Surgery as well as a Master of Education (MEd) in Learning and Teaching in Higher Education in 2014. Kurt was then appointed Clinical Senior University Teacher/Honorary

Consultant in Oral Surgery by the University of Glasgow Dental School, where he is currently coordinator of the 4th Year BDS undergraduate course and the postgraduate Master of Science (MSc) in Oral and Maxillofacial Surgery programme.

Andrea Sherriff

Dr Andrea Sherriff graduated from Strathclyde University with a BSc (Hons) in Mathematical Sciences. She received her PhD in Statistics and Modelling Sciences in 1996, after which she took up a post as a statistician with the Avon Longitudinal Study of Parents and Children (ASLPAC) in the department of Social Medicine, University of Bristol.

After a 3-year Wellcome Trust Training Fellowship in Mathematical Biology, and a secondment to the Laboratory of Statistical Genetics at Rockefeller University, New York City, she took up a Research Fellow post at the Paediatric Epidemiology and Child Health (PEACH) Unit at Glasgow University.
In 2008, Andrea was appointed a Lecturer in Statistics at the University of Glasgow Dental School before promotion to Senior Lecturer (2010). She is Director of Postgraduate Student affairs at the Dental School, and leads the TORCH programme of research within the Dental School. She has experience working on large linked secondary dataset within the National Safe Haven. She provides support for research design and statistical analysis to the research groups within the Dental School, and leads on the teaching of Evidence-Based Practice to dental undergraduate and postgraduate students.

Michael McEwan

Dr Michael McEwan was awarded PhD in Applied Mathematics from the University of St Andrews.
Much of his early career focused on research in mathematically modelling solar phenomena before shifting to teaching mathematics, physics and statistics in higher education at Glasgow International College in 2006. Having gained a Master of Education (MEd), he was then appointed University Lecturer/Senior Academic and Digital Development Adviser at the University of Glasgow Learning and Teaching Centre, where he is responsible for the coordination of the Postgraduate Certificate in Academic Practice (PGCAP).

8. Are arrangements for the provision of clinical facilities to handle emergencies necessary? If so, briefly describe the arrangements made.

This study does not encompass clinical work and therefore no clinical facilities are required.

9. In cases where subjects will be identified from information held by another party (e.g., a doctor or hospital), describe how you intend to obtain this information. Include, where appropriate, which Multi Centre Research Ethics Committee or Local Research Ethics Committee will be applied to.

LEP data are owned and controlled by NES, who have fully supported the study and are willing to contribute and assist where necessary. A data sharing agreement is currently being drawn up between the University of Glasgow and NES to allow for transfer of LEP data transferred via a secure NHS.NET email onto a drive

owned by the University of Glasgow. This will be done in accordance with recent policy changes in data protection (GDRP) within the University of Glasgow.

Once LEP data has been linked to the relevant LIFTUPP© data and pseudonymised by a third party and stored on the University of Glasgow J-Drive (see section 15), the emails containing LEP data will be deleted. NES have reviewed and are satisfied with the University of Glasgow's data security protocol. A copy of the data security protocol and confirmation of approval from NES have been provided alongside this application form.

9. Specify whether subjects will include students or others in a dependent relationship and, where possible, avoid recruiting students who might feel to be, or be construed to be, under obligation to volunteer for a project. This is most likely to be when a student is enrolled on a course where the investigator is a teacher. In these circumstances, the recruitment could be carried out by one of the other investigators or a suitably qualified third party.

The quantitative component involves secondary analysis of pseudonymised retrospective student assessment data. No students will be identified during analysis for this part of the study.
For the qualitative component, undergraduate dental students will be invited to participate via email by the PhD student. The invitation will contain a participant information leaflet detailing the purpose of the project, its methods and how their data will be used, anonymised and stored. They will also be given assurances that:

- their participation is entirely voluntary
- no prejudice will be set between those who chose to participate and those who do not
- they are free to withdraw from the study up to (and including) the date on which focus group/semi-structured interview conversations are transcribed*
- their responses will have no impact on their academic records and progress
- data gathered from focus groups/semi-structured interviews will be used for research purposes only.

To further reduce any potential anxiety caused by dependant relationships, students will participate in their own focus groups (i.e., there will be no other teaching staff present apart for the lead investigator).

*Transcription of the focus group/semi-structured interview discussions will remove any references to (i.e., anonymise) participants (see section 15), so they will no longer be able to withdraw after this process has been completed.

11. Specify whether the research will include children or participants with mental illness, physical disability or intellectual disability. If so, please explain the necessity of involving these individuals as research subjects and include documentation of the

suitability of those researchers who will be in contact with children (e.g., Disclosure Scotland or membership of the PVG Scheme).

The study will not include children or participants with mental illness, physical disability or intellectual disability.

12. Will payment or other incentive, such as a gift or free services, be made to any research subject? If so, please specify, and state the level of payment to be made and/or the source of the funds/gift/free service to be used. Please explain the justification for offering an incentive.

Undergraduate students who volunteer to take part in the focus group(s) will be offered pizza and soft drinks as a token of appreciation for their participation. These incentives will be made available over the duration of the focus groups.

No incentives will be made to faculty members taking part in the study.

Amendment (October 2020): Incentives will be no longer be made to key stakeholders since the focus groups will take place online.

13. Please give details of how consent is to be obtained. A copy of the proposed consent form, along with a separate information sheet, written in simple, non-technical language MUST ACCOMPANY THIS PROPOSAL FORM.

Potential participants for the study focus groups/semi-structured interviews will be provided with a participant information leaflet (via email from the PhD student) that outlines the purpose of the project, the methods and how the gathered data will be used. It will also signify participation is voluntary, they can withdraw at any time and their recorded data will be anonymised and securely stored (as detailed in previous sections).

Those interested in participating are to respond to the PhD student via email. The PhD student will then provide each volunteering participant with a written consent form. This form will need to be signed by participants and returned to the PhD student for countersigning to complete the consent process.
Copies of the participant information leaflets and written consent forms have been submitted alongside this application form.

Amendment (October 2020): Alternative participant information leaflets and consent forms will need to be provided for online focus groups. Copies of these participant information leaflets and consent forms are attached to this document.
Consent forms will need to be signed electronically due to COVID-19 social distancing measures.

Participants will be asked to refer to Microsoft's privacy policies within the information leaflets and their invitation email (https://www.microsoft.com/en-gb/trust-center/privacy?rtc=1).

14. Comment on any cultural, social or gender-based characteristics of the subjects which have affected the design of the project or may affect its conduct.

No cultural, social or gender-based characteristics have affected the design of this study or should affect its conduct.

15. Please state (i) who will have access to the data, (ii) how the data will be stored, how will access be restricted, and (iii) what measures will be adopted to maintain the confidentiality of the research subjects and to comply with data protection requirements.

i) A third-party University of Glasgow academic staff member will originally have access to the quantitative OSCE, MSA, MCQ, LIFTUPP© and LEP data in order to link and pseudonymised the data prior to analysis. This party will otherwise have no future access to the data.

The linked and pseudonymised data will then be made accessible to the researchers listed in section 7 for analysis.

Qualitative data from focus groups/semi-structured interviews of key stakeholders will only be accessible to the researchers listed in section 7.

ii) OSCE, MSA, MCQ, LIFTUPP© and LEP data will be transferred to a secure networked data drive (J-drive) folder at the University of Glasgow by their respective data controllers via secure email servers. These emails will be deleted once the data have been transferred to the J-Drive. LEP data will be transferred in three blocks (Cohort 1 – August 2018, Cohort 2 – August 2019 and Cohort 3 – August 2020). Each block will be linked and pseudonymised by the third party academic staff member before being stored on a separate folder on university secure drive that can only be accessed by the researchers. No additional individuals will be given authorised access to the pseudonymised data folder.

Audio recordings from focus groups/semi-structured interviews will be transferred to the researchers' secure university directory drive. The discussions will then be transcribed by the lead researcher and the audio recordings will then be deleted from the device they were originally recorded on.

iii) OSCE, MSA, MCQ, LIFTUPP© and LEP data features will be not be available to the research team once linkage and pseudonymisation has taken place and the emails transferring the data have been deleted.

Pseudonymised data will not be stored on any other format or in any other location other than the J-Drive, to which only the researchers (and the third party data linker) will have authorised access.

Audio recordings and transcriptions from focus groups/semi-structured interviews will also be stored on the University of Glasgow's secure networked J-Drive. Transcriptions will contain no identifiable data (e.g., names). Instead, participants will be assigned an alphanumeric code that will only be known to the researchers.

In regard to (ii) above, please clarify (tick one) how the data will be stored:

☐ **(a) in a fully anonymised form (link to subject broken),**

☑ **(b) in a linked anonymised form (data +/- samples linked to subject identification number but subject not identifiable to researchers), or**

☐ **(c) in a form in which the subject could be identifiable to researcher.**

If data are stored in linked anonymised form, please state who will have access to the code and personal information about the subject.

Codes and personal information for participants involved in the quantitative study component will be accessible to an experienced member of University of Glasgow academic staff (not otherwise involved in the study), who will act as the third party for linking and pseudonymising all forms of assessment data.

The data will be held securely for a period of ten years after the completion of the research project, or for longer if specified by the research funder or sponsor, in accordance with the University's Code of Good Practice in Research. (http://www.gla.ac.uk/media/media_227599_en.pdf ☑ **Please tick**

Amendments (October 2020):

i) Qualitative data (i.e., audio and video recordings and transcripts) from focus groups with key stakeholders will only be accessible to the PhD student and his project supervisors (see section 7 of original application).

ii) Focus group audio and video will initially be saved to the PhD researcher's Microsoft Stream account which is part of the University of Glasgow's Office 365 package. Recordings made via Microsoft Teams are automatically saved to Microsoft Stream and will allow the recordings to be transcribed. Only the PhD researcher will be able to access the recordings uploaded to Microsoft Stream.
Once the transcriptions have been completed, they will be saved to the University of Glasgow's secure networked J-Drive.

Once the content of the transcripts has been verified by the PhD student, the recordings uploaded to Microsoft Stream will be deleted as will the backup audio file on the Dictaphone.

iii) Only the focus group host (i.e., the PhD student) will be able to initiate audio and video recordings via Microsoft Teams.

Once the transcriptions have been completed, checked, and verified, the audio and video recordings of the focus groups will be deleted.

Transcriptions will contain no identifiable data (e.g., names). Instead, participants will be assigned an alphanumeric code that will only be known to the researchers. Deletion of the audio and video recording will provide greater anonymity to the transcripts.

Transcripts will be kept for 10-years in accordance with the University of Glasgow's research polices.

16. To your knowledge, will the intended group of research subjects be involved in other research? If so, please justify.
To our knowledge, the intended groups of researcher subjects will not be involved in other research.

17. Proposed starting date: 1st June 2018

Expected completion date: 1st October 2023

18. Please state location(s) where the project will be carried out.
University of Glasgow Dental School.

19. Please state briefly any precautions being taken to protect the health and safety of researchers and others associated with the project (as distinct from the research subjects), e.g., where blood samples are being taken.
There are no perceived health and safety risks to the researchers and others within this project as it is analytic.

20. Please state all relevant sources of funding or support for this study.
The project has been supported by NHS Education for Scotland (NES), who have agreed to extend the lead researcher's postgraduate specialist training by an additional 3.5-years (8.5-years in total) to allow this study to be conducted.

Registration fees are supported by the Dorothy Geddes Studentship, which is awarded by the University of Glasgow Dental School (for up to 5-years postgraduate research study).

Additional sources of funding will be sought to cover any educational fees incurred during the course of study, e.g., the TC White Young Researcher Award offered by the Royal College of Physicians and Surgeons of Glasgow.

21a). Are there any conflicts of interest related to this project for any member of the research team? This includes, but is not restricted to, financial or commercial interests in the findings. If so, please explain these in detail and justify the role of the research team. For each member of the research team please complete a declaration of conflicts of interest below.

There are no conflicts of interest to declare, however the researchers are aware of the potential implications of two members of the team having a duel role as both researchers and assessors of undergraduate students (see section 5).

Researcher Name: _____Jamie Dickie_____ conflict of interest Yes / ***No***
If yes, please detail below

Researcher Name: _____Kurt Naudi_____ conflict of interest Yes / ***No***
If yes, please detail below

Researcher Name: _____Andrea Sherriff_____ conflict of interest Yes / ***No***
If yes, please detail below

Researcher Name: _____Michael McEwan_____ conflict of interest Yes / ***No***
If yes, please detail below

21b). If there are any conflicts of interest, please describe these in detail and justify conducting the proposed study.

The researchers have reflected fully on this and consider that there are no conflicts of interest to be declared for the commencement of this study.

22. How do you intend to disseminate the findings of this research?

Present at local, national and international education meetings (oral and poster formats), e.g., the University of Glasgow Learning and Teaching Conference, Glasgow Dental School's annual Postgraduate Research Prize Seminars, the Scottish Oral Health Research Collaboration (SOHRC) and the Association for Dental Education in Europe (ADEE).

Provide feedback to NES, colleagues in LIFTUPP© user group and LIFTUPP© Limited.

Produce academic papers for submission to medical education journals, e.g., The European Journal of Dental Education.

I confirm that have read the University of Glasgow's Data Protection Policy.
[http://www.gla.ac.uk/services/dpfoioffice/policiesandprocedures/dpa-policy/]
Please initial box ☑

Name ___ SIGNTAURE REDACTED _____ Date ___6/11/2020____

(Proposer of research)
Please type your name on the line above.

For student projects:

I confirm that I have read and contributed to this submission and believe that the methods proposed and ethical issues discussed are appropriate.

I confirm that the student will have the time and resources to complete this project.

Name _____SIGNTAURE REDACTED _____ Date __6/11/2020__
(Supervisor of student)

Please type your name on the line above.

Please upload the completed and <u>signed</u> form, along with other required documents by logging in to the Research Ethics System at - https://frontdoor.spa.gla.ac.uk/login/

ii) Original ethical approval confirmation - 24th October 2018

Dear Dr Kurt Busuttil Naudi

MVLS College Ethics Committee
Project Title: Longitudinal assessment of dental undergraduates
Project No: 200170146

The College Ethics Committee has reviewed your application and has agreed that there is no objection on ethical grounds to the proposed study.

We are happy therefore to approve the project, subject to the following conditions.

- Project end date as stipulated in original application.
- The data should be held securely for a period of ten years after the completion of the research project, or for longer if specified by the research funder or sponsor, in accordance with the University's Code of Good Practice in Research: (http://www.gla.ac.uk/media/media_227599_en.pdf)
- The research should be carried out only on the sites, and/or with the groups defined in the application.
- Any proposed changes in the protocol should be submitted for reassessment, except when it is necessary to change the protocol to eliminate hazard to the subjects or where the change involves only the administrative aspects of the project. The Ethics Committee should be informed of any such changes.
- For projects requiring the use of an online questionnaire, the University has an Online Surveys account for research. To request access, see the University's application procedure at https://www.gla.ac.uk/research/strategy/ourpolicies/useofonlinesurveystoolforresearch/ .
- You should submit a short end of study report to the Ethics Committee within 3 months of completion.

Yours sincerely

Dr Terry Quinn

Terry Quinn
FESO, MD, FRCP, BSc (hons), MBChB (hons)
Senior Lecturer / Honorary Consultant

College of Medicine, Veterinary & Life Sciences
Institute of Cardiovascular and Medical Sciences
New Lister Building, Glasgow Royal Infirmary
Glasgow
G31 2ER
terry.quinn@glasgow.gla.ac.uk
Tel – 0141 201 8519

iii) Updated ethical approval confirmation – 12th November 2020

RE: Amendments - Project number 200170146

ⓘ You replied on Thu 12/11/2020 11:00

MA

MVLS Ethics Admin
Thu 12/11/2020 10:01
To: Jamie Dickie; MVLS Ethics Admin
Cc: Kurt Naudi; Michael McEwan

Hi Jamie

The Chair has approved the requested amendments. Please treat this email as confirmation.

Regards
Neil

Neil Allan
MVLS Ethics Committee Administrator

Direct line: 0141 330 5206
email is the only reliable form of contact at this time

Institute of Infection, Immunity & Inflammation
College of Medical, Veterinary & Life Sciences
Glasgow Biomedical Research Centre
Room 314, Sir Graeme Davies Building
University of Glasgow
120 University Place
Glasgow G12 8TA
The University of Glasgow, charity number SC004401

iv) GDPR Legitimate Interests Assessment

PhD Project – GDPR Legitimate Interests Assessment (LIA)

Jamie Dickie – Clinical Lecturer/Honorary Specialty Registrar in Restorative Dentistry, PhD student

Premise

This Doctor of Philosophy (PhD) project aims to investigate longitudinal assessment in the undergraduate Bachelor of Dental Surgery (BDS) curriculum to establish its utility as an assessment tool and its association with postgraduate performance. It is aligned with the top priority by the Scottish Oral Health Research Collaboration (SOHRC) for dental education research – "*The role of assessments in identifying competence*" (Ajjawi et al., 2017).

The study involves linking undergraduate assessment data from an electronic longitudinal system (known as LIFTUPP©) and summative, standalone examinations used by the University of Glasgow Dental School. It will also obtain postgraduate longitudinal assessment data from Longitudinal Evaluations of Performance (LEPs), which are used in Scottish postgraduate dental vocational training (VT) schemes. Data will be sourced from two cohorts (the graduating classes of 2017 and 2018), before being linked and pseudonymised by a third-party staff member of the University of Glasgow. The data will then be made available to the researchers for analyses. Personal data (first and surnames, date of birth and matriculation numbers) are required by the third party to initially link the various data sets. For this reason, it is the responsibility of the researchers to ensure that the project is compliant with the European Union's (EU's) new General Data Protection Regulations (GDPR).

Since the project will retrospectively analyse secondary data that has been pseudonymised, the researchers believe that it is possible to conduct the study without the consent of the intended participants (who are now former students). However, for such circumstances, the UK Information Commissioner's Office (ICO) has suggested that researchers need to justify that there is a legitimate interest for the study. ICO's legitimate interests provision is broken down into a three-part test:

1. Purpose test - Are you pursuing a legitimate interest?
2. Necessity test - Is the processing necessary for that purpose?
3. Balancing test - Do the individual's interests override the legitimate interest?

The following document lists the series of questions associated with each of these three-part test sections and provides the researchers' responses in relation to this project.

1. Purpose Test

A) Why do you want to process the data?

Processing these data will allow Glasgow Dental School (and other dental schools) to begin evaluating the validity of the LIFTUPP© assessment system in determining the development of competent clinical practice of undergraduate students. Glasgow Dental School has made (and continues to make) significant investments in implementing the LIFTUPP© system into their undergraduate programme. However, since LIFTUPP© has only been in use since the 2014/15 academic year, the extent to which these data provide any meaningful contribution to the assessment of graduating dentist's competence has yet to be investigated. Evaluating this assessment method against other methods and future clinical performance is important to ensure dental schools produce competent practitioners. Furthermore, the Quality Assurance Agency (QAA) Scotland advocates that evaluation of assessment methods for the purpose of their enhancement is an indicator of good educational practice (Quality Assurance Agency for Higher Education, 2011).

B) What benefit do you expect to get from the processing?

The main benefit of the project will be to contribute evidence on the validity of using longitudinal assessment to determine dental students' clinical competence, and if shown to be valid, these methods can be used in the assessment of undergraduates with confidence. On the other hand, if this study identifies issues or concerns with this method of assessment, measures can be taken within curricula to address these. Presentations and publications will be produced to disseminate the study findings among the academic community. These will be informative to educational institutes (both within and outside of dentistry) that use, or are considering the use of, longitudinal assessment methods. In turn, this will allow educational institutes to make evidence-based decisions on what assessment methods they wish to adopt.

Since this investigation will be carried out by means of a PhD project, the study process will serve as research training for the main researcher (Jamie Dickie) and successful completion would result in the award of a PhD qualification from the University of Glasgow.

C) Who else benefits from the processing (third parties/the public)?

Future students will benefit as they can assured that they have been assessed using methods that have been validated for the purpose they are being used for.

Various groups benefit from ensuring that dental students have been thoroughly assessed using the most valid, appropriate and robust methods according to available evidence. These include:

- The UK regulator of dentistry, the General Dental Council (GDC) - who will be provided with proof that graduating dental students are competent clinicians and ready to be admitted onto the professional register.
- Dental Vocational Trainers – who can be confident that they are recruiting "safe beginners" as their Vocational Dental Trainees (VDPs). This will assure them that their VDP will be a competent clinician but is also able to assess their own capabilities and limitations, act within these boundaries and knows when to request support and advice from a trainer.
- The public – who are ultimately protected from being treated by individuals who are not fit to practise dentistry (see answer to question D).

D) How important are those benefits?

Collectively, these benefits are very important to the UK National Health Service (NHS), which has significant interest in patient safety and protecting the public. Part of safeguarding patients is to ensure that healthcare practitioners are appropriately trained and assessed before being permitted onto professional registers, thus preventing the public from being treated by individuals who are not fit to practise dentistry. The importance of clinical education, training and assessment, particularly at undergraduate level, was reiterated by the publication of the Francis Report (2013).

Dental schools should continually review their assessment methods and seek to adopt the best practice in accordance with the available evidence. Longitudinal performance data have recently been proposed as one of the strongest means of assessing student competence in medical-based subjects compared to more traditional standalone/single encounter assessment methods (Albino et al. 2008; Dawson et al., 2017). However, there are currently no robust evaluations of longitudinal assessments using objective outcome measures within dentistry and, therefore, there is a need for further studies on this topic.

Furthermore, the unit cost to the UK public purse of training a dental student is at least £250,000. Accurate assessment of student competence is a vital component of the BDS course. It is essential that we are confident of the utility of our assessment methods and that they represent good value for money.

E) What would the impact be if you couldn't go ahead?

Since evaluation of assessment methods is critical for good educational practice, not proceeding with this project is to ignore the expectations placed on higher education institutes by the QAA. The study may also have several other potential benefits in terms of economics, regulation of the dental profession and protection of the public, which could be missed if it were not conducted. An economic benefit would stem from the study being able to help determine the value of the LIFTUPP© system. If the system appears to have little educational benefit, then Glasgow Dental School (and other dental schools) may decide that they are best directing their funding elsewhere. Alternatively, if there is evidence that LIFTUPP© data can be used to for a multitude of purposes (determining competence in this case), then it will serve as a basis for additional research that could inform and maximise the system's potential.

Potential benefits for regulation of the dental profession and protection of the public have been discussed in the answers to questions C and D. If this study were not carried out, dental schools will be unsure as to whether they are using the best available assessment methods for detecting individuals who are not fit to practise dentistry which puts the public at risk of harm.

F) What is the intended outcome for individuals?

The project will have no effect on the participants since their assessment data will be analysed retrospectively (i.e., participants will have already graduated from dental school and subsequently completed their postgraduate vocational training). Therefore, undergraduate student and postgraduate vocational trainee progress outcomes cannot be influenced.
Furthermore, no identifiable personal data will be published in any subsequent presentations, publications or reports.

G) Are you complying with other relevant laws and industry guidelines/codes?

Yes – we are complying with the University of Glasgow's Data Security Protocol.

H) Are there any ethical issues with the processing?

The main ethical considerations are *i)* confidentiality and *ii)* two of the researchers' dual roles as assessors and researchers.

i) As discussed in the answer to question A, participant undergraduate and postgraduate assessment LEP data will need to be linked for comparisons to be made and allow narratives on the development of competent dental practice to be established. For the datasets to be linked correctly, personal identifying factors (first and surnames, date of

birth and matriculation numbers) will need to be made available. However, to preserve confidentiality, the researchers will not have access to the personal identifying data. Instead a third party who has no links to the study will perform the linkage and provide the study team with a pseudonymised linked dataset

ii) Though highly unlikely, there is a small chance that two of the researchers (Jamie Dickie and Kurt Naudi) may recognise LIFTUPP© and examination performance data patterns for outliers, as their academic responsibilities involve analysing these data sets in an identifiable format to contribute to decisions on student progress. However, even if the unlikely event of student identification were to occur, the researchers can have no influence on student or VDP progression (from both an academic and career prospective) as the participants would have already graduated from Glasgow Dental School and completed their postgraduate vocational training period.

Ethical approval has been sought through the University of Glasgow's College of Medical, Veterinary & Life Sciences (MVLS) Ethics Committee for Non-Clinical Research Involving Human Subjects. Initial feedback from the committee suggested they were satisfied with steps taken by the researchers to address these ethical issues and full approval would be granted on receipt of written confirmation from the relevant data controllers (i.e., Glasgow Dental School for undergraduate data; NHS Education for Scotland for postgraduate data) which permits the researchers to progress with the study.

I) Are you processing for fraud prevention, IT security or any of the purposes highlighted by the GPDR?
No.

2 – Necessity Test

A) Will the processing actually help you achieve your purpose?
Yes. Since an assessment system is being evaluated, data generated by that system need to be processed so they can be quantitatively and qualitatively analysed. Undertaking statistical modelling and analyses and any subsequent evaluations can only be completed if the data are processed as described in the premise of this document.

The results generated by statistical modelling and analyses can then be formatted and presented to key stakeholders within dental education to investigate how they impact on and/or could be used to improve dental student assessment.

B) Is the processing proportionate to that purpose?

Yes. The data management and processing are standard for this type of project and completely proportionate. Data must be processed to create derived variables necessary to answer the questions. Large amounts of data must be aggregated and summarized to produce meaningful interpretation.

C) Can you achieve your purpose without processing the data, or processing less data?

No. A substantial amount of data processing, aggregation and statistical analyses need to be performed to achieve the aims. As it stands, the datasets are too large to do this without necessary processing and statistical analyses. All the data requested are required to perform the analyses.

D) Can you achieve your purpose by processing the data in another more obvious or less intrusive way?

No. Given the nature of the data, there are no obvious or less intrusive ways to process the data to answer the research questions proposed by this study.

3 – Balancing test

As a minimum consider:

i) The nature of the personal data you want to process;

Identifiable information (first name, surname, date of birth and matriculation numbers) will only be used to link undergraduate and postgraduate assessment data by a third party. Once this has been completed, the data will be pseudonymised (i.e., the identifiable information will be removed) before being made available to the researchers for processing and analyses. The data that will be processed and analysed is historic, secondary assessment data that can no longer influence or impact on the progression of current or former students.

ii) The reasonable expectations of the individual;

The UK Quality Code for Higher Education states that higher education providers are expected to seek to enhance the quality of their programmes by continually monitoring and reviewing them (Quality Assurance Agency, 2011). This process can involve using data on student progression and achievement. Therefore, students should reasonably expect that their assessment data could be used by their educational institutions to review and improve assessment processes.

iii) The likely impact of the processing on the individual and whether any safeguards can be put in place to mitigate negative impacts.

The project will have no impact on the participants. See answer to question F in the ***Purpose Test*** section.

Conclusion/Outcome

From the answers given to the three-part test, there appears to be a legitimate interest for this study to be conducted. There is minimal risk to the participants and it will contribute towards improving future assessment within dental education, as the information gathered will help inform dental assessors and regulators on the use of longitudinal data as a method of assessment. Ultimately, this will help protect the public by ensuring that dental students are being assessed using the best, evidence-based methods available before their entry onto the professional registers.

References

Ajjawi, R., Barton, K.L., Dennis, A.A., Rees, C.E., (2017). Developing a national dental education research strategy: priorities, barriers and enablers. *BMJ Open*, 7: e013129.

Albino, J.E et al. (2008). Assessing dental students' competence: best practice recommendations in the performance assessment literature and investigation of current practices in predoctoral dental education. *Journal of Dental Education*, 72(12), pp.1405-1435.

Dawson, L. et al. (2017). Calling for a re-evaluation of the data required to credibly demonstrate a dental student is safe and ready to practice. *European Journal of Dental Education,* 21(2), pp.130-135.

Francis, R. (2013) Report of the Mid Staffordshire NHS Foundation Trust Public Inquiry. London: The Stationery office.

GDC, (2015). General Dental Council: Preparing for Practice. London.

Quality Assurance Agency, 2011. UK Quality Code for Higher Education, part B: Assuring and enhancing academic quality.

v) Letter from Head of Glasgow Dental School

University of Glasgow

Dental School

JB/ew

13th September 2018

TO WHOM IT MAY CONCERN

Dear Sir / Madam

Re: MVLS Ethics Committee Application No200170146 submitted by Dr Kurt Naudi

I am writing to confirm that in the context of GDPR, Mr Jamie Dickie, the PhD student associated with this study, has completed an extensive Legitimate Interests Assessment which in my view satisfies the requirements of a legitimate interest for this study.

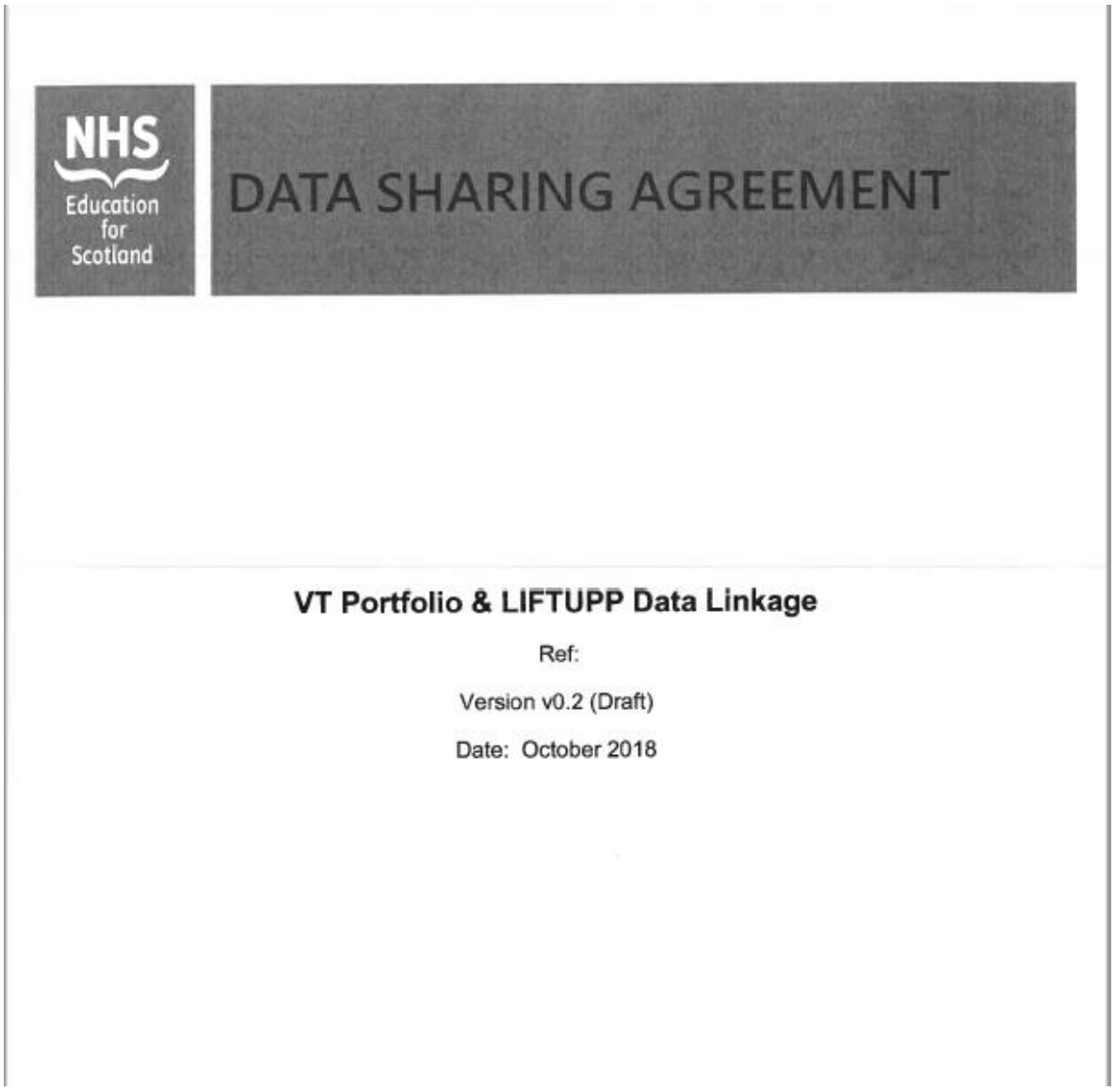
From a University perspective, therefore, I am content that utilisation of the stored data for the purposes identified in the protocol is entirely acceptable.

Yours faithfully

[redacted]

Jeremy Bagg

vi) Data sharing agreement – University of Glasgow/NHS Education for Scotland



NHS Education for Scotland

DATA SHARING AGREEMENT

VT Portfolio & LIFTUPP Data Linkage

Ref:

Version v0.2 (Draft)

Date: October 2018

Contents

1 Parties, Scope and Purpose

1.1 Name and details of the parties who agree to share information

Legal name of parties to DSA	Short name of the party	Head Office address	ICO Registration
NHS Education for Scotland	NES	Westport 102 West Port Edinburgh EH3 9DN	Z7921413
The University Court of the University Glasgow	UoG	University Avenue Glasgow G12 8QQ	Z6723578

1.2 Business and legislative drivers.

1.2.1 Purpose of the information sharing

Purpose description	Primary or secondary purpose
The purposes of this information sharing agreement is to support a research project linking undergraduate and postgraduate dental student assessment data undertaken by a PhD Student by the sharing of NES LEPs data. The study involves linking undergraduate assessment data from an electronic longitudinal system (known as LIFTUPP©) and summative, standalone examinations used by the University of Glasgow Dental School to postgraduate longitudinal assessment data from Longitudinal Evaluations of Performance (LEPs), which are used in Scottish postgraduate dental vocational training (VT) schemes. Data will be sourced from two cohorts (the graduating classes of 2017 and 2018), before being linked and pseudonymised by a third-party staff member of the University of Glasgow. The data will then be made available to the researchers for analyses. Processing these data will allow Glasgow Dental School (and other dental schools) to begin evaluating the validity of the LIFTUPP© assessment system in determining the development of competent clinical practice of undergraduate students. Glasgow Dental School has made (and continues to make) significant investments in implementing the LIFTUPP© system into their undergraduate programme. However, since LIFTUPP© has only been in use since the 2014/15 academic year, the extent to which these data provide any meaningful contribution to	

the assessment of graduating dentist's competence has yet to be investigated. Evaluating this assessment method against other methods and future clinical performance is important to ensure dental schools produce competent practitioners.	

Indicate how the data controllers will decide upon changes in the purposes of the sharing	Jointly or independently
	Jointly.

The instructions for reaching agreement on changes in the purposes of the sharing is described in the [Name of the Instructions] listed in Appendix 1 Instructions.

1.2.2 Legal basis for the processing and constraints

If sharing personal data:	
Article 6 conditions met	Article 9 conditions met
NES: 6(1)(b) – processing is necessary for the performance of a contract with the data subject or to take steps to enter into a contract; or 6(1)(e) – processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller.	N/A – no special category data is being used as part of this project
University of Glasgow: 6 (1) (e) – processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller.	N/A – no special category data is being used as part of this project

2 Description of the information to be shared

Data category	Data Controller status	PD* / SCD*
Student Name	Data Controller – NES	PD
Date of Birth	Data Controller – NES	PD
Gender	Data Controller – NES	PD
Cohort Dates	Data Controller – NES	PD
LEP data - including: - LEP number	Data Controller – NES	PD

- Location - Grade - Code of post - Trainer ID - Status - LEP date - Details of Encounter - Case complexity - Examination (LEP grade/rating) - Clinical Judgement (LEP grade/rating) - Technical ability (LEP grade/rating) - Communication (LEP grade/rating) - Professionalism (LEP grade/rating) - Knowledge (level and application) (LEP grade/rating) - Organisation (LEP grade/rating) -Trainee's Insight into Performance (LEP grade/rating)		

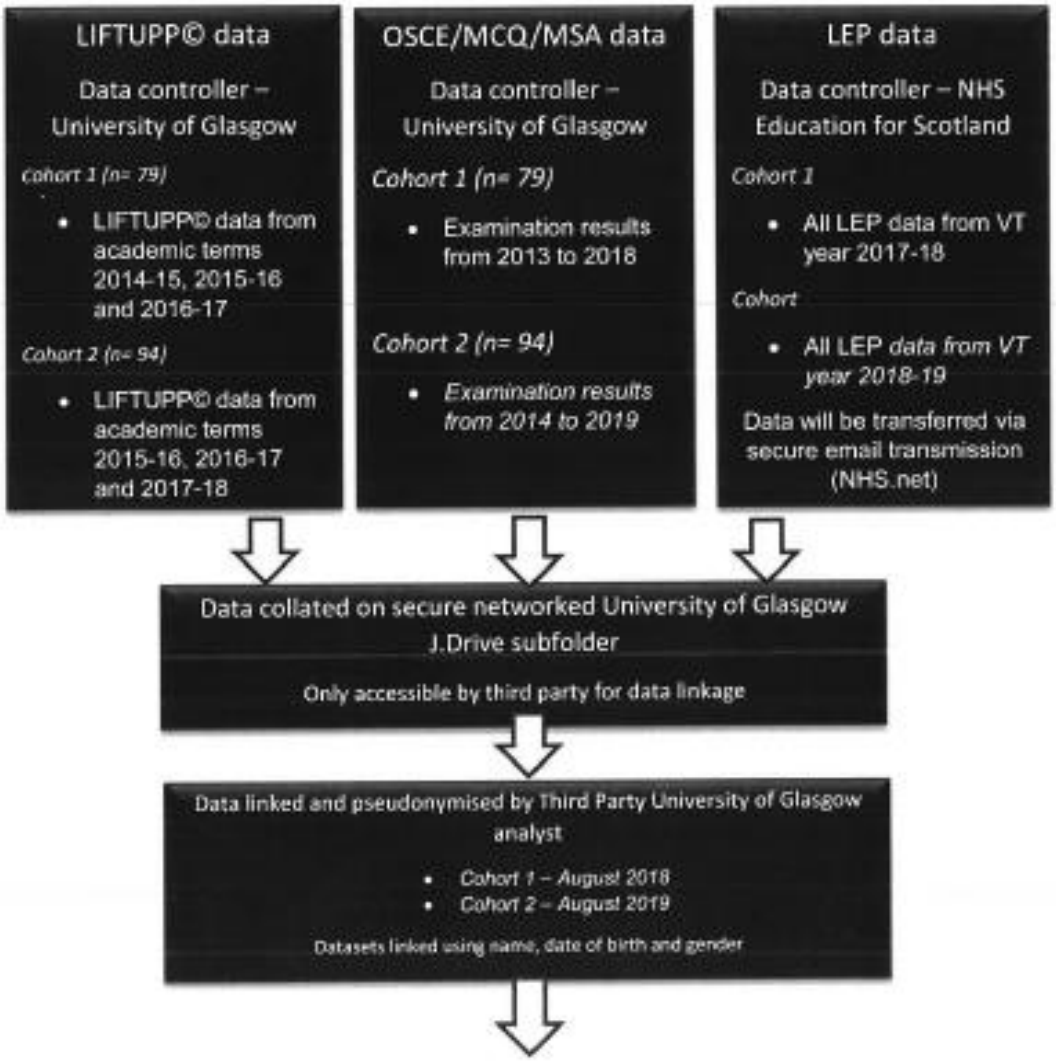
(*) PD – Personal Data as defined within the General Data Protection Regulation (Art. 4(1))

(*) SCD – Special Category Data as defined within the General Data Protection Regulation (Art. 9(1)).

The parties agree this is the minimum amount of data needed to properly fulfil the purposes of this agreement.

3 Description and manner of information sharing

3.1 Data flows



Pseudonymised data sent to 2nd secure networked J Drive subfolder

Data only accessible by Jamie Dickie and Andrea Sherriff after transfer

Analysed by researcher and supervisors

- *Cohort 1 data – deposited August 2018*
- *Cohort 2 data – deposited August 2019*

No disclosive data will be reported/published

3.2 How data/information is to be accessed, processed and used

Data use description	Associated work instructions, policy or procedure (listed in Appendix 1) if applicable
Personal identifiable data will be only used to link undergraduate and postgraduate assessment data by a third party within the University of Glasgow (an independent analyst from within the Unit). The PhD researcher will only have access to the data once data linkage and pseudonymisation has been completed.	• Public Benefit & Privacy Panel for Health & Social Care (PBPP) application submitted • NES Information Governance and Security policies and procedures • UoS Information Governance Policies and Procedures
Analysed (aggregated non-disclosive) data will be published as part of the PHD student's thesis, academic papers and reports, which may be presented and shared within academic forums	

4 Impact assessments and preparatory work

- University of Glasgow – COHS Data Security Protocol
- University of Glasgow – UNIT signed security agreement
- The University of Glasgow's College of Medical, Veterinary & Life Sciences (MVLS) Ethics Committee for Non-Clinical Research Involving Human Subjects (copy of ethics application, reviewer response and study proposal provided)

5 Fair processing

5.1 List of relevant Fair Processing Notice(s)

- NES Data Protection Notice (https://www.nes.scot.nhs.uk/privacy-and-data-protection.aspx)
- UoG Data Protection Notice (https://www.gla.ac.uk/media/media_590481_en.pdf

5.2 Impact on people interests

Impact description	Control measure
Risk of small number identification in published outputs	Adhere to University Statistical Disclosure Policy
Risk of accidental disclosure of stored personal data during data transfer	• Ensure data is transferred securely. • All NHS Scotland Staff are required to undertake mandatory IG Training • UoG - Medical Research Council (MRC) information governance training

Risk of accidental disclosure of stored personal data due to unauthorised access to the data store	• NES Information Governance Policies and Procedures
	  ○ Information Governance Policy
	  ○ Data Protection Management Procedures
	  ○ Corporate Information Security Policy
	• UoG Policies and Procedures
	  ○ Regulations for the Use of University ICT Systems and Facilities.
	  ○ Information Security: Confidential Data policy
	  ○ Network service, Systems and Data Communications Monitoring policy.
	  ○ Guidelines and Procedures for Blocking Network Access.
	  ○ Network Connection policy
	  ○ Antivirus and Firewalls policy.
	  ○ Glasgow University Identification (GUID) Password policy
	  ○ Mobile Device Encryption policy
	  ○ Computer Equipment Disposal policy
	  ○ Incident Handling policy.
	  ○ University of Glasgow Community Oral Health Section – Confidential Data Security Protocol.

5.3 Consent decisions

Consent will not be obtained as the research project will only retrospectively analyse secondary data that has been pseudonymised.

6 Accuracy of the information

6.1 Agreed steps to ensure the accuracy of any data shared.

- All parties sharing data under this agreement are responsible for the quality of the data they are sharing.
- Before sharing data, NES will check that the information being shared with the University of Glasgow is accurate and up to date to the best of their knowledge.

6.2 Agreed arrangements for any challenges to the accuracy of information

- If a complaint is received about the accuracy of personal data which affects the datasets shared with partners in this agreement, an updated replacement dataset will be communicated to the partners. The partners will replace the out of date data with the revised data.
- Partners are independently responsible for ensuring processes are in place to allow individuals to challenge the accuracy of information.

7 Data retention

7.1 Retention periods and purpose

NES:

- Partners to this agreement undertake that information shared under this agreement will only be used for the specific purpose for which it was shared, in line with this agreement. It must not be shared for any other purpose outside of this agreement.
- In each case, the originating organisation remains the primary information owner and record keeper for the information that is shared.
- The retention period for the information held by NES will be in line with NES policies and procedures and the NHS Scotland Code of Practice for Records Management.
- The University of Glasgow will not release the personal data shared within this agreement to any third party without obtaining the express written authority of NES. Outputs from the aggregated pseudonymised data will be published as part of the PhD thesis, academic papers and reports and presented to academic forums.

University of Glasgow

- Ethics approval stipulates that the data must be stored for 10 years after the end of the research project in accordance with Glasgow University's code of good practice in research.

7.2 Secure disposal of information

NES:

- The following destruction processes will be used when the information is no longer required:
  - Confidentially and securely destroyed in line with local NES policies and procedures.
- Electronic files will be data cleansed on an annual basis in line with NES policies and procedures; During the annual data cleansing process information held will be audited and deleted if no longer required to maintain EUGDPR compliance.
- Microsoft uses best practice procedures and a wiping solution that is NIST 800-88 (National Institute of Standards & Technology Special Publication 800-88, Guidelines for Media Sanitization) compliant. The appropriate means of disposal is determined by the asset type. Records of the destruction are retained and audited through the ISO process. All Windows Azure services utilize approved media storage and disposal management services.

University of Glasgow:

- When the data are no longer required it will be confidentially and securely destroyed in line with University of Glasgow policies and procedures.
- Personal identifiable data will be destroyed as soon as it is no longer required for the purposes of this project.
- NES will be notified of the specific details on how this will be performed and confirmed as soon as a pending enquiry with the UoG IT services has been fulfilled.

8 The rights of individuals

8.1 Subject access request

Under the General Data Protection Regulation / UK Data Protection Act 2018 a data subject (or authorised individuals acting on their behalf) has the right to make a Subject Access Request and to receive a copy of the personal data relating to them which is processed by an organisation. Dealing with such requests is the responsibility of each individual data controller. Communication must take place speedily to ensure the request is processed within the statutory one-month time period.

8.2 Freedom of Information (Scotland) Act – Information Requests

All the Parties are Scottish public authorities for purposes of the Freedom of Information (Scotland) Act 2002 and must respond to any request for recorded information made to them in a permanent form (such as letter or email). This would include an obligation to respond to requests about information sharing practices and procedures such as the arrangements under this Protocol. It should be noted that the actual personal information exchanged between the Parties will, in almost every case, itself be exempt from disclosure under the freedom of information legislation.

Any request for information submitted to either organisation will be processed under the organisations existing FOISA handling procedures, passing up through the organisations internal review process where appropriate.

8.3 Further Rights pf Data Subjects

8.3.1 Art. 15 GDPR Right of access by the data subject

1. The data subject shall have the right to obtain from the controller confirmation as to whether or not personal data concerning him or her are being processed, and, where that is the case, access to the personal data and the following information:

    1. the purposes of the processing;
    2. the categories of personal data concerned;
    3. the recipients or categories of recipient to whom the personal data have been or will be disclosed, in particular recipients in third countries or international organisations;
    4. where possible, the envisaged period for which the personal data will be stored, or, if not possible, the criteria used to determine that period;
    5. the existence of the right to request from the controller rectification or erasure of personal data or restriction of processing of personal data concerning the data subject or to object to such processing;
    6. the right to lodge a complaint with a supervisory authority;
    7. where the personal data are not collected from the data subject, any available information as to their source;
    8. the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.

2. Where personal data are transferred to a third country or to an international organisation, the data subject shall have the right to be informed of the appropriate safeguards pursuant to Article 46 relating to the transfer.

3. The controller shall provide a copy of the personal data undergoing processing. For any further copies requested by the data subject, the controller may charge a reasonable fee based on administrative costs. Where the data subject makes the request by electronic means, and unless otherwise requested by the data subject, the information shall be provided in a commonly used electronic form.

4. The right to obtain a copy referred to in paragraph 3 shall not adversely affect the rights and freedoms of others.

8.3.2 Art. 16 GDPR Right to rectification

The data subject shall have the right to obtain from the controller without undue delay the rectification of inaccurate personal data concerning him or her. Taking into account the purposes of the processing, the data subject shall have the right to have incomplete personal data completed, including by means of providing a supplementary statement.

8.3.3 Art. 21 GDPR Right to object

1. The data subject shall have the right to object, on grounds relating to his or her particular situation, at any time to processing of personal data concerning him or her which is based on point (e) or (f) of Article 6(1), including profiling based on those provisions. The controller shall no longer process the personal data unless the controller demonstrates compelling legitimate grounds for the processing which override the interests, rights and freedoms of the data subject or for the establishment, exercise or defense of legal claims.

2. Where personal data are processed for direct marketing purposes, the data subject shall have the right to object at any time to processing of personal data concerning him or her for such marketing, which includes profiling to the extent that it is related to such direct marketing.
3. Where the data subject objects to processing for direct marketing purposes, the personal data shall no longer be processed for such purposes.
4. At the latest at the time of the first communication with the data subject, the right referred to in paragraphs 1 and 2 shall be explicitly brought to the attention of the data subject and shall be presented clearly and separately from any other information.
5. In the context of the use of information society services, and notwithstanding Directive 2002/58/EC, the data subject may exercise his or her right to object by automated means using technical specifications.
6. Where personal data are processed for scientific or historical research purposes or statistical purposes pursuant to Article 89(1), the data subject, on grounds relating to his or her particular situation, shall have the right to object to processing of personal data concerning him or her, unless the processing is necessary for the performance of a task carried out for reasons of public interest.

8.4 Direct Marketing

Direct marketing is not involved in this agreement.

8.5 Automated decisions

No automated decisions are involved in this agreement – in the context of this agreement "Automated decisions" refer to decisions made using shared information with no human intervention.

9 Security

NES:

The security measures put in place within NES ensure that:

- Only authorised individuals can access, alter, disclose or destroy data. This is achieved through the following work instructions, policies and procedures (Appendix 1): NES Information Security Policy, NES Information Governance Policy;
- Authorised individuals act only within the scope of their authority. This is achieved through the following work instructions, policies and procedures - NES Information Security Policy, NES Information Governance Policy;
- If personal data is accidentally lost, altered or destroyed, it can be recovered to prevent any damage or distress to the individuals concerned. This is achieved through the following work instructions, policies and procedures - NES Information Security Policy, NES Information Governance Policy;
- Breaches of security leading to accidental, unlawful destruction, loss, alteration, unauthorised disclosure of, or access to, personal data transmitted, stored, or otherwise processed must be reported within 72 hours of the breach being identified in line with each partner organisations' incident reporting procedures and EU GDPR regulations;
- Significant data breaches involving personal information provided by partners under this DSA should be notified to the partner that originally provided the information;

- All signatories must have appropriate technical and organisational measures in place to ensure that any personal data shared between partners is handled and processed in accordance with the requirements of the General Data Protection Regulation/UK Data Protection Act 2018, Privacy and Electronic Communication Regulations (PECR) as well as ePrivacy and EU GDPR when they become enforceable law.

UoG:

The UoG security measures ensure that:

- Only authorised individuals can access, review, analysis, alter, disclose or destroy data. This is achieved through the following protocols and policies: UoG Regulations for the Use of University ICT Systems and Facilities, UoG Information Security: Confidential Data Policy, Glasgow University Identification (GUID) Password Policy, UoG Mobile Device Encryption Policy, UoG Network Service, Systems and Data Communications Monitoring Policy, UoG Guidelines and Procedures for Blocking Network Access and UoG Community Oral Health Section – Confidential Data Security Protocol;
- Data access activity is regularly monitored and audited as per the UoG Network Service, Systems and Data Communications Monitoring policy and the UoG Community Oral Health Section – Confidential Data Security Protocol;
- Data are stored in a secure physical and technical environment, as determined by the UoG Information Security Policy, UoG Network Connection Policy and UoG Antivirus and Firewalls Policy and UoG Community Oral Health Section – Confidential Data Security Protocol;
- There is no unauthorised copying of data (UoG Regulations for the Use of University ICT Systems and Facilities);
- Data breaches or information security incidents are reported and handled in a timely, structured and appropriate manner (UoG Incident Handling Policy);
- Records will not be retained for longer than necessary (in line with the EU GDPR).
- Following completion of all study analyses, data will be transferred to the University Records Centre for archiving for a period of 10-years as per the UoG Ethics Code of Good Practice.. NOTE: Disclosive information will not be stored as part of the archived data.

The security controls applicable by each organisation will be:		Jointly agreed between the parties
	X	Independently decided by each party

10 International transfers of personal data

Personal data shared in line with this agreement will be transferred to		EEA countries only
		Out with EEA
	X	Will not be transferred outside the UK

11 Implementation of the data sharing agreement

October 2018

11.1 Dates when information sharing commences/ends

Start: October 2018
End: January 2022

11.2 Training and communications

NES:

- All NHSS staff must complete mandatory safe information handling training via Learnpro or equivalent via local Health Board training packages.
- NES staff adhere to NHS Education for Scotland Confidentially and Information Governance policies and procedures.

University of Glasgow:

- Jamie Dickie, Andrea Sherriff and third-party data analyst have completed Medical Research Council (MRC) information governance training. These were completed before the introduction of the GDPR. The MRC course has since been withdrawn to be updated to incorporate the introduction of the GDPR. Once it becomes available again, the researchers will complete the updated course.

11.3 Information sharing instructions and security controls

All signatories must have appropriate technical and organisational measures in place to ensure that any personal data shared between partners is handled and processed in accordance with the requirements of the Data Protection Act 2018, EU GDPR, and Privacy and Electronic Communication Regulations (PECR).

11.4 Publication and transparency

[Indicate how and when the information sharing agreement (ISA) will be published (or state the security or other reasons that would prevent this.]

- The data collected will not be published other than anonymised compliance reports and presentations as part of the PhD thesis, to the academic community and in academic journals
- This agreement is available on request
- NHS Education for Scotland Privacy statement can be found on the NES website

12 Sign-off and responsibilities

12.1 Name of accountable officer(s)

[Insert name of all Accountable Officers (e.g. Chief Executive Officer) and if it exists Senior Information Risk Owner.]

Accountable Officer Name	Post title	Organisation
Caroline Lamb	Chief Executive	NHS Education for Scotland

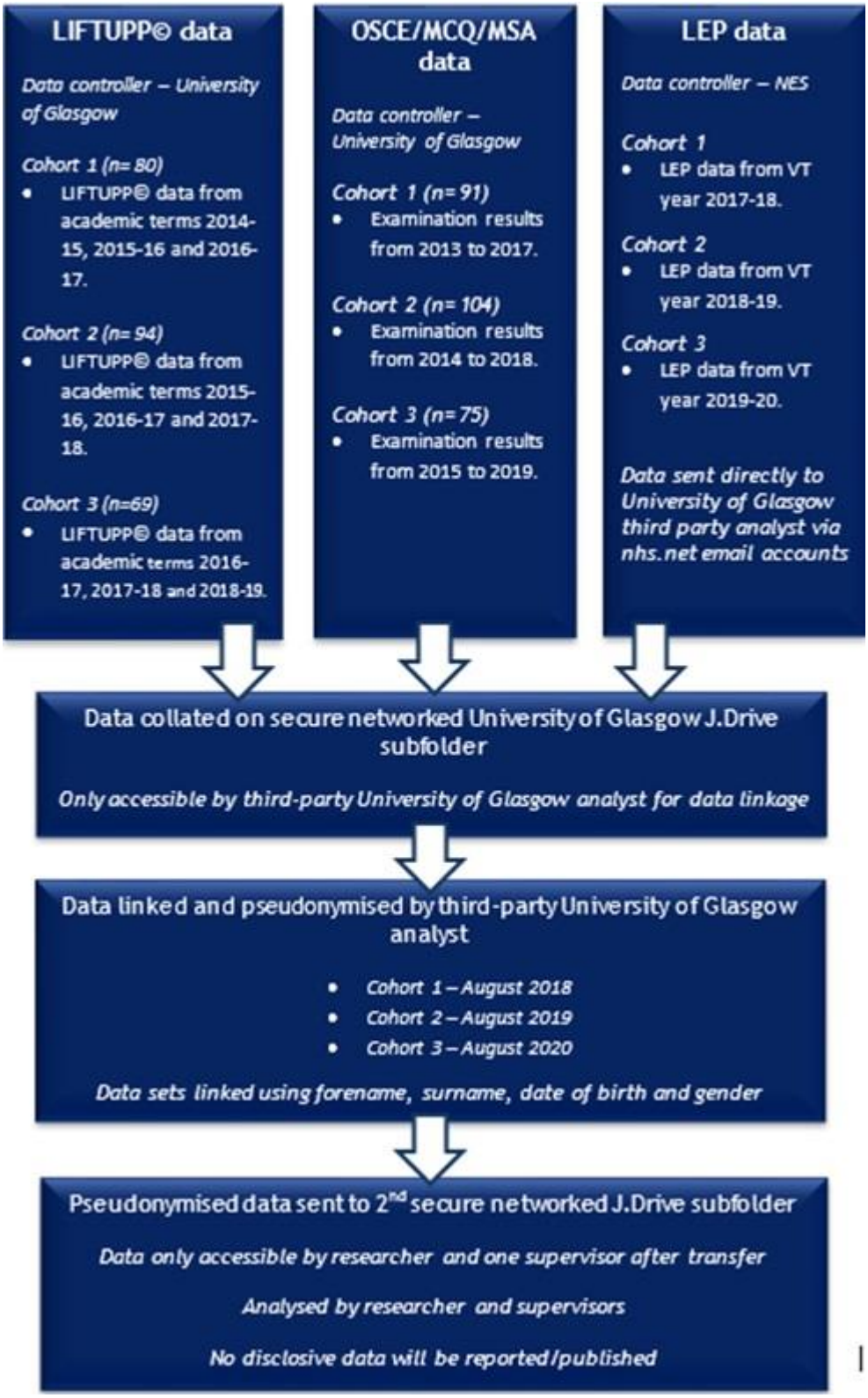
Senior Information Risk Owner Name	Post title	Organisation
Christopher Wroath	Digital Director/Senior Information Risk Own (SIRO)	NHS Education for Scotland

12.2 Lead practitioner

[Insert the names of the lead Information Governance (IG) / Data Protection Officer who have been tasked with preparing and completing the ISA on behalf of the organisation(s) participating.]

Lead IG Practitioner Name	Post title	Organisation
Tracey Gill	Information Governance & Security Lead / DPO	NHS Education for Scotland

Updated data flow chart for inclusion of third cohort (Vocational training year 2019/20) data



vii) Email confirmation: Review of University of Glasgow/NHS Education for Scotland data sharing agreement for inclusion of 2019/20 Vocational Dental Practitioner Longitudinal Evaluation of Performance data

From: Patrick Maitland-Cullen <Patrick.Maitland-Cullen@nes.scot.nhs.uk>
Sent: 30 July 2020 12:15
To: Jamie Dickie <Jamie.Dickie@glasgow.ac.uk>; Tracey Gill <tracey.gill@nes.scot.nhs.uk>
Subject: RE: [External] 2019/20 VDP LEP data

Hi Jamie

The existing DSA says in Section 11 that the sharing covered runs up to 2022.

I don't think there is any significant change to what was agreed. We have simply made a clarification of something implicit in the 2018 agreement, in my view.

So long as Tracey is happy on our side, and your IG person too, I can't see grounds for a substantive amendment and signed agreement to that.

Kind regards

Patrick.

Patrick Maitland-Cullen
Manager - Information Governance

NHS Education for Scotland
102 Westport
West Port
Edinburgh
EH3 9DN

Tel: 0131 656 4299 ext 4299

Email: patrick.maitland-cullen@nes.scot.nhs.uk

From: Patrick Maitland-Cullen <Patrick.Maitland-Cullen@nes.scot.nhs.uk> Sent: 30 July 2020 11:40 To: Jamie Dickie <Jamie.Dickie@glasgow.ac.uk>; Tracey Gill <tracey.gill@nes.scot.nhs.uk> Cc: Freedom of Information and Data Protection <foidp@nes.scot.nhs.uk> Subject: RE: [External] 2019/20 VDP LEP data

Hi Jamie

That is extremely helpful – especially the updated flow chart which unpacks what was already agreed.

I was a bit slow on the uptake regarding a third cohort.

In light of your response, I've no further comments on the final round of sharing. These emails can be kept on file as a record of review and DSA update.

I hope the study goes well.

Kind regards

Patrick.

Patrick Maitland-Cullen
Manager - Information Governance

NHS Education for Scotland
102 Westport
West Port
Edinburgh
EH3 9DN

Tel: 0131 656 4299 ext 4299

Email: patrick.maitland-cullen@nes.scot.nhs.uk

viii) Data protection impact assessment – Focus groups

Data Protection Impact Assessment: List of Requirements

Before submitting your DPIA template to the DP Office, please ensure that you completed and addressed all relevant points below. **The DP Office will not review your DPIA unless you can demonstrate engagement with or reference to the checklist and its attending documents and requirements.**

☒ If you are claiming that your data is anonymous, are there any potential data linkages that would allow someone to identify your data subjects? Note that simply removing a name does not constitute anonymisation. Have you considered the impact of other potential identifiers e.g., you are studying individuals with an uncommon medical condition and also working with gender, age, and location data or other factors that narrow your population and potentially lead to identification?

☒ Have you determined if your data truly anonymous or is it pseudonymous? If you hold an identifier key(s) that would allow you or another party to identify your masked data then it is pseudonymous and therefore must be treated as personal data.

☒ Have you determined whether you/the University is a data controller or a data processor for this project?

☒ Have you included specific details regarding your data flow – where is the data coming from, who are you sharing access with, where is it going upon project completion?

☒ How will you gather, store and access the data? Will you require third parties to assist you? Have you detailed this within the DPIA?

☒ If you are sharing your personal data outwith the University, is a data sharing agreement in place or do you need one?

☒ Have you completed a privacy notice to inform data subjects on the intended use of their personal data?

☒ Have you completed the University's online Data Protection and/or Information Security trainings? (These trainings are mandatory for staff.)

☒ Have you completed a research data management plan? Have you reviewed the DMP and DPIA Workflow chart?

☒ Have you considered ways to reduce potential risk and have you demonstrated practical compliance, including:

- pseudonymisation
- data minimisation
- storage limitation
- access restrictions
- technical solutions (e.g. encryption)

- organisational measures (e.g. policies, procedures and workflows to comply with GDPR requirements)

Step 1: Identify the need for a DPIA

Explain broadly what the service/project aims to achieve and what type of processing it involves. You may find it helpful to refer or link to other documents, such as a project proposal. Summarise why you identified the need for a DPIA.
The aim of the PhD project is to investigate the validity of longitudinal data in the assessment of undergraduate dental students. Part of this investigation will involve conducting focus group discussions with key stakeholders on how assessment within dental education can be enhanced based on statistical analyses of a longitudinal assessment system (known as LIFTUPP©) and more established assessment methods. Recruitment for the focus groups will require personal data - such as participant names, email address and either job titles or student year of study - to be obtained. Focus groups were originally going to be conducted in person, However, due to the COVID-19 outbreak, it is currently not possible to conduct face-to-face focus groups. As a contingency, focus groups meetings will conducted online using Microsoft Teams video conferencing technology. Audio and video recordings of the conversations between key stakeholders will be made using the recording features available on Microsoft Teams. These recordings will automatically be uploaded to Microsoft Stream and will be transcribed using the transcription feature available within this platform. Audio and video recordings of human participants are regarded as personal data under GDPR. Therefore, there is a need for a DPIA since the research will require and record personal data as part of the recruitment and focus group processes respectively.

Step 2: Describe the processing

2.1 Describe the nature of the processing: how will you collect, use, store and delete data? What is the source of the data? Will you be sharing data with anyone? You might find it useful to refer to a flow diagram or another way of describing data flows. What types of processing identified as likely high risk are involved?

Key stakeholders will be emailed by the PhD researcher inviting them to participate in the study. Those who wish to participate will be asked to respond and provide their name, email address and either their job title or year of study. A list of details from volunteering respondents will be complied and stored in a folder on the University of Glasgow's secure networked J-Drive. This folder will only accessible by the PhD researcher and one of his supervisory team (Dr Michael McEwan). Upon completion of the recruitment process, the original email responses will be deleted from the PhD researcher's University of Glasgow email account.

Audio and video recordings made via Microsoft Teams online video conferencing software will be produced using the platform's recording feature. The recordings will be automatically uploaded to Microsoft Stream
– an online cloud storage system. Both Microsoft Teams and Microsoft Stream have been institutional approved by the University of Glasgow. Both platforms have been integrated into the University of Glasgow's Office 365 package, which means they have enterprise-grade security and the compliance required (https://docs.microsoft.com/en-us/microsoftteams/security-compliance-overview

A backup copy of the audio will be made using an Olympus digital voice recorder placed next to the PhD researcher's computer speakers.

The recordings uploaded to Microsoft Stream will then be transcribed using the automated transcription feature available within the platform. Once the accuracy of the transcriptions has been reviewed by the PhD researcher, copies of the transcriptions will be saved to the University of Glasgow's secure networked J-Drive. The original audio and video recordings will be deleted from Microsoft Stream as will the backup audio recording from the digital voice recorder.

Personal identifiable data (such as names and job titles) will be not be included in the transcripts. Instead, numerical codes will be assigned to the participants to initially provide pseudonymisation. The transcripts will be analysed by the PhD researcher and the findings will be included in a thesis report which will be submitted to the University of Glasgow for the award of a PhD degree. Selected direct quotations from the anonymised focus group participants will be included within the main text of the thesis as well as subsequent academic publications.

No identifiable data will be included in the thesis report and subsequent academic publications. Instead, a summary of roles covered will be presented to provide. context on the "make up" of the focus groups.

Pseudonymised transcripts of the focus group discussions will be included as appendices of the thesis report.

The audio/video recordings will not be seen/heard by anyone other than the PhD researcher and his supervisory team.

2.2 Describe the scope of the processing: what is the nature of the data, and does it include special category or criminal offence data? How much data will you be collecting and using? How often? How long will you keep it? How many individuals are affected? What geographical area does it cover?

Personal data (names, email address, job title/BDS year of study) will be required for the focus group recruitment process. This data will only be collected from those who volunteer to take part in the study. Focus group discussion data will include the key stakeholders' thoughts and opinions on how assessment within dental education can be enhanced based on the statistical analysis of multiple assessments methods (NOTE: These analyses were completed earlier in the study by the PhD researcher).

No special category or criminal offence data will be collected.

Two focus groups will be arranged. One will consist of current undergraduate BDS (dental) students and recently qualified (2017-2020) dentists who have been assessed with the LIFTUPP© system. The other focus group will consist of faculty - from both the University of Glasgow Dental School and other UK Dental School's - and representatives from NHS Education for Scotland (NES). Each focus group will contain 6–10 participants, meaning up to 12-20 individuals will be affected depending on final recruitment numbers.

Focus groups are expected to last between 1.5 and 2-hours. This will equate to approximately 360MB of data per focus group.

The transcripts produced from the focus groups will be kept for 10-years in accordance with the University of Glasgow's Code of Good Practice in Research. (http://www.gla.ac.uk/media/media_227599_en.pdf).

2.3 Describe the context of the processing: what is the nature of your relationship with the individuals? How much control will they have? Would they expect you to use their data in this way? Do they include children or other vulnerable groups? Are there prior concerns over this type of processing or security flaws? Is it novel in any way? What is the current state of technology in this area? Are there any current issues of public concern that you should factor in? Are you signed up to any approved code of conduct or certification scheme (once any have been approved)?

The PhD researcher is employed as a Clinical Lecturer at the University of Glasgow dental school. Therefore, there is a dependent relationship between the PhD researcher and the undergraduate dental students who will be invited to participate in the study.

There is also a professional working relationship between the PhD researcher and recently qualified dentists, dental school faculty and NES representatives who will be invited to participate.

All individuals invited to take part in the study will notified that participation is entirely voluntary. Those who agree to participate will also be able to withdraw from the study up until the focus group discussions have been transcribed and the original audio/video recordings have been deleted. They will no longer be able to withdraw beyond this point as the transcriptions will have been anonymised and it will not be possible to determine what was said by individual participants. All potential participants will be informed about this using an information leaflet (written in plain language) which will be provided as part of the invitation to take part in the study.

The information leaflet will also provide details on the purpose of the research project, its methods, what data will be required and how their data will be used, anonymised and stored.

In addition, those who volunteer to participant will be provided with a privacy notice which will reiterate how their data are to be used.

Regarding data security, Microsoft Teams and Microsoft Stream have been institutional approved by the University of Glasgow. Both platforms have been integrated into the University of Glasgow's Office 365 package, which means they have enterprise-grade security and the compliance required (https://docs.microsoft.com/en-us/microsoftteams/security-compliance-overview

No children or vulnerable groups will be included in the study.

2.4 Describe the purposes of the processing: what do you want to achieve? What is the intended effect on individuals? What are the benefits of the processing for you, and more broadly?

The purpose of the data processing is to obtain opinions from key stakeholders on how findings from the analyses of multiple assessment methods can be used to enhance assessment within dental education.

Focus groups discussions will not influence undergraduate or career progression of the volunteering participants. However, individuals who wish to participant will need to commit 1.5-2 hours of their time to attend the focus groups.

The PhD researcher will personally benefit as they will use the findings from the data processing to form part of a thesis report which will be submitted to the University of Glasgow for the award of a PhD degree.

There are also wider benefits to the public. The data processing will contribute evidence on the use of longitudinal assessment in dental education and help make recommendations for future research. This will assist dental schools in reviewing their assessment methods to ensure they are adopting best practice in accordance with the available evidence. Ultimately this will safeguard patients by ensuring dental practitioners are appropriately trained and assessed before being permitted onto professional registers, thus preventing the public from being treated by individuals who are not fit to practise.

Furthermore, the unit cost to the UK public purse of training a dental student is at least £250,000. It is therefore essential that dental schools are confident of the utility of their assessment methods and that they represent good value for money.

Step 3: Consultation process

Consider how to consult with relevant stakeholders: describe when and how you will seek individuals' views – or justify why it's not appropriate to do so. Who else do you need to involve within your organisation? Do you need to ask your processors to assist? Do you plan to consult information security experts, or any other experts?

Key stakeholders will be invited to participate in the study via email by the PhD researcher 2-3 weeks before the focus groups are scheduled to take place. An information leaflet on the study will be attached to the invitation for participation. This information leaflet – written in plain language - will provide details on the purpose of the study, its methods and how participant data will be used, anonymised and stored. It will also invite key stakeholders to contact the PhD researcher with any questions, queries and/or concerns they may have irrespective if they wish to participate or not.

The University of Glasgow's data protection department has been consulted to ensure data collection, processing and storage are secure and compliant with the European Union's General Data Protection Regulations (GDPRs) respectfully.

Ethical approval for the study was previously granted by the University of Glasgow's MVLS ethics committee. However, following the COVID-19 outbreak, an amended application was resubmitted to request approval for focus groups to take place online using Microsoft Teams) video conferencing software as a contingency.

Step 4: Assess necessity and proportionality

Describe compliance and proportionality measures, in particular: what is your lawful basis for processing? Does the processing actually achieve your purpose? Is there another way to achieve the same outcome? How will you prevent function creep? How will you ensure data quality and data minimisation? What information will you give individuals? How will you help to support their rights? What measures do you take to ensure processors comply? How do you safeguard any international transfers?

The study protocol was previously subjected to the UK Information Commissioner's Office's (ICO's) "*three-part test*". The test concluded that, in accordance with GDPR, there was a legitimate interest for the study to progress since risk to the participants was minimal and the study would contribute towards improving future assessment within dental education, as the information gathered will help inform dental assessors and regulators on the use of longitudinal data as a method of assessment. Ultimately, this contributes to public protection by ensuring dental students are being assessed with the best, evidence-based methods available before their entry onto the professional registers (see Step 2.4 above).

Processing the data will help achieve the purpose of the study (outlined above in Step 2.4) as the opinions of key stakeholders will contribute qualitative evidence towards investigating the validity of longitudinal assessment in dental education.

A similar outcome could be achieved using one-to-one interviews. However, focus group discussions will promote exchanges of viewpoints and discussions between key stakeholders, which may generate data that may not have been captured by one-to-one interviews.

A topic guide will be used to prevent focus group conversations deviating from key areas of discussion. This will also help function creep, contribute to data quality and minimisation through ensuring data necessary for answering the study research questions are obtained.

The host of the meeting (i.e., the PhD researcher) will use Microsoft Teams settings to ensure no video footage is recorded until after participants have entered the online meeting room and choose to activate their video footage themselves. This also gives participants the option on whether they wish to appear on the video recording of the meeting or not.

As previously discussed in Step 2.3, key stakeholders invited for participation will be issued with an information leaflet will also provide details on the purpose of the research project, its methods, what data will be required and how their data will be used, anonymised and stored. Those who volunteer to participate will be given a privacy notice which will reiterate when, why, where and how their personal data are used by the University of Glasgow and the PhD researcher. The privacy notice will also inform participants of their rights and how to exercise them.

Privacy notices must be electronically signed as part of the written consent process for participation.

All key stakeholders will be informed that their participation is voluntary, and should they choose to participate, they can withdraw from the study up until the focus group transcriptions are completed. Once transcribed and checked by the PhD researcher, the audio and/or video recording(s) of the focus group discussions on Microsoft Steams and the voice recorded will be deleted. This will anonymise the transcripts since any references to participants (such as names and job titles) will be not be included within their content. Therefore, it will no longer be possible to determine what was said by each individual participant.

Step 5: Identify and assess risks

Describe the source of risk and nature of potential impact on individuals. Include associated compliance and corporate risks as necessary.	**Likelihood of harm**	**Severity of harm**	**Overall risk**
	Remote, possible or probable	Minimal, significant or severe	Low, medium or high
i. *Confidentiality* - Personal data (name, email address and, where appropriate, job title or BDS year of study) will be required as part of the focus group recruitment process. Audio and video recordings of human participants are also regarded as personal data under GDPR regulations.	Remote	Minimal	Low
ii. *Data security and non-disclosive information* – All personal data (Names, email addressed, job title/BDS year of study and focus group recordings) need to be processed, stored and protected appropriately as they are all personal data.	Possible	Significant	Medium
iii. *Privacy* - If online focus groups are conducted, some participants may join the discussion from their own homes and therefore their privacy may be compromised if video communication is used. Other individual's privacy may be compromised due to unintended viewers.	Possible	Significant	Medium
iv. *Dependent relationships* - Undergraduate dental students will be invited to participate in the focus groups. Invitations for participation will be sent by the PhD researcher, who will also chair the focus group discussions. Since the PhD researcher is a teacher known to the students at the University of Glasgow Dental School, some students may feel to be, or be construed to be, under obligation to volunteer for participation in the project.	Possible	Minimal	Low

Step 6: Identify measures to reduce risk

Identify additional measures you could take to reduce or eliminate risks identified as medium or high risk in step 5				
Risk	**Options to reduce or eliminate risk**	**Effect on risk**	**Residual risk**	**Measure approved**
		Eliminated, reduced or accepted	Low, medium or high	Yes/no
i) Confidentiality	Participant focus group responses will be anonymised during the transcriptions of the recorded discussions. Numeric codes will be used instead of any disclosive information (such as names and job title). Whilst the PhD researcher and his supervisory team will make every effort to ensure participation is anonymised, due to the nature of focus groups, this assurance cannot be made on behalf of the other participants. The study information leaflet will inform all individuals invited for participation of this.	Reduced	Low	
ii) Data security and non-disclosive information	Personal data obtained as part of the recruitment process (i.e., participant names, email address and either job titles or student year of study) will be stored on the University of Glasgow's secure networked J-Drive and only be accessible to the PhD student and his supervisors. Audio and video recordings and transcripts from focus groups with key stakeholders will only be accessible to the PhD student and his project supervisors. Transcriptions will contain no identifiable data (e.g., names). Instead, participants will be assigned an alphanumeric code that will only be known to the researchers. Once the transcriptions have been completed, checked and verified, the audio and video recordings of the focus groups will be deleted. The transcriptions will be stored on the University of Glasgow's secure	Reduced	Medium	

	networked J-Drive and only be accessible to the PhD student and his supervisors. Deletion of the original audio and video recordings will provide greater anonymity to the transcripts			
iii)Privacy	Participants could also use the "virtual background" feature which can increase privacy by blocking out the background of the room from which they are broadcasting. The host of the meeting (i.e., the PhD researcher) will use Microsoft Team's settings to ensure no video footage can be recording can be made until after participants have entered the online meeting room and then choose to activate their video footage themselves. This also gives participants the option on whether they wish to appear on the video recording of the meeting or not.	Reduced	Medium	
iv)Dependent relationships	Undergraduate students invited for participation will be given assurances that: • Participation is entirely voluntary; • No prejudice will be set between those who chose to participate and those who did not; • Participants are free to withdraw from the study up to (and including) the date on which focus group conversations are transcribed (see Step 4 above); • Responses will not impact academic records and/or student progress; • Data gathered from the focus groups will only be used for research purposes; • Responses will be anonymised as part of the transcription process. These assurances will be reiterated to those who express an interest in taking part in the study through one-to-one discussion prior to signing a consent form for participation.	Accepted	Low	

Step 7: Data Protection & FOI Office recommendations

DP & FOI Office advice provided:		DP & FOI Office should advise on compliance and step 6 measures
Summary of DPO advice: No formal DPO advice required following email contact with University of Glasgow data protection department.		

Step 8: Sign off and record outcomes

(To be completed by the PI/Project Lead)

Item	**Name/date**	**Notes**
Measures approved by:	Jamie Dickie	Integrate actions back into project plan, with date and responsibility for completion
Residual risks approved by:	Jamie Dickie	The ICO must be consulted where high risks are identified and cannot be mitigated.
DPO advice accepted or overruled by:		If overruled, you must explain your reasons
Comments:		
Consultation responses reviewed by:		If your decision departs from individuals' views, you must explain your reasons

Comments:		
This DPIA will be kept under review by:		The DPO should also review ongoing compliance with DPIA and a copy of the most recent version should be sent to the DP&FOI Office.

ix) Data management plan

University of Glasgow
Data Management Plan template for PGR students

1. Overview	
Student name	Jamie Dickie
Supervisor name	Kurt Naudi, Andrea Sherriff, Michael McEwan
Project title	Longitudinal assessment of undergraduate dental students: Building a validity argument
Funder & award number	NA
Project Summary	This PhD project aims to investigate longitudinal assessment in the undergraduate Bachelor of Dental Surgery (BDS) curriculum to establish its utility as an assessment tool and its association with postgraduate performance. It is aligned with the top priority by the Scottish Oral Health Research Collaboration (SOHRC) for dental education research. The study involves linking and secondary analysis of existing undergraduate student assessment data from an electronic longitudinal system (known as LIFTUPP©), summative examinations used by the University of Glasgow Dental School and postgraduate longitudinal assessment data from Longitudinal Evaluations of Performance (LEPs), which are used in Scottish postgraduate dental vocational training (VT) schemes. Data will be sourced from three cohorts (the graduating classes of 2017, 2018 and 2019), before being linked and pseudonymised by a third-party data analyst (based at the University of Glasgow). The pseudonymised data will then be made available to the researchers for analysis. Processing these data will allow Glasgow Dental School to begin evaluating the validity of the LIFTUPP© assessment system in determining the development of competent clinical practice of undergraduate students. Following statistical analyss of student assessment data, the findings will be presented to key stakeholders within dental education as part of focus group discussions. Data produced by these discussions will be used to recommend how assessment within dental education could be enhanced.

2. Data
What types of data will be collected or created?
• All student assessment data are secondary data, having been collated for other purposes. The following are quantitative in nature ▪ Undergraduate longitudinal clinical performance assessment (LIFTUPP©) data;

- Postgraduate longitudinal clinical performance assessment (LEP) data;
- Undergraduate professional examination results.

- Qualitative data will be created and collected through focus group discussions with key stakeholders in dental education. These data will be audibly recorded and transcribed in verbatim.

UPDATE (JUNE 2020):
Due to the COVID-19 outbreak, it may not be possible to conduct focus groups with key stakeholders in person should social distancing restrictions remain in place by September 2020. As a contingency, focus group discussions could take place online via video conference software (such as ZOOM or Microsoft Teams). If this approach is to be adopted, video and audio recordings of the discussions and the subsequent transcriptions will be made using recording and automated transcription features available within ZOOM or Microsoft Teams.

UPDATE (NOVEMBER 2020):
It will not be possible to conduct focus groups with key stakeholders in person due to social distancing restrictions. Focus group discussions will now take place online via Microsoft Teams. Audio and video recordings of the discussions will be made using recording available within Microsoft Teams. These recordings will be automatically transferred onto Microsoft Stream and transcripts will be produced using the automated transcription feature available within Microsoft Stream.

What formats will you use?

- Undergraduate longitudinal clinical performance assessment (LIFTUPP©) data – Microsoft Excel spreadsheets (.xls) and Stata statistical software (.dta)
- Postgraduate longitudinal clinical performance assessment data (LEP) - Microsoft Excel spreadsheets (.xls) and Stata statistical software (.dta)
- Undergraduate professional examination results - Microsoft Excel spreadsheets (.xls) and Stata statistical software (.dta)
- Focus group session audio recordings (MP4)
- Transcripts of focus group sessions (Unconfirmed at present – likely to be Microsoft Word (.doc)

UPDATE (JUNE 2020)

- Video and audio recordings of online focus groups (MP4)
- Transcriptions made via ZOOM or Microsoft Teams (.txt)

UPDATE (NOVEMBER 2020)

- Video and audio recordings of online focus groups (MP4)
- Transcriptions made via Microsoft Stream (.doc)

How much data will you collect?

- Undergraduate longitudinal clinical performance assessment (LIFTUPP©) data – 136 MB (2 cohorts x 68 MB)
- Postgraduate longitudinal clinical performance assessment (LEP) data – 4 MB (2 cohorts x 2 MB)
- Graphs, tables and working files based on student assessment data – Approximately 600 MB
- Undergraduate professional examination results – 300 KB (2 cohorts x 150 KB)

- Focus group session recordings – Approximately 460 MB (Amendment (June 2020) – this refers to audio recordings of face-to-face focus groups)
- Transcripts of focus group sessions – *Unknown at present*

UPDATE (JUNE 2020)

- Video and audio recordings of online focus groups – Approximately 360 MB (2 x 180 MB)

UPDATE (NOVEMBER 2020)

- Video and audio recordings of online focus groups – Approximately 360 MB (2 x 180 MB)
- Transcripts of focus group discussions – 2 x 100 KB

3. Documentation

How will the data be documented and described?

Appropriate documentation and descriptions for all original quantitative data files to be used within this project have previously been fixed since they secondary data.

Audio recordings of focus group discussions will be made and transcribed in verbatim. Transcriptions will be documented with the following metadata:

- Title
- Project abstract
- Date of focus group
- Location of focus group
- Details on the roles of key stakeholder focus group participants (e.g., student, senior dental school faculty, clinic teacher/supervisor etc.)
- Creator
- Format
- File type
- Language
- Methodology used to generate data (including details of equipment and software used)
- Key words on subject and data content

Update (JUNE 2020):
Transcriptions made via ZOOM or Microsoft Teams will also be documented using the above metadata.

Update (NOVEMBER 2020):
Transcriptions made Microsoft Stream will also be documented using the above metadata.

Are there any standards for this in your field of research?

This project will follow the data documentation recommendations outlined by the UK Data Archive (UKDA). The UKDA recommendations are based on the standards provided by the Data Documentation Initiative (DDI), which are widely used, international standards for describing data from the social, behavioural, and economic sciences.

The EU's General Data Protection Regulations (GDPR) will also be adhered to.

4. Ethics and Intellectual Property
Who owns the data in your project?
All undergraduate assessment data (longitudinal and examination results) are owned by the University of Glasgow. Postgraduate longitudinal assessment data are owned by NHS Education for Scotland (NES).
Detail any ethical, legal or commercial considerations relating to your research data
The main ethical considerations relating to the research data are i) confidentiality, ii) data security and non-disclosive information, iii) two of the researchers' dual roles as assessors and researchers and iv) dependent relationships. i) *Confidentiality* - Participant assessment data (at both undergraduate and postgraduate levels) need to be linked for comparisons to be made and to allow narratives on the development of competent dental practice to be established. For the datasets to be linked correctly, personal identifying factors (foreman, surname, date of birth and gender) need to be made available. Personal data (name, email address and, where appropriate, job title or BDS year of study) will be required as part of the focus group recruitment process. Audio and video recordings of human participants are also regarded as personal data under GDPR regulations. ii) *Data security and non-disclosive information* – For the quantitative component of the study, identifying data (forename, surname, date of birth and gender) are only to be used by the third-party analyst to link the assessment data sets. Raw data sets containing this information are to be transferred to the third-party analyst by their data controllers (University of Glasgow and NES) via secure nhs.net email accounts. Once data linkage has been completed, personal identifiers will be removed to produce a pseudonymised data set. The linked pseudonymised data set will be stored on the University of Glasgow's secure networked J-Drive and will only be assessible to the researcher and one of supervisory team. For the qualitative component of the study, names, email addressed, job title/BDS year of study and focus group recordings) need to be processed, stored and protected appropriately as they all personal data. iii) *Dual roles of two researchers* - The researcher and one of supervisory team could recognise LIFTUPP© and undergraduate examination performance data patterns as their academic responsibilities involve analysing these data sets to contribute to decisions on student progress. *iv)* *Dependent relationships* - Undergraduate dental students will be invited to participate in focus groups as part of the study. The discussions held in the focus groups will be

chaired by the lead researcher, who is also a teacher known to the students at the University of Glasgow Dental School. Therefore, some students may feel to be, or be construed to be, under obligation to volunteer for participation in the project.

How will these concerns be dealt with?

i) *Confidentiality* – To preserve confidentiality, the researcher and project supervisors will not be privy to any disclosive participant information. Instead a third-party University of Glasgow analyst, who had no links to the study, will perform the linkage and provide the study team with a pseudonymised linked data-set having removed any identifying data.

Participants' focus group responses will be anonymised during the transcription of the recorded discussions. Numeric codes will be used instead of any disclosive information (such as names and job title).

ii) *Data security and non-disclosive information* - Since personal data are required to initially link the various quantitative datasets without participant consent it was important to ensure that the study was compliant with the European Union's (EU's) General Data Protection Regulations (GDPR).

The study protocol was subjected to the UK Information Commissioner's Office's (ICO's) "*three-part test*", which determines if there is a lawful basis for processing personal data. This test concluded that, in accordance with GDPR, there was a legitimate interest for the study to progress since risk to the participants was minimal and the study would contribute towards improving future assessment within dental education, as the information gathered will help inform dental assessors and regulators on the use of longitudinal data as a method of assessment.

Further details on how both quantitative and qualitative data will be securely stored, retained and shared are given in sections 5, 6 and 7 respectively.

iii) *Dual roles of two researchers* - Working with pseudonymised data will reduce the possibility of the researchers who are also assessors of undergraduate students being able to identify individuals' assessment data. LEP data will not identifiable since the researcher and project supervisors have no role in postgraduate VT assessment.

The size of the pseudonymised data set for this research project mitigates the risk of LIFTUPP© and undergraduate examination performance data patterns being recognised, as it is likely that only data extremes (i.e., exceptionally good or poor performance compared to peers) will be attributable to students previously known to the researcher and one of supervisory team. Despite this potential eventuality, student progress outcomes cannot be influenced since the assessment data will be analysed retrospectively.

iv) *Dependent relationships* - Focus group participants are to be provided with information leaflets which will explain (in plain language) the purpose of the study, its

methods and how their data would be used, anonymised and stored. They will also be given assurances that:

- Participation is entirely voluntary;
- No prejudice will be set between those who chose to participate and those who did not;
- Participants are free to withdraw from the study up to (and including) the date on which focus group conversations were transcribed (NOTE: Transcription of the focus group discussions will remove any references to participants, meaning they will no longer be able to withdraw after this process had been completed);
- Responses will not impact academic records and/or student progress;
- Data gathered from the focus groups will be used for research purposes;
- Responses will be anonymised as part of the transcription process.

These assurances will be reiterated to those who express an interest in taking part in the study through one-to-one discussion prior to signing a consent form for participation.

UPDATE (JUNE 2020):

Information on data security for both ZOOM and Microsoft Teams is currently being sought from the University of Glasgow data security services. This document will be updated once the appropriate data security information has been obtained.

A data protection and impact assessment (DPIA) is currently being completed to ensure online focus group methods are compliant with data security regulations and that the rights and interests of data subjects are protected.

UPDATE (November 2020):

Information on data security for Microsoft Teams and Stream has now been obtained from the University of Glasgow data security services. Both applications are part of the University of Glasgow's Office 365 package and therefore are covered by enterprise-grade security. (https://docs.microsoft.com/en-us/microsoftteams/security-compliance-overview)

A data protection and impact assessment (DPIA) has now been completed.

5. Storage and organisation
How will the data be named, organised and structured?
All data obtained and generated by this project will be stored on the University of Glasgow's secure networked J-Drive in a folder only accessible to the student and one of supervisory team.

Master files of the linked pseudonymised assessment data will be kept in a folder entitled "Linked Data". These master files are never to be edited. Copies of these files will be transferred into an additional subfolder entitled "Working data files". It is these copy/working files that will be edited as part of the data cleansing and analysis processes.

Any further copies made to the copy/working files will be given an updated version number within its file name.

Results created by data analysis will be saved as unique files. These name of these files will contain a brief description of their content and, if appropriate, a version number.

Focus group audio recordings and transcriptions will be stored in subfolder entitled "Focus groups". Audio recordings will be deleted from the recording devices (an Olympic digital voice recorder and the PhD student's iPhone) and J-Drive server once transcriptions have been completed.

UPDATE (JUNE 2020):
Should online focus groups be required, data storage will depend on which online video conferencing software is used.

If ZOOM was to be used, focus group audio and video recordings will be made locally (i.e., directly) to a folder on the University of Glasgow's secure networked J-Drive which is only accessible to the student and one of his supervisory team (Dr Andrea Sherriff). Once the transcriptions of the discussions have been created and checked, the recordings will be deleted from the J-Drive.

If Microsoft Teams as to be used, focus group audio and video recordings will be saved to Microsoft Stream – an online cloud service which is part of the University of Glasgow's Office365 package with enterprise-grade security. Once the transcriptions of the discussions have been created and checked, the recordings will be deleted from Microsoft Stream.

NOTE: Further data security details both ZOOM and Microsoft Teams are currently being sought from the University of Glasgow data security services. This document will be updated once the appropriate data security information has been obtained

UPDATE (NOVEMBER 2020):
Since Microsoft Teams is to be used for the focus group discussions, audio and video recordings will be saved to Microsoft Stream – an online cloud service which is part of the University of Glasgow's Office365 package with enterprise-grade security. Once the transcriptions of the discussions have been created and checked, the recordings will be deleted from Microsoft Stream.

How will the data be stored for the duration of the project?

Unprocessed, linked and pseudonymised, audio file and transcription data will be only stored on the University of Glasgow secure networked J-Drive.

UPDATE (JUNE 2020):
Audio and video recordings and transcriptions made via ZOOM will be only stored on the University of Glasgow secure networked J-Drive.

Audio and video recordings and transcriptions made via Microsoft Teams will be stored on the Microsoft Stream cloud – which is integrated with the University of Glasgow's Office365 package. Transcriptions produced from the recordings will be stored on the University of Glasgow secure networked J-Drive.

UPDATE (NOVEMBER 2020):
Transcriptions produced from the recordings will be stored on the University of Glasgow secure networked J-Drive.

How will the data be backed up during the project?

In accordance with the University of Glasgow's backups policy, data redundancy is implemented via multi-site server placement, server clustering, RAID disks, volume shadow copies, and tape backup.

Does access to the data need to be controlled for the duration of the project?

Yes.

Who has the right to access the data during the project?

Access to a J-Drive subfolder used to store the unprocessed assessment data will only be permitted to Dr Alex McMahon (the third-party University of Glasgow analyst) for data linkage and pseudonymisation of the data. A separate J-Drive subfolder for storing the linked pseudonymised assessment data, focus group audio files and transcriptions will only be accessible by Jamie Dickie (the PhD student) and Andrea Sherriff (one of the project supervisors). This means there will be a shared file space that can only be accessed by these two researchers. Access to all J-Drive subfolders folders is password protected.

UPDATE (JUNE 2020):
Audio and video recordings and transcriptions made via ZOOM will be stored in a folder on the University of Glasgow's secure networked J-Drive. This folder will only be accessible to the PhD researcher and one of his supervisor team (Dr Andrea Sherriff).

NOTE: Further details on accessibility to Microsoft Stream are currently being sought from the University of Glasgow data security services. This document will be updated once the appropriate data security information has been obtained.

UPDATE (NOVEMBER 2020):
The audio and video recordings uploaded to Microsoft Stream will only be accessible to the PhD researcher via their password protected University of Glasgow login.

The transcripts will be accessible to the PhD researcher and one of his supervisor team (Dr Andrea Sherriff) on the University of Glasgow's secure networked J-Drive.

6. Deposit and long-term preservation

Which data should be retained long-term?

All data used for this project will be retained long-term (10-years), except for the focus group audio files. These files will be deleted from the recording devices and J-Drive server once they have been transcribed.

UPDATE (JUNE 2020):
Audio and video recordings made via ZOOM or Microsoft Teams will be deleted from the J-Drive or Microsoft Stream (respectively) upon completion of the transcripts

The transcripts will be retained long-term (10-years).

UPDATE (NOVEMBER 2020):
Audio and video recordings will be deleted from Microsoft Stream upon completion of the transcripts

The transcripts will be still be retained long-term (10-years) (see below).

How long will data be retained for?

In accordance with the University of Glasgow Ethics code of good practice, data to be retained will be stored for **10-years** after the end of the research project.

Where will the data be archived at the end of the project?

On the University of Glasgow's secure networked J-Drive.

What formats will the data be archived in?

- Undergraduate longitudinal clinical performance assessment data – Microsoft Excel spreadsheets (.xls) and Stata statistical software (.dta)
- Postgraduate longitudinal clinical performance assessment data - Microsoft Excel spreadsheets (.xls) and Stata statistical software (.dta)
- Undergraduate professional examination results - Microsoft Excel spreadsheets (.xls) and Stata statistical software (.dta)
- Focus group session recordings (*MP4*)
- Transcripts of focus group sessions (*Unconfirmed at present – likely to be Microsoft Word (.doc) or .txt)*

  UPDATE (JUNE 2020)
- Audio and video recordings of focus groups (MP4)

  UPDATE (NOVEMBER 2020)
- Transcripts of focus group sessions (.pdf)

7. Data sharing

Is any of the data suitable for sharing?

No. Ethical approval and a data sharing agreement between NES and the University of Glasgow were granted on the basis that strict data usage rules were placed. These rules did not permit the sharing of data used for this project.

How will the data be shared?
Not applicable.
Who should be able to access and use the shared data?
Not applicable.

8. Implementation
Who is responsible for implementing this plan?
Jamie Dickie (the researcher)
How will this plan be kept up-to-date?
The data management plan will be reviewed by the researcher alongside their project supervisory team. It is unlikely that any changes will be made to the data management plan as ethical approval and a data sharing agreement were granted on the assurance that the data were managed and stored using the processes described in the answers to the above questions. However, if any alterations were to be proposed, the data controllers (University of Glasgow and NES) and the MVLS Ethics Committee are to be informed and consulted. UPDATE (JUNE 2020): This data management plan has been updated to provide details relating to methodological changes to the study which may need to be implemented as contingencies to the COVID-19 outbreak. The potential need to conduct focus groups online has required amendments to the original ethical approval document to be submitted to the MVLS ethics committee. Further updates to this document will be required once details surrounding online video conferencing data security have been received from the University of Glasgow data protection services. UPDATE (NOVEMBER 2020): This data management plan has been updated to provide details relating to methodological changes to the study which were required as contingencies to the COVID-19 outbreak.
What actions are necessary to implement this plan?
Ethical approval has already been granted by the University of Glasgow MVLS Ethics Committee based on the data management steps described above. The data controllers have been contacted via email and made aware that transfer of assessment data sets to the third-party analysis must be via secure nhs.net email addresses. IT support at the University of Glasgow have also previously been contacted to provide storage space on the University of Glasgow's secure networked J-Drive that can only be accessed by the researcher and one of the supervisory team. The researcher will need to obtain a suitably encrypted audio recording device ahead of conducting the focus groups. UPDATE (JUNE 2020): Amendments to the project's original ethical approval application were submitted to the MVLS ethics committee in April 2020 to notify them of a potential change to methodological approach and the additional ethical issues that have been taken into consideration. The MVLS committee has responded and requested further assurances on the security of online video conferencing

systems. They have also requested that a privacy notice for participants and a DPIA are completed. The PhD researcher is currently working to fulfil these requests. UPDATE (NOVEMBER 2020): Amendments to the project's ethical approval application were submitted to the MVLS ethics committee in November 2020 to notify them of the changes to methodological approach and the additional ethical issues that were taken into consideration. The MVLS committee responded and confirmed that ethical approval was granted.
What training or further information are needed to implement this plan?
The researcher has previously contacted the Data Protection team for advice on GDPR. They have subsequently completed a Legitimate Interests Assessment to ensure that the project is compliant with GDPR. This Legitimate Interests Assessment was submitted as part of the project's ethical approval application.

Appendix 8 – Additional undergradute examination analysis data

Cohort 1: Summary statistics for all BDS examination results with available numeric (percentile) scores. MCQ = Multiple-choice question. MSA = Multiple-short answer. OSCE = Objective structured clinical examination. SD = Standard deviation.

BDS Year	Examination	Number of student participants (n)	Mean (%)	SD (%)	Min (%)	Q1 (%)	Median (%)	Q3 (%)	Max (%)
1	Written (2 x MCQ)	90	65.93	9.13	30.32	61.76	66.65	71.55	87.5
	OSCE	89	67.63	7.49	50.60	63.98	66.90	72.38	84.64
	Aggregated	90	66.79	7.61	40.46	62.55	66.75	71.86	86.07
2	MCQ	88	72.48	6.69	55.00	69.00	73.00	76.00	90.00
	MSA	88	69.2	8.48	33.33	64.72	70.00	74.44	88.89
	OSCE	88	82.19	5.69	67.90	78.04	82.44	86.04	99.25
	Aggregated	88	74.62	5.81	52.51	71.15	75.06	78.40	87.32
3	Anatomy	87	73.05	13.11	45.63	63.13	72.50	83.75	97.50
	MCQ	87	73.41	6.64	58.00	70.00	73.00	77.00	91.00
	MSA	87	67.65	7.73	44.50	62.50	68.50	73.00	87.00
	OSCE	87	76.18	6.88	54.00	73.00	76.67	80.33	92.67
	Aggregated	87	72.57	7.00	58.79	66.74	72.42	77.61	88.69
4	MSA (final)	84	75.28	6.21	60.88	72.19	75.63	79.25	89.38
5	OSCE (final)	82	71.54	4.61	56.37	69.38	71.94	73.97	81.43

Cohort 2: Summary statistics for all BDS examination results with available numeric (percentile) scores. MCQ = Multiple-choice Question. MSA = Multiple-short answer. OSCE = Objective structured clinical examination. SD = Standard deviation.

BDS Year	Examination	Number of student participants (n)	Mean (%)	SD (%)	Min (%)	Q1 (%)	Median (%)	Q3 (%)	Max (%)
1	Written (2 x MCQ)	92	65.78	9.74	43.22	60.17	65.84	71.13	89.86
	OSCE	92	76.51	6.41	54.17	72.36	77.36	80.94	92.92
	Aggregated	92	71.14	7.59	48.77	67.19	71.25	75.91	89.09
2	MCQ	92	71.93	7.89	44.00	68.50	72.00	77.00	88.00
	MSA	92	76.24	8.26	54.44	70.83	76.94	82.50	92.22
	OSCE	92	80.39	5.50	67.39	76.97	80.65	84.79	90.19
	Aggregated	92	76.19	6.10	60.60	73.03	76.13	80.09	87.75
3	Anatomy	92	77.62	15.07	23.13	70.63	80.63	88.13	100.00
	MCQ	92	77.47	9.07	50.00	72.00	80.00	84.00	92.00
	MSA	92	70.43	9.46	45.25	64.75	71.00	76.38	89.00
	OSCE	92	74.88	6.79	46.30	70.93	75.37	79.07	88.15
	Aggregated	92	75.10	8.75	44.67	70.53	75.82	81.71	91.67
4	MSA (final)	91	71.94	5.93	53.63	68.38	72.13	75.50	84.63
5	OSCE (final)	93	71.39	4.94	57.62	68.26	71.48	74.98	81.61

Cohort 3: Summary statistics for all BDS examination results with available numeric (percentile) scores. MCQ = Multiple-choice question. MSA = Multiple-short answer. OSCE = Objective structured clinical examination. SD = Standard deviation.

BDS Year	Examination	Number of student participants (n)	Mean (%)	SD (%)	Min (%)	Q1 (%)	Median (%)	Q3 (%)	Max (%)
1	Written (2 x MCQ)	73	73.87	8.30	53.67	67.33	73.00	79.83	91.17
	OSCE	73	77.99	6.74	56.67	73.33	79.38	82.71	90.83
	Aggregated	73	75.93	7.09	59.42	70.33	76.54	81.29	91.00
2	MCQ	72	72.94	8.08	52.00	68.00	72.00	90.00	91.00
	MSA	72	70.37	7.16	52.78	65.28	70.83	74.72	86.11
	OSCE	72	80.55	5.50	67.26	77.02	80.91	85.06	91.17
	Aggregated	72	74.62	5.53	59.94	71.17	74.33	78.72	87.19
3	Anatomy	72	78.29	12.39	28.75	72.19	80.63	87.19	96.25
	MCQ	72	84.39	5.20	65.00	82.00	85.00	88.00	92.00
	MSA	72	76.58	6.58	56.00	73.75	77.13	80.75	91.75
	OSCE	72	72.13	7.18	48.13	68.92	72.81	76.77	85.02
	Aggregated	72	77.85	6.56	51.93	74.68	78.80	81.83	88.30
4	MSA (final)	72	79.84	6.29	61.12	75.94	80.25	84.75	92.25
5	OSCE (final)	69	74.94	4.03	63.10	72.39	74.90	77.63	82.08

Cohort 1: Summary statistics for thirds of aggregated early examination (BDS1/2/3) performance and fifths of final examination (BDS4/5) performance. Most frequent grade emboldened.

		n	Mean	SD	Min	Median	Max	Grades
BDS1	***Overall***	90	66.79	7.61	40.46	66.75	86.07	
	T1	30	58.78	4.96	40.46	60.14	63.7	
	T2	30	66.77	1.61	63.92	66.75	69.83	
	T3	30	74.82	4.19	70.10	73.49	86.07	
BDS2	***Overall***	88	74.62	5.81	52.51	75.06	87.32	
	T1	30	68.63	3.93	52.51	70.05	72.01	
	T2	29	74.71	1.65	72.02	75.13	77.50	
	T3	29	80.73	2.98	77.89	79.31	87.32	
BDS3	***Overall***	87	72.57	7.00	58.79	72.42	88.69	
	T1	30	65.10	3.16	58.79	65.80	69.03	
	T2	28	72.46	2.17	69.30	72.51	76.23	
	T3	29	80.41	3.49	76.27	78.78	88.69	
BDS4	***Overall***	84	75.28	6.21	60.88	75.63	89.38	
	Q1	17	65.69	2.40	60.88	66.13	68.50	3x C3, **6x D1**, 4x D2, 2x D3, 2x E1
	Q2	17	73.01	1.56	69.25	73.63	74.88	5x B3, **9x C1**, 2x C2, 1x C3
	Q3	17	75.90	0.64	75.00	75.63	77.00	8x B2, **9x B3**
	Q4	18	78.63	0.88	77.13	78.38	79.88	**14x B1**, 4x B2
	Q5	15	83.96	2.13	81.25	83.63	89.28	**15x A5**
BDS5	***Overall***	82	71.54	4.61	56.37	71.94	81.43	
	Q1	17	65.01	2.84	56.37	65.61	68.20	2x C, **14x D**, 1x F
	Q2	16	69.52	0.57	68.27	69.62	70.29	**16x C**
	Q3	17	71.95	0.66	70.33	71.94	72.78	8x B, **9x C**
	Q4	16	73.68	0.68	72.83	73.44	74.89	**16x B**
	Q5	16	77.95	1.86	75.41	77.47	81.43	**14x A**, 2x B

T1 = lowest 33% T3 = top 33% Q1 = lowest 20% Q5 = top 20%
SD = Standard deviation

Cohort 2: Summary statistics for thirds of aggregated early examination (BDS1/2/3) performance and fifths of final examination (BDS4/5) performance. Most frequent grade emboldened.

		n	Mean	SD	Min	Median	Max	Grades
BDS1	***Overall***	92	71.14	7.59	48.77	71.25	89.09	
	T1	31	63.13	4.99	48.77	64.71	68.35	
	T2	31	71.29	1.87	68.37	71.33	74.57	
	T3	30	79.28	3.79	74.61	78.76	89.09	
BDS2	***Overall***	92	76.19	6.10	60.60	76.13	87.75	
	T1	31	69.57	3.93	60.60	70.22	74.26	
	T2	31	76.43	1.42	74.34	76.35	79.20	
	T3	30	82.77	2.68	79.21	82.62	87.75	
BDS3	***Overall***	92	75.10	8.75	44.67	75.82	91.67	
	T1	31	65.41	6.37	44.67	67.91	72.17	
	T2	31	76.14	2.02	72.28	75.88	79.97	
	T3	30	84.03	3.13	80.06	83.09	91.67	
BDS4	***Overall***	91	71.94	5.93	53.63	72.13	84.63	
	Q1	19	63.88	3.78	53.63	64.75	67.50	1x E3, 1x E2, 2x E1, 1x D3, **6x D2**, 4x D1, 4x C3
	Q2	20	69.50	0.81	68.38	69.38	70.63	7x C3, **13x C2**
	Q3	18	72.21	0.73	70.75	72.38	73.13	1x C2, **17x C1**
	Q4	16	74.77	1.18	73.25	74.69	77.00	**10x B3**, 6x B2
	Q5	18	80.38	2.39	77.13	79.88	84.63	1x B2 6xB1, **5x A5**, 2x A4, 4x A3
BDS5	***Overall***	93	71.39	4.94	57.62	71.48	81.61	
	Q1	19	64.51	2.61	57.62	65.31	67.42	1x E, **12x D**, 6x C
	Q2	19	68.76	0.74	67.45	68.93	69.63	**19x C**
	Q3	18	71.51	0.96	70.13	71.48	73.10	7x C, **11x B**
	Q4	19	74.32	0.74	73.19	74.48	75.33	**19x B**
	Q5	18	78.19	1.83	75.43	77.97	81.61	3x B, **15x A**

T1 = lowest 33% T3 = top 33% Q1 = lowest 20% Q5 = top 20%
SD = Standard deviation

Cohort 3: Summary statistics for thirds of overall early examination (BDS1/2/3) performance and fifths of final examination (BDS4/5) performance. Most frequent grade emboldened.

		n	Mean	SD	Min	Median	Max	Grades
BDS1	***Overall***	73	75.93	7.09	59.42	76.54	91.00	
	T1	25	67.92	3.42	59.42	68.65	72.23	
	T2	24	76.53	2.07	73.48	76.57	79.96	
	T3	24	83.67	2.87	79.98	83.34	91.00	
BDS2	***Overall***	72	74.62	5.53	59.94	74.33	87.19	
	T1	24	68.79	3.11	59.94	69.45	72.06	
	T2	24	74.33	1.68	72.11	74.33	76.95	
	T3	24	80.74	2.68	77.04	80.21	87.19	
BDS3	***Overall***	72	79.84	6.29	61.12	80.25	92.25	
	T1	24	70.94	6.23	51.92	72.49	76.64	
	T2	24	78.83	1.14	76.85	78.80	81.06	
	T3	24	83.77	2.32	81.34	83.10	88.30	
BDS4	***Overall***	72	79.84	6.29	61.12	80.25	92.25	
	Q1	15	70.67	3.87	61.13	71.63	75.00	1x E, 1x D, 11x C, 2x B
	Q2	14	76.91	1.02	75.13	77.00	78.13	14x B
	Q3	16	80.57	1.29	78.50	80.44	83.00	3x B, 13x A
	Q4	13	84.32	0.61	83.38	84.50	85.13	13x A
	Q5	14	87.60	2.23	85.25	87.13	92.25	14x A
BDS5	***Overall***	69	74.94	4.03	63.10	74.90	82.08	
	Q1	14	69.27	2.21	63.10	70.01	71.18	1x D, 9x C, 4x B
	Q2	14	72.82	0.64	71.62	72.95	73.82	14x B
	Q3	14	75.08	0.88	73.91	75.00	76.26	14x B
	Q4	14	77.38	0.66	76.35	77.42	78.55	4x B, 10x A
	Q5	13	80.57	0.89	79.14	80.58	82.08	13x A

T1 = lowest 33% T3 = top 33% Q1 = lowest 20% Q5 = top 20%
SD = Standard deviation

Cohort 1: Cross tabulations between thirds of mean aggregated performance in early examinations (BDS1/2/3) and fifths of performance in each of the final examinations (BDS4/5).

		BDS4							BDS5						
		Q1 % (n)	***Q2*** % (n)	***Q3*** % (n)	***Q4*** % (n)	***Q5*** % (n)	p-value (Fisher's exact test)	C-statistic	***Q1*** % (n)	***Q2*** % (n)	***Q3*** % (n)	***Q4*** % (n)	***Q5*** % (n)	p-value (Fisher's exact test)	C-statistic
BDS1	***T1***	32.00 (8)	28.00 (7)	24.00 (6)	8.00 (2)	8.00 (2)	<0.01	0.77	24.00 (6)	28.00 (7)	16.00 (4)	20.00 (5)	12.00 (3)	0.37	0.65
	T2	20.00 (6)	20.00 (6)	23.33 (7)	33.33 (10)	3.33 (1)			27.59 (8)	17.24 (5)	24.14 (7)	17.24 (5)	13.79 (4)		
	T3	7.14 (2)	14.29 (4)	14.29 (4)	21.43 (6)	42.86 (12)			7.41 (2)	14.81 (4)	22.22 (6)	22.22 (6)	33.33 (9)		
BDS2	***T1***	44.44 (12)	22.22 (6)	14.81 (4)	14.81 (4)	3.70 (1)	<0.01	0.80	40.74 (11)	25.93 (7)	14.81 (4)	11.11 (3)	7.41 (2)	<0.01	0.78
	T2	17.86 (5)	21.43 (6)	28.57 (8)	25.00 (7)	7.14 (2)			22.22 (6)	7.41 (2)	29.63 (8)	29.63 (8)	11.11 (3)		
	T3	0.00 (0)	17.24% (5)	17.24 (5)	24.14 (7)	41.38 (12)			0.00 (0)	25.00 (7)	17.86 (5)	17.86 (5)	39.29 (11)		
BDS3	***T1***	51.72 (15)	24.14 (7)	17.24 (5)	3.45 (1)	3.45 (1)	<0.01	0.82	46.43 (13)	17.86 (5)	21.43 (6)	10.71 (3)	3.57 (1)	<0.01	0.81
	T2	7.41 (2)	22.22 (6)	29.63 (8)	37.04 (10)	3.70 (1)			14.81 (4)	33.33 (9)	25.93 (7)	18.52 (5)	7.41 (2)		
	T3	0.00 (0)	14.29 (4)	14.29 (4)	25.00 (7)	46.43 (13)			0.00 (0)	7.41 (2)	14.81 (4)	29.63 (8)	48.15 (13)		

T1 = lowest 33% T3 = top 33% Q1 = lowest 20% Q5 = top 20%

Cohort 2: Cross tabulations between thirds of mean aggregated performance in each of the early examinations (BDS1/2/3) and fifths of performance in each of the final examinations (BDS4/5).

		BDS4							BDS5						
		Q1 % (n)	***Q2*** % (n)	***Q3*** % (n)	***Q4*** % (n)	***Q5*** % (n)	p-value (Fisher's exact test)	C-statistic	***Q1*** % (n)	***Q2*** % (n)	***Q3*** % (n)	***Q4*** % (n)	***Q5*** % (n)	p-value (Fisher's exact test)	C-statistic
BDS1	***T1***	32.14 (9)	39.29 (11)	14.29 (4)	14.29 (4)	0.00 (0)	<0.01	0.70	28.57 (8)	25.00 (7)	21.43 (6)	21.43 (6)	3.57 (1)	0.11	0.75
	T2	23.33 (7)	13.33 (4)	26.67 (8)	23.33 (7)	13.33 (4)			10.34 (3)	24.14 (7)	24.14 (7)	24.14 (7)	17.24 (5)		
	T3	10.00 (3)	10.00 (3)	20.00 (6)	13.33 (4)	46.67 (14)			16.67 (5)	13.33 (4)	16.67 (5)	16.67 (5)	36.67 (11)		
BDS2	***T1***	51.85 (14)	33.33 (9)	7.41 (2)	7.41 (2)	0.00 (0)	<0.01	0.74	30.77 (8)	15.38 (4)	30.77 (8)	19.23 (5)	3.85 (1)	<0.01	0.82
	T2	9.68 (3)	25.81 (8)	25.81 (8)	29.03 (9)	9.68 (3)			12.90 (4)	38.71 (12)	16.13 (5)	25.81 (8)	6.45 (2)		
	T3	6.67 (2)	3.33 (1)	26.67 (8)	13.33 (4)	50.00 (15)			13.33 (4)	6.67 (2)	16.67 (5)	16.67 (5)	46.67 (14)		
BDS3	***T1***	46.43 (13)	39.29 (11)	7.14 (2)	7.14 (2)	0.00 (0)	<0.01	0.82	29.63 (8)	29.63 (8)	29.63 (8)	11.11 (3)	0.00 (0)	<0.01	0.69
	T2	12.90 (4)	25.81 (8)	38.71 (12)	19.35 (6)	3.23 (1)			19.35 (6)	25.81 (8)	16.13 (5)	25.81 (8)	12.90 (4)		
	T3	6.67 (2)	0.00 (0)	13.33 (4)	23.33 (7)	56.67 (17)			6.67 (2)	10.00 (3)	16.67 (5)	23.33 (7)	43.33 (13)		

T1 = lowest 33% T3 = top 33% Q1 = lowest 20% Q5 = top 20%

Cohort 3: Cross tabulations between thirds of mean aggregated performance in each of the early examinations (BDS1/2/3) and fifths of performance in each of the final examinations (BDS4/5).

		BDS4							BDS5						
		Q1 % (n)	***Q2*** % (n)	***Q3*** % (n)	***Q4*** % (n)	***Q5*** % (n)	p-value (Fisher's exact test)	C-statistic	***Q1*** % (n)	***Q2*** % (n)	***Q3*** % (n)	***Q4*** % (n)	***Q5*** % (n)	p-value (Fisher's exact test)	C-statistic
BDS1	***T1***	44.00 (11)	20.00 (5)	8.00 (2)	12.00 (3)	16.00 (4)	0.03	0.52	34.78 (8)	8.70 (2)	21.74 (5)	17.39 (4)	17.39 (4)	0.16	0.54
	T2	8.33 (2)	25.00 (6)	25.00 (6)	16.67 (4)	25.00 (6)			17.39 (4)	26.09 (6)	17.39 (4)	21.74 (5)	17.39 (4)		
	T3	9.09 (2)	9.09 (2)	36.36 (8)	27.27 (6)	18.18 (4)			0.00 (0)	28.57 (6)	23.81 (5)	23.81 (5)	23.81 (5)		
BDS2	***T1***	41.67 (10)	25.00 (6)	4.17 (1)	12.50 (3)	16.67 (4)	<0.01	0.57	40.91 (9)	9.09 (2)	13.64 (3)	18.18 (4)	18.18 (4)	0.03	0.63
	T2	16.67 (4)	25.00 (6)	16.67 (4)	25.00 (6)	16.67 (4)			17.39 (4)	26.09 (6)	26.09 (6)	21.74 (5)	8.70 (2)		
	T3	4.17 (1)	8.33 (2)	45.83 (11)	16.67 (4)	25.00 (6)			0.00 (0)	26.09 (6)	21.74 (5)	21.74 (5)	30.43 (7)		
BDS3	***T1***	50.00 (12)	33.33 (8)	8.33 (2)	8.33 (2)	0.00 (0)	<0.01	0.61	45.45 (10)	18.18 (4)	18.18 (4)	18.18 (4)	0.00 (0)	<0.01	0.73
	T2	8.33 (2)	16.67 (4)	25.00 (6)	20.83 (5)	29.17 (7)			13.04 (3)	26.09 (6)	17.39 (4)	26.09 (6)	17.39 (4)		
	T3	4.17 (1)	8.33 (2)	33.33 (8)	25.00 (6)	29.17 (7)			0.00 (0)	17.39 (4)	26.09 (6)	17.39 (4)	39.13 (9)		

T1 = lowest 33% T3 = top 33% Q1 = lowest 20% Q5 = top 20%

Appendix 9 – Additional LIFTUPP© analysis data

i. Descriptive statistics

Cohort 1: Summary statistics for the number of clinical assessments completed per student within and across Bachelor of Dental Surgery (BDS) academic years.

BDS year	*Total LIFTUPP© assessments completed (n)*	*Mean*	*Standard deviation*	*Minimum*	*Q1*	*Median*	*Q3*	*Maximum*
3	2,756	34.45	11.14	11	27	33	42.5	62
4	4,044	50.55	12.85	26	39.5	50	60	79
5	12,399	154.99	30.37	89	138	152.5	174	237
All	19,199	239.99	41.4	151	212.5	240.5	266.5	349

Cohort 2: Summary statistics for the number of clinical assessments completed per student within and across Bachelor of Dental Surgery (BDS) academic years.

BDS year	*Total LIFTUPP© assessments completed (n)*	*Mean*	*Standard deviation*	*Minimum*	*Q1*	*Median*	*Q3*	*Maximum*
3	2,746	31.93	11.16	13	24	29	38	71
4	5,225	60.76	14.60	32	50	59	69	111
5	12,341	143.50	38.83	75	113	136	174	249
All	20,312	236.19	45.82	157	197	232	269	346

Cohort 3: Summary statistics for the number of clinical assessments completed per student within and across Bachelor of Dental Surgery (BDS) academic years.

BDS year	*Total LIFTUPP© assessments completed (n)*	*Mean*	*Standard deviation*	*Minimum*	*Q1*	*Median*	*Q3*	*Maximum*
3	2,938	43.21	12.73	25	32	41	50.5	81
4	4,806	70.68	15.72	26	62.5	70	77	115
5	13,073	192.25	41.50	118	157.5	189.50	224.50	289
All	20,817	306.13	47.85	204	270.5	305.5	341	430

ii. *Group-based trajectory modelling - Models returned without errors*

BIC^2 = Bayesian information criterion for the total number of participants.

BIC^3 = Bayesian information criterion for the total number of observations.

Data models listed from least to most negative BIC^2.

Censored normal data distribution

Cohort 1

Number of groups	Model	BIC^2 (n = 80)	BIC^3 (n = 19,199)	Contains at least X students per group, where X =			
				5	*10*	*15*	*20*
4	1 3 2 3	-22537.16	-22583.74	✓	✓	✗	✗
4	3 3 3 1	-22540.91	-22590.23	✗	✗	✗	✗
4	2 3 1 2	-22570.69	-22614.54	✗	✗	✗	✗
4	1 3 1 2	-22571.67	-22612.77	✓	✗	✗	✗
4	2 3 1 0	-22597.27	-22635.63	✗	✗	✗	✗
4	1 1 3 0	-22598.49	-22634.11	✗	✗	✗	✗
4	2 3 1 1	-22599.46	-22640.56	✗	✗	✗	✗
4	1 3 2 1	-22600.34	-22641.45	✓	✗	✗	✗
4	1 1 3 1	-22600.68	-22639.05	✗	✗	✗	✗
4	1 3 1 1	-22600.68	-22639.05	✗	✗	✗	✗
4	2 1 3 0	-22601.32	-22639.68	✗	✗	✗	✗
4	2 2 1 3	-22602.14	-22645.99	✓	✓	✓	✗
4	2 0 2 3	-22615.96	-22657.07	✓	✗	✗	✗
4	2 2 0 3	-22615.96	-22657.07	✓	✗	✗	✗
4	1 2 0 3	-22621.16	-22659.52	✓	✓	✗	✗
4	1 2 1 3	-22624.06	-22665.17	✓	✓	✗	✗
4	3 1 1 2	-22629.22	-22670.33	✗	✗	✗	✗
4	3 2 2 1	-22631.97	-22675.81	✗	✗	✗	✗

3	1 1 3	-22639.40	-22669.54	✓	✓	✗	✗
3	1 3 2	-22643.53	-22676.41	✓	✗	✗	✗
3	3 1 2	-22643.53	-22676.41	✓	✓	✗	✗
3	3 2 2	-22651.08	-22686.70	✓	✗	✗	✗
4	3 2 0 0	-22654.56	-22690.19	✗	✗	✗	✗
3	2 3 2	-22655.17	-22690.80	✗	✗	✗	✗
4	3 2 0 1	-22656.75	-22695.12	✗	✗	✗	✗
4	1 3 0 1	-22660.36	-22695.99	✗	✗	✗	✗
4	3 1 0 0	-22662.18	-22695.07	✗	✗	✗	✗
4	1 3 1 0	-22662.25	-22697.87	✗	✗	✗	✗
4	2 2 2 2	-22666.32	-22710.17	✓	✗	✗	✗
3	3 0 3	-22671.26	-22704.14	✓	✓	✗	✗
4	2 2 1 2	-22672.90	-22714.00	✓	✗	✗	✗
3	0 3 2	-22674.41	-22704.55	✓	✗	✗	✗
3	1 3 0	-22681.57	-22708.97	✗	✗	✗	✗
3	2 3 0	-22681.70	-22711.85	✗	✗	✗	✗
3	1 3 1	-22683.76	-22713.90	✗	✗	✗	✗
4	2 2 0 2	-22684.95	-22723.31	✓	✗	✗	✗
4	1 2 2 2	-22687.39	-22728.49	✗	✗	✗	✗
3	3 0 2	-22697.69	-22727.83	✓	✗	✗	✗
3	3 1 0	-22701.38	-22728.78	✓	✓	✓	✗
4	2 2 2 0	-22709.28	-22747.64	✗	✗	✗	✗
3	3 1 1	-22711.30	-22741.44	✗	✗	✗	✗
4	2 2 2 1	-22711.47	-22752.57	✗	✗	✗	✗
4	1 2 1 2	-22716.25	-22754.62	✗	✗	✗	✗
4	1 2 2 1	-22719.72	-22758.08	✗	✗	✗	✗
3	0 3 0	-22721.17	-22745.83	✗	✗	✗	✗
4	1 2 0 2	-22721.67	-22757.29	✗	✗	✗	✗
2	1 3	-22724.34	-22746.26	✓	✓	✓	✗
2	3 1	-22724.34	-22746.26	✓	✓	✓	✗

4	1 1 1 2	-22733.65	-22769.28	✓	✓	✗	✗
4	2 2 1 0	-22737.05	-22772.67	✗	✗	✗	✗
4	2 2 1 1	-22739.24	-22777.60	✗	✗	✗	✗
3	2 2 2	-22740.19	-22773.07	✓	✓	✓	✗
4	2 2 0 0	-22745.35	-22778.23	✗	✗	✗	✗
4	1 2 1 0	-22745.49	-22778.37	✗	✗	✗	✗
3	1 2 2	-22745.87	-22776.02	✓	✓	✓	✗
3	2 1 2	-22745.87	-22776.02	✓	✓	✓	✗
4	2 2 0 1	-22747.54	-22783.16	✗	✗	✗	✗
4	1 2 1 1	-22747.68	-22783.30	✗	✗	✗	✗
4	1 0 2 0	-22752.73	-22782.88	✗	✗	✗	✗
2	3 2	-22753.81	-22778.47	✓	✗	✗	✗
4	1 2 0 0	-22754.40	-22784.55	✗	✗	✗	✗
4	1 2 0 1	-22754.92	-22787.81	✗	✗	✗	✗
3	3 0 0	-22755.74	-22780.40	✗	✗	✗	✗
3	3 0 1	-22757.93	-22785.33	✗	✗	✗	✗
4	2 0 2 1	-22768.73	-22804.36	✓	✗	✗	✗
3	0 2 2	-22773.50	-22800.90	✓	✓	✗	✗
3	2 0 2	-22773.50	-22800.90	✓	✓	✗	✗
3	1 2 0	-22792.82	-22817.48	✓	✓	✓	✗
4	1 1 1 1	-22794.50	-22827.38	✗	✗	✗	✗
3	2 2 0	-22801.78	-22829.19	✗	✗	✗	✗
3	2 2 1	-22803.97	-22834.12	✗	✗	✗	✗
3	1 2 1	-22811.29	-22838.69	✗	✗	✗	✗
2	3 0	-22811.54	-22830.72	✗	✗	✗	✗
3	0 2 1	-22812.34	-22837.00	✓	✓	✗	✗
3	0 0 2	-22819.06	-22840.98	✓	✓	✗	✗
3	0 2 0	-22819.06	-22840.98	✓	✓	✗	✗
3	2 1 0	-22822.41	-22847.07	✓	✓	✓	✓
3	2 1 1	-22862.98	-22890.38	✗	✗	✗	✗

2	2 2	-22865.00	-22886.92	✓	✓	✓	✓
1	3	-22876.40	-22890.10	✓	✓	✓	✓
3	2 0 0	-22877.07	-22899.00	✗	✗	✗	✗
3	0 1 1	-22877.15	-22899.07	✓	✓	✗	✗
3	2 0 1	-22879.26	-22903.93	✗	✗	✗	✗
3	0 0 1	-22881.98	-22901.17	✓	✓	✗	✗
3	0 1 0	-22881.98	-22901.17	✓	✓	✗	✗
3	1 0 0	-22881.98	-22901.17	✓	✓	✗	✗
2	2 1	-22903.10	-22922.29	✓	✓	✓	✓
2	0 2	-22912.39	-22928.84	✓	✓	✗	✗
2	2 0	-22914.01	-22930.45	✓	✓	✓	✓
3	1 1 1	-22928.65	-22953.32	✗	✗	✗	✗
3	1 0 1	-22942.24	-22964.17	✗	✗	✗	✗
2	1 1	-22955.87	-22972.31	✓	✓	✓	✓
2	1 0	-22974.58	-22988.28	✓	✓	✓	✗
2	0 1	-23015.29	-23028.99	✓	✓	✗	✗
1	2	-23041.21	-23052.18	✓	✓	✓	✓
1	1	-23146.74	-23154.96	✓	✓	✓	✓
3	0 0 0	-23283.82	-23300.27	✗	✗	✗	✗
2	0 0	-23337.82	-23348.78	✓	✓	✓	✓
1	0	-23525.33	-23530.81	✓	✓	✓	✓

Cohort 2

Number of groups	Model	BIC[2] (n = 86)	BIC[3] (n = 20,312)	Contains at least X students per group, where X =			
				5	*10*	*15*	*20*
2	3 3	-22440.56	-22467.88	✓	✓	✓	✓
2	1 3	-22449.02	-22470.88	✓	✓	✓	✗
2	3 2	-22452.12	-22476.71	✓	✓	✓	✗
2	3 1	-22453.90	-22475.76	✓	✓	✓	✗
2	0 3	-22473.22	-22492.35	✓	✓	✗	✗
2	3 0	-22526.04	-22545.17	✓	✓	✗	✗
2	1 1	-22634.80	-22651.19	✓	✓	✓	✓
2	2 1	-22635.68	-22654.80	✓	✓	✓	✓
2	2 2	-22637.89	-22659.75	✓	✓	✓	✓
2	0 1	-22675.29	-22688.96	✓	✓	✓	✓
2	0 2	-22677.12	-22693.52	✓	✓	✓	✓
2	1 0	-22701.56	-22715.22	✓	✓	✓	✗
2	2 0	-22702.88	-22719.28	✓	✓	✓	✗
1	1	-22792.06	-22800.25	✓	✓	✓	✓
1	2	-22793.57	-22804.50	✓	✓	✓	✓
1	3	-22904.83	-22918.49	✓	✓	✓	✓
2	0 0	-22915.72	-22926.65	✓	✓	✓	✓
1	0	-23075.92	-23081.39	✓	✓	✓	✓

Cohort 3

Number of groups	Model	BIC[2] (n = 68)	BIC[3] (n = 20,817)	Contains at least X students per group, where X =			
				5	*10*	*15*	*20*
2	3 2	-22062.14	-22087.90	✓	✓	✓	✓
2	1 3	-22071.31	-22094.21	✓	✓	✓	✓
2	3 1	-22071.31	-22094.21	✓	✓	✓	✓
2	3 0	-22149.76	-22169.79	✓	✗	✗	✗
3	2 2 1	-22176.73	-22208.21	✓	✓	✓	✓
2	2 2	-22198.63	-22221.52	✓	✓	✓	✓
2	1 1	-22206.97	-22224.14	✓	✓	✓	✓
2	2 1	-22209.06	-22229.09	✓	✓	✓	✓
2	0 2	-22264.45	-22281.62	✓	✓	✓	✓
2	2 0	-22264.45	-22281.62	✓	✓	✓	✓
2	0 1	-22276.14	-22290.45	✓	✓	✓	✓
2	1 0	-22276.14	-22290.45	✓	✓	✓	✓
1	2	-22412.64	-22424.09	✓	✓	✓	✓
1	1	-22415.24	-22423.83	✓	✓	✓	✓
1	3	-22470.02	-22484.33	✓	✓	✓	✓
2	0 0	-22516.52	-22527.97	✓	✓	✓	✓
1	0	-22708.99	-22714.71	✓	✓	✓	✓
2	0 0	-22516.52	22527.97	✓	✓	✓	✓

Bernoulli data distribution

Threshold LIFTUPP© development indicator = 4

Cohort 1

Number of groups	Model	BIC^2 (n = 80)	BIC^3 (n = 19,199)	Contains at least X students per group, where X =			
				5	*10*	*15*	*20*
1	0	-5686.26	-5688.98	✓	✓	✓	✓
1	1	-5501.68	-5507.13	✓	✓	✓	✓
1	2	-5492.66	-5500.83	✓	✓	✓	✓
1	3	-5417.16	-5428.06	✓	✓	✓	✓

Cohort 2

Number of groups	Model	BIC^2 (n = 86)	BIC^3 (n = 20,318)	Contains at least X students per group, where X =			
				5	*10*	*15*	*20*
1	0	-6180.09	-6182.83	✓	✓	✓	✓
1	1	-5931.22	-5936.68	✓	✓	✓	✓
1	2	-5933.44	-5941.64	✓	✓	✓	✓
1	3	-5905.25	-5916.18	✓	✓	✓	✓

Cohort 3

Number of groups	Model	BIC^2 (n = 80)	BIC^3 (n = 19,199)	Contains at least X students per group, where X =			
				5	*10*	*15*	*20*
1	0	-5549.26	-5552.12	✓	✓	✓	✓
1	1	-5339.07	-5344.79	✓	✓	✓	✓
1	2	-5325.21	-5333.79	✓	✓	✓	✓
1	3	-5299.08	-5310.52	✓	✓	✓	✓

Threshold LIFTUPP© development indicator = 5

Cohort 1

Number of groups	Model	BIC^2 (n = 80)	BIC^3 (n = 19,199)	Contains at least X students per group, where X =			
				5	*10*	*15*	*20*
2	3 2	-12432.70	-12454.62	✓	✓	✓	✗
2	1 3	-12443.08	-12462.26	✓	✓	✓	✗
2	0 3	-12479.40	-12495.85	✓	✓	✗	✗
2	3 0	-12479.40	-12495.85	✓	✓	✗	✗
2	3 1	-12505.11	-12524.29	✓	✓	✓	✗
2	2 2	-12511.56	-12530.74	✓	✓	✓	✗
2	0 2	-12557.52	-12571.22	✓	✓	✗	✗
2	2 0	-12557.52	-12571.22	✓	✓	✗	✗
1	3	-12587.31	-12598.27	✓	✓	✓	✓
2	2 1	-12615.01	-12631.45	✓	✓	✓	✗
2	1 1	-12653.09	-12666.79	✓	✓	✓	✗
1	2	-12679.02	-12687.24	✓	✓	✓	✓
2	0 1	-12684.83	-12695.79	✓	✓	✓	✓
2	1 0	-12779.30	-12790.26	✗	✗	✗	✗
1	1	-12816.48	-12821.96	✓	✓	✓	✓
2	0 0	-13092.22	-13100.44	✓	✓	✓	✗
1	0	-13248.97	-13251.72	✓	✓	✓	✓

Cohort 2

Number of groups	Model	BIC[2] (n = 86)	BIC[3] (n = 20,312)	Contains at least X students per group, where X =			
				5	10	15	20
2	1 3	-13441.60	-13460.72	✓	✓	✓	✓
2	3 1	-13441.60	-13460.72	✓	✓	✓	✓
2	3 2	-13443.36	-13465.21	✓	✓	✓	✓
2	3 3	-13444.46	-13469.05	✓	✓	✓	✓
2	0 3	-13495.47	-13511.86	✓	✓	✓	✓
2	3 0	-13541.52	-13557.91	✓	✓	✓	✗
2	1 1	-13612.97	-13626.63	✓	✓	✓	✓
2	2 1	-13614.89	-13631.28	✓	✓	✓	✓
2	1 2	-13615.46	-13631.85	✓	✓	✓	✗
2	2 2	-13616.59	-13635.71	✓	✓	✓	✓
2	0 1	-13618.53	-13629.46	✓	✓	✓	✓
2	0 2	-13619.43	-13633.10	✓	✓	✓	✓
1	3	-13635.88	-13646.81	✓	✓	✓	✓
2	1 0	-13694.20	-13705.13	✓	✓	✓	✗
2	2 0	-13696.22	-13709.88	✓	✓	✓	✗
1	1	-13779.38	-13784.84	✓	✓	✓	✓
1	2	-13780.33	-13788.53	✓	✓	✓	✓
2	0 0	-13823.53	-13831.73	✓	✓	✓	✗
1	0	-13974.75	-13977.48	✓	✓	✓	✓

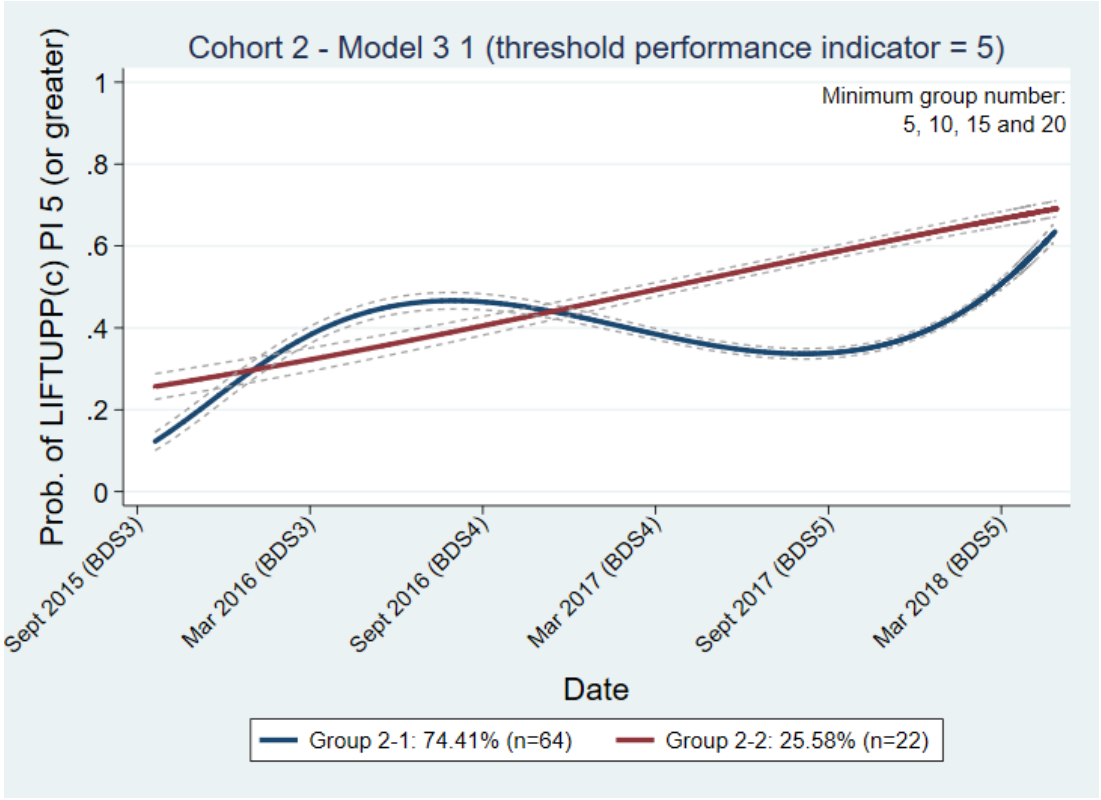
Cohort 3

Number of groups	Model	BIC^2 (n = 68)	BIC^3 (n = 20,817)	Contains at least X students per group, where X =			
				5	10	15	20
2	3 2	-13592.67	-13615.56	✓	✓	✓	✓
2	1 3	-13594.60	-13614.64	✓	✓	✓	✓
2	0 3	-13608.87	-13626.04	✓	✓	✓	✓
2	3 1	-13673.16	-13693.19	✓	✓	✗	✗
1	3	-13773.26	-13784.71	✓	✓	✓	✓
2	2 1	-13784.27	-13801.44	✓	✓	✓	✓
2	2 2	-13786.33	-13806.37	✓	✓	✓	✓
2	0 2	-13812.81	-13827.12	✓	✓	✓	✓
2	2 0	-13812.81	-13827.12	✓	✓	✓	✓
2	1 1	-13824.50	-13838.81	✓	✓	✓	✓
2	0 1	-13863.54	-13874.99	✓	✓	✓	✓
1	2	-14035.00	-14043.59	✓	✓	✓	✓
1	1	-14054.59	-14060.31	✓	✓	✓	✓
2	0 0	-14106.42	-14115.00	✓	✓	✓	✓
1	0	-14302.81	-14305.67	✓	✓	✓	✓

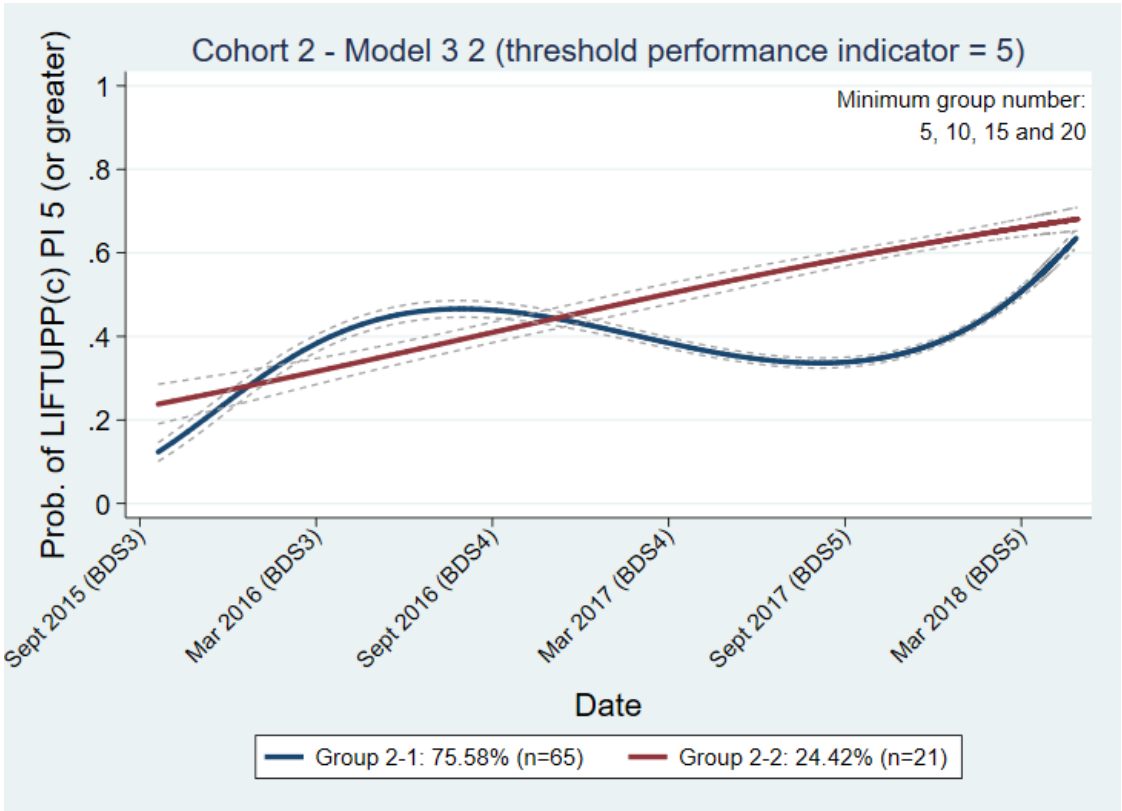
iii. Alternative LIFTUPP© models

Cohort 2

Model 3 1

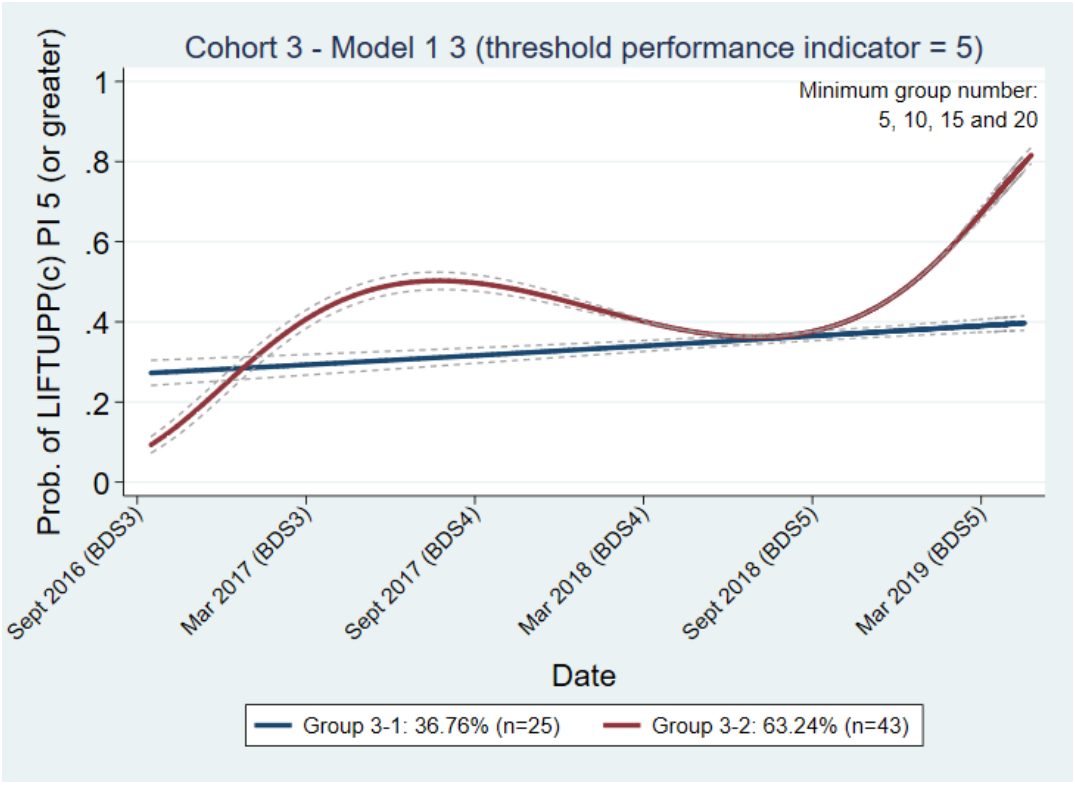


Model 3 2



Cohort 3

Model 1 3



Appendix 10 – Additional longitudinal evaluation of performance analysis data

i. Descriptive statistics

Summary statistics for the number of clinical longitudinal evaluation of performance (LEP) assessments completed per vocational dental practitioner (VDP).

Cohort	n	Number of LEP assessments	Mean	SD	Min	Q1	Median	Q3	Max
1	67	2,294	34.24	5.06	19.00	31.00	35.00	37.00	51.00
2	70	2,839	40.56	5.27	32.00	37.00	40.00	43.00	56.00
3	60	1,956	32.60	5.64	22.00	28.50	33.00	37.00	48.00

SD = Standard deviation

ii. *Group-based trajectory modelling - Models returned without errors*

BIC^2 = Bayesian information criterion for the total number of participants.

BIC^3 = Bayesian information criterion for the total number of observations.

Data models listed from least to most negative BIC^2.

Bernoulli data distribution: Threshold LEP score = 4

Cohort 1

Number of groups	Model	BIC^2 (n = 67)	BIC^3 (n = 2,294)	Contains at least X students per group, where X =			
				5	*10*	*15*	*20*
1	1	-419.99	-423.52	✓	✓	✓	✓
1	2	-421.82	-427.12	✓	✓	✓	✓
1	3	-423.91	-430.98	✓	✓	✓	✓
1	0	-443.55	-445.32	✓	✓	✓	✓

Cohort 2

Number of groups	Model	BIC^2 (n = 70)	BIC^3 (n = 2,839)	Contains at least X students per group, where X =			
				5	*10*	*15*	*20*
1	1	-381.88	-385.58	✓	✓	✓	✓
1	2	-383.98	-389.54	✓	✓	✓	✓
1	3	-385.46	-392.87	✓	✓	✓	✓
1	0	-431.63	-433.48	✓	✓	✓	✓

Cohort 3

Number of groups	Model	BIC^2 (n = 60)	BIC^3 (n = 1,956)	Contains at least X students per group, where X =			
				5	*10*	*15*	*20*
1	1	-275.46	-278.94	✓	✓	✓	✓
1	2	-277.11	-282.34	✓	✓	✓	✓
1	3	-279.06	-286.03	✓	✓	✓	✓
1	0	-300.59	-302.33	✓	✓	✓	✓

Bernoulli data distribution: Threshold LEP score = 5

Cohort 1

Number of groups	Model	BIC² (n = 67)	BIC³ (n = 2,294)	Contains at least X students per group, where X =			
				5	*10*	*15*	*20*
3	1 2 1	-867.01	-882.91	✗	✗	✗	✗
4	1 2 1 3	-868.18	-892.91	✗	✗	✗	✗
3	2 1 2	-868.83	-886.50	✗	✗	✗	✗
3	1 1 3	-868.95	-886.61	✗	✗	✗	✗
3	2 1 3	-868.98	-888.42	✗	✗	✗	✗
3	1 2 2	-869.10	-886.76	✗	✗	✗	✗
4	1 2 2 3	-870.60	-897.10	✗	✗	✗	✗
4	2 3 2 3	-872.10	-902.13	✗	✗	✗	✗
4	3 1 3 3	-872.61	-902.64	✗	✗	✗	✗
3	1 0 2	-873.24	-887.37	✗	✗	✗	✗
3	2 0 2	-874.38	-890.28	✗	✗	✗	✗
4	2 2 0 3	-874.59	-899.32	✗	✗	✗	✗
4	2 3 0 3	-876.67	-903.17	✗	✗	✗	✗
4	2 0 0 2	-878.30	-897.74	✗	✗	✗	✗
3	3 1 2	-880.46	-899.90	✗	✗	✗	✗
3	3 1 3	-881.49	-902.69	✗	✗	✗	✗
2	1 1	-882.80	-891.63	✗	✗	✗	✗
2	2 2	-885.72	-898.08	✗	✗	✗	✗
2	2 3	-887.26	-901.39	✗	✗	✗	✗
2	3 2	-887.81	-901.95	✗	✗	✗	✗
2	3 3	-889.35	-905.25	✗	✗	✗	✗
2	0 2	-891.42	-900.25	✓	✓	✓	✓
3	0 2 3	-892.98	-910.65	✗	✗	✗	✗
3	3 2 3	-893.09	-916.05	✗	✗	✗	✗

3	0 3 3	-895.08	-914.52	✗	✗	✗	✗
3	0 1 3	-907.29	-923.19	✗	✗	✗	✗
1	1	-913.69	-917.22	✓	✓	✓	✓
1	2	-915.51	-920.81	✓	✓	✓	✓
1	3	-917.49	-924.55	✓	✓	✓	✓
1	0	-1018.89	-1020.66	✓	✓	✓	✓

Cohort 2

Number of groups	**Model**	**BIC^2 (n = 70)**	**BIC^3 (n = 2,839)**	**Contains at least X students per group, where X =**			
				5	***10***	***15***	***20***
1	1	-1007.10	-1010.80	✓	✓	✓	✓
1	3	-1008.67	-1016.08	✓	✓	✓	✓
1	2	-1008.85	-1014.40	✓	✓	✓	✓
1	0	-1125.26	-1127.11	✓	✓	✓	✓

Cohort 3

Number of groups	**Model**	**BIC^2 (n = 60)**	**BIC^3 (n = 1,956)**	**Contains at least X students per group, where X =**			
				5	***10***	***15***	***20***
1	1	-731.50	-734.99	✓	✓	✓	✓
1	3	-732.61	-739.58	✓	✓	✓	✓
1	2	-732.87	-738.10	✓	✓	✓	✓
1	0	-805.26	-807.00	✓	✓	✓	✓

Bernoulli data distribution: Threshold LEP score = 6

Cohort 1

Number of groups	Model	BIC^2 (n = 67)	BIC^3 (n = 2,294)	Contains at least X students per group, where X =			
				5	*10*	*15*	*20*
3	2 1 1	-1265.20	-1281.10	✗	✗	✗	✗
3	1 2 1	-1265.83	-1281.73	✗	✗	✗	✗
3	3 1 1	-1265.88	-1283.55	✗	✗	✗	✗
4	1 1 1 1	-1266.29	-1285.72	✗	✗	✗	✗
3	2 2 1	-1266.98	-1284.64	✗	✗	✗	✗
3	2 1 2	-1267.08	-1284.75	✗	✗	✗	✗
3	3 1 2	-1267.77	-1287.20	✗	✗	✗	✗
3	1 3 1	-1267.86	-1285.53	✗	✗	✗	✗
4	1 1 1 2	-1268.25	-1289.45	✗	✗	✗	✗
4	1 2 1 1	-1268.27	-1289.47	✗	✗	✗	✗
4	3 3 1 1	-1268.28	-1294.78	✗	✗	✗	✗
4	1 3 1 1	-1268.49	-1291.46	✗	✗	✗	✗
4	3 1 2 2	-1268.78	-1295.28	✗	✗	✗	✗
3	2 2 2	-1268.86	-1288.29	✗	✗	✗	✗
4	3 1 3 1	-1268.89	-1295.39	✗	✗	✗	✗
3	2 3 1	-1269.01	-1288.44	✗	✗	✗	✗
4	2 2 1 1	-1269.03	-1292.00	✗	✗	✗	✗
4	1 2 0 1	-1269.31	-1288.75	✗	✗	✗	✗
3	3 3 1	-1269.48	-1290.68	✗	✗	✗	✗
4	1 2 1 2	-1269.69	-1292.65	✗	✗	✗	✗
3	1 3 2	-1269.74	-1289.18	✗	✗	✗	✗
3	3 1 3	-1269.79	-1290.99	✗	✗	✗	✗
4	3 3 1 2	-1270.19	-1298.45	✗	✗	✗	✗
4	2 3 1 1	-1270.26	-1295.00	✗	✗	✗	✗

4	3 3 2 1	-1270.31	-1298.58	✗	✗	✗	✗
4	0 2 2 1	-1270.37	-1291.57	✗	✗	✗	✗
4	1 2 0 2	-1270.37	-1291.57	✗	✗	✗	✗
4	2 2 0 1	-1270.37	-1291.57	✗	✗	✗	✗
4	1 1 3 1	-1270.41	-1293.38	✗	✗	✗	✗
4	1 2 2 1	-1270.41	-1293.38	✗	✗	✗	✗
4	3 0 2 1	-1270.69	-1293.66	✗	✗	✗	✗
4	3 1 3 2	-1270.74	-1299.01	✗	✗	✗	✗
4	1 3 2 3	-1270.80	-1299.07	✗	✗	✗	✗
3	2 3 2	-1270.89	-1292.09	✗	✗	✗	✗
4	2 2 1 2	-1270.90	-1295.64	✗	✗	✗	✗
4	2 2 2 1	-1270.98	-1295.72	✗	✗	✗	✗
4	1 0 2 2	-1271.19	-1292.39	✗	✗	✗	✗
4	0 3 1 2	-1271.29	-1294.26	✗	✗	✗	✗
4	1 3 0 1	-1271.35	-1292.55	✗	✗	✗	✗
3	3 3 2	-1271.37	-1294.33	✗	✗	✗	✗
4	1 2 2 2	-1271.46	-1296.19	✗	✗	✗	✗
4	2 1 3 1	-1271.52	-1296.25	✗	✗	✗	✗
3	1 3 3	-1271.76	-1292.96	✗	✗	✗	✗
4	1 3 1 2	-1271.92	-1296.66	✗	✗	✗	✗
4	1 3 2 2	-1272.00	-1298.50	✗	✗	✗	✗
4	3 1 2 1	-1272.09	-1296.83	✗	✗	✗	✗
4	1 3 2 1	-1272.11	-1296.84	✗	✗	✗	✗
4	1 2 1 3	-1272.16	-1296.89	✗	✗	✗	✗
4	3 3 2 2	-1272.18	-1302.21	✗	✗	✗	✗
4	1 1 3 2	-1272.19	-1296.92	✗	✗	✗	✗
4	1 3 0 3	-1272.23	-1296.97	✗	✗	✗	✗
4	3 3 0 1	-1272.23	-1296.97	✗	✗	✗	✗
4	2 2 0 2	-1272.25	-1295.21	✗	✗	✗	✗
4	1 3 3 3	-1272.41	-1302.44	✗	✗	✗	✗

4	3 3 3 1	-1272.41	-1302.44	✗	✗	✗	✗
4	1 3 0 2	-1272.42	-1295.39	✗	✗	✗	✗
4	2 3 0 1	-1272.42	-1295.39	✗	✗	✗	✗
4	2 3 1 2	-1272.84	-1299.34	✗	✗	✗	✗
3	2 3 3	-1272.91	-1295.88	✗	✗	✗	✗
4	2 2 1 3	-1272.92	-1299.42	✗	✗	✗	✗
4	2 3 2 1	-1272.94	-1299.44	✗	✗	✗	✗
4	1 2 0 3	-1273.21	-1296.17	✗	✗	✗	✗
4	1 2 2 3	-1273.32	-1299.82	✗	✗	✗	✗
3	3 3 3	-1273.39	-1298.12	✗	✗	✗	✗
4	1 3 3 2	-1273.81	-1302.08	✗	✗	✗	✗
4	1 3 1 3	-1273.94	-1300.44	✗	✗	✗	✗
4	2 3 3 1	-1273.95	-1302.21	✗	✗	✗	✗
4	3 3 0 2	-1274.10	-1300.60	✗	✗	✗	✗
4	1 3 3 1	-1274.20	-1300.70	✗	✗	✗	✗
4	2 2 0 3	-1274.27	-1299.00	✗	✗	✗	✗
4	1 1 3 3	-1274.28	-1300.78	✗	✗	✗	✗
4	2 3 0 2	-1274.30	-1299.03	✗	✗	✗	✗
4	3 3 3 2	-1274.38	-1306.19	✓	✗	✗	✗
4	2 3 1 3	-1274.86	-1303.13	✗	✗	✗	✗
4	2 1 3 3	-1275.40	-1303.67	✗	✗	✗	✗
4	3 3 1 3	-1276.13	-1306.16	✗	✗	✗	✗
4	3 3 3 3	-1276.30	-1309.86	✗	✗	✗	✗
4	2 3 0 3	-1276.32	-1302.82	✗	✗	✗	✗
4	2 3 2 3	-1276.83	-1306.86	✗	✗	✗	✗
4	2 3 3 2	-1276.90	-1306.93	✗	✗	✗	✗
4	3 1 3 3	-1278.09	-1308.12	✗	✗	✗	✗
4	2 3 3 3	-1278.92	-1310.72	✗	✗	✗	✗
2	1 1	-1280.60	-1289.43	✓	✓	✓	✓
2	2 1	-1282.40	-1293.00	✓	✓	✓	✓

3	1 0 1	-1282.43	-1294.79	✗	✗	✗	✗
2	3 1	-1282.77	-1295.13	✓	✓	✓	✓
2	2 2	-1284.29	-1296.65	✓	✓	✓	✓
3	2 0 1	-1284.39	-1298.52	✓	✗	✗	✗
2	1 3	-1284.53	-1296.89	✓	✓	✓	✓
3	0 2 1	-1284.72	-1298.86	✗	✗	✗	✗
2	3 2	-1284.73	-1298.87	✓	✓	✓	✓
3	0 2 2	-1285.32	-1301.22	✓	✗	✗	✗
3	2 0 2	-1285.57	-1301.47	✓	✗	✗	✗
3	3 0 3	-1286.02	-1305.45	✓	✗	✗	✗
2	3 3	-1286.19	-1302.09	✓	✓	✓	✓
3	0 3 1	-1286.76	-1302.66	✗	✗	✗	✗
3	0 3 2	-1287.08	-1304.75	✓	✗	✗	✗
3	3 0 2	-1287.47	-1305.14	✓	✗	✗	✗
3	0 3 3	-1288.74	-1308.17	✗	✗	✗	✗
4	2 2 0 0	-1289.41	-1308.85	✗	✗	✗	✗
3	0 0 1	-1307.24	-1317.84	✗	✗	✗	✗
2	1 0	-1315.03	-1322.10	✓	✓	✓	✓
2	2 0	-1316.88	-1325.71	✓	✓	✓	✓
2	3 0	-1318.85	-1329.45	✓	✓	✓	✓
3	1 0 0	-1319.23	-1329.83	✓	✗	✗	✗
2	0 1	-1323.64	-1330.71	✓	✗	✗	✗
2	0 2	-1325.53	-1334.36	✓	✗	✗	✗
2	0 3	-1327.52	-1338.12	✓	✗	✗	✗
3	0 1 0	-1327.85	-1338.45	✗	✗	✗	✗
3	0 2 0	-1329.73	-1342.10	✗	✗	✗	✗
3	0 3 0	-1331.72	-1345.85	✗	✗	✗	✗
1	1	-1347.10	-1350.64	✓	✓	✓	✓
1	2	-1349.20	-1354.50	✓	✓	✓	✓
1	3	-1351.26	-1358.32	✓	✓	✓	✓

2	0 0	-1447.76	-1453.06	✓	✓	✓	✓
1	0	-1505.68	-1507.44	✓	✓	✓	✓

Cohort 2

Number of groups	Model	BIC^2 (n = 70)	BIC^3 (n = 2,839)	Contains at least X students per group, where X =			
				5	*10*	*15*	*20*
4	2 3 1 3	-1505.93	-1535.55	✓	✓	✗	✗
4	1 3 1 2	-1505.96	-1531.88	✓	✓	✗	✗
4	1 1 1 2	-1506.00	-1528.21	✓	✓	✗	✗
4	1 1 3 3	-1507.33	-1535.10	✓	✓	✗	✗
4	2 2 1 3	-1507.63	-1535.40	✓	✓	✗	✗
4	2 3 2 3	-1507.66	-1539.13	✓	✓	✗	✗
4	2 1 1 2	-1507.93	-1532.00	✓	✓	✗	✗
4	1 3 2 3	-1508.37	-1537.99	✓	✓	✗	✗
4	1 2 1 3	-1509.15	-1535.07	✓	✓	✗	✗
4	2 2 2 3	-1509.25	-1538.87	✓	✓	✗	✗
4	3 2 2 3	-1510.38	-1541.85	✓	✗	✗	✗
4	1 3 3 3	-1510.48	-1541.95	✓	✓	✗	✗
4	3 1 3 3	-1510.48	-1541.95	✗	✗	✗	✗
r4	3 3 3 3	-1511.34	-1546.52	✓	✓	✗	✗
4	2 1 3 2	-1511.90	-1539.67	✓	✓	✗	✗
4	2 2 2 2	-1511.90	-1539.67	✓	✓	✗	✗
4	2 3 3 3	-1512.49	-1545.81	✓	✗	✗	✗
4	3 0 1 3	-1513.09	-1539.01	✗	✗	✗	✗
4	2 1 3 3	-1513.13	-1542.75	✓	✓	✗	✗
3	1 2 3	-1513.19	-1533.56	✓	✓	✗	✗
4	1 0 2 3	-1513.98	-1538.05	✗	✗	✗	✗
4	0 3 2 2	-1514.45	-1540.37	✗	✗	✗	✗
3	3 1 2	-1514.62	-1534.99	✓	✓	✗	✗

3	1 3 3	-1515.29	-1537.50	✓	✓	✗	✗
4	0 3 2 3	-1515.98	-1543.75	✗	✗	✗	✗
3	3 1 3	-1516.27	-1538.49	✓	✓	✗	✗
3	1 2 1	-1516.42	-1533.09	✓	✓	✗	✗
4	1 1 0 2	-1516.44	-1536.81	✓	✗	✗	✗
3	2 3 3	-1516.78	-1540.85	✓	✓	✗	✗
4	1 2 0 3	-1517.52	-1541.59	✓	✓	✗	✗
3	1 2 2	-1517.73	-1536.25	✓	✓	✗	✗
4	2 0 3 3	-1517.91	-1545.68	✗	✗	✗	✗
4	0 3 3 3	-1518.08	-1547.70	✗	✗	✗	✗
4	1 2 0 2	-1518.38	-1540.59	✓	✗	✗	✗
3	2 1 2	-1519.00	-1537.52	✓	✓	✗	✗
4	2 3 0 2	-1519.12	-1545.04	✓	✗	✗	✗
3	2 2 1	-1519.29	-1539.65	✗	✗	✗	✗
3	2 2 2	-1519.29	-1539.65	✓	✓	✗	✗
4	3 2 1 1	-1519.32	-1545.24	✗	✗	✗	✗
r4	3 2 2 2	-1519.71	-1549.33	✗	✗	✗	✗
3	1 3 2	-1520.26	-1540.62	✓	✓	✗	✗
4	3 3 1 2	-1520.27	-1549.90	✓	✗	✗	✗
4	0 1 3 3	-1521.10	-1547.02	✗	✗	✗	✗
4	3 2 0 2	-1521.46	-1547.38	✓	✗	✗	✗
3	2 2 3	-1521.82	-1544.04	✓	✓	✓	✗
3	1 1 1	-1522.27	-1537.08	✓	✓	✗	✗
4	0 1 1 3	-1523.46	-1545.67	✗	✗	✗	✗
4	0 1 0 3	-1525.31	-1545.67	✗	✗	✗	✗
4	1 0 0 3	-1525.31	-1545.67	✗	✗	✗	✗
4	2 1 3 1	-1526.07	-1551.99	✗	✗	✗	✗
4	0 2 0 3	-1526.21	-1548.42	✗	✗	✗	✗
4	0 2 0 2	-1527.41	-1547.77	✗	✗	✗	✗
3	3 0 3	-1529.99	-1550.36	✓	✓	✗	✗

3	0 1 3	-1530.62	-1547.28	✗	✗	✗	✗
3	2 0 2	-1531.77	-1548.43	✓	✗	✗	✗
4	0 1 3 0	-1532.73	-1553.09	✗	✗	✗	✗
4	0 1 3 1	-1534.85	-1557.07	✗	✗	✗	✗
4	0 0 2 1	-1535.06	-1553.58	✗	✗	✗	✗
3	1 0 1	-1535.74	-1548.70	✓	✓	✗	✗
3	2 0 1	-1537.58	-1552.39	✓	✓	✗	✗
2	3 2	-1538.91	-1553.72	✓	✓	✓	✓
2	1 3	-1539.61	-1552.57	✓	✓	✓	✓
2	2 2	-1540.15	-1553.11	✓	✓	✓	✓
2	3 3	-1540.41	-1557.07	✓	✓	✓	✓
3	1 2 0	-1542.42	-1557.23	✗	✗	✗	✗
2	3 1	-1542.76	-1555.72	✓	✓	✓	✓
2	1 1	-1542.92	-1552.17	✓	✓	✓	✓
3	1 3 0	-1543.86	-1560.52	✗	✗	✗	✗
3	2 2 0	-1544.40	-1561.06	✗	✗	✗	✗
2	2 1	-1544.42	-1555.53	✓	✓	✓	✓
3	3 3 0	-1544.66	-1565.02	✗	✗	✗	✗
3	2 3 0	-1545.80	-1564.31	✗	✗	✗	✗
3	3 1 0	-1547.01	-1563.67	✗	✗	✗	✗
3	1 1 0	-1547.16	-1560.12	✗	✗	✗	✗
3	2 1 0	-1548.67	-1563.48	✗	✗	✗	✗
3	0 0 3	-1550.06	-1564.87	✓	✗	✗	✗
4	0 0 0 3	-1554.30	-1572.82	✗	✗	✗	✗
3	0 0 2	-1578.46	-1591.42	✓	✓	✗	✗
2	0 3	-1580.81	-1591.92	✓	✗	✗	✗
2	0 2	-1581.71	-1590.97	✓	✗	✗	✗
2	0 1	-1584.44	-1591.85	✓	✗	✗	✗
2	3 0	-1590.07	-1601.18	✓	✓	✗	✗
3	0 1 2	-1592.94	-1607.75	✗	✗	✗	✗

2	2 0	-1593.33	-1602.59	✓	✓	✗	✗
3	3 0 0	-1594.32	-1609.13	✗	✗	✗	✗
2	1 0	-1597.15	-1604.55	✓	✓	✗	✗
3	2 0 0	-1597.58	-1610.54	✗	✗	✗	✗
1	3	-1641.12	-1648.52	✓	✓	✓	✓
1	2	-1642.65	-1648.21	✓	✓	✓	✓
1	1	-1645.16	-1648.87	✓	✓	✓	✓
2	0 0	-1789.35	-1794.90	✓	✓	✓	✓
3	0 0 1	-1795.72	-1806.83	✗	✗	✗	✗
1	0	-1866.72	-1868.57	✓	✓	✓	✓

Cohort 3

Number of groups	**Model**	**BIC^2 (n = 60)**	**BIC^3 (n = 1,956)**	**Contains at least X students per group, where X =**			
				5	***10***	***15***	***20***
4	1 0 0 3	-1123.64	-1142.81	✗	✗	✗	✗
3	3 0 3	-1126.03	-1145.19	✓	✗	✗	✗
4	2 0 2 3	-1126.86	-1151.25	✗	✗	✗	✗
4	3 0 2 3	-1128.69	-1154.82	✗	✗	✗	✗
2	1 3	-1128.86	-1141.06	✓	✓	✓	✓
4	2 2 1 3	-1129.81	-1155.94	✗	✗	✗	✗
2	1 1	-1130.08	-1138.80	✓	✓	✓	✓
2	2 2	-1130.18	-1142.38	✓	✓	✓	✓
2	3 2	-1132.05	-1145.99	✓	✓	✓	✓
4	3 2 2 3	-1132.05	-1161.66	✗	✗	✗	✗
2	2 1	-1132.08	-1142.53	✓	✓	✓	✓
2	3 3	-1132.89	-1148.57	✓	✓	✓	✓
4	3 0 1 3	-1133.65	-1158.04	✗	✗	✗	✗
4	3 3 2 3	-1133.93	-1165.29	✗	✗	✗	✗
3	2 3 3	-1133.96	-1156.61	✗	✗	✗	✗

2	3 1	-1134.02	-1146.22	✓	✓	✓	✓
3	3 1 3	-1135.22	-1156.12	✓	✗	✗	✗
3	0 0 3	-1144.57	-1158.50	✓	✗	✗	✗
2	1 0	-1145.85	-1152.82	✓	✓	✗	✗
2	2 0	-1147.28	-1155.99	✓	✓	✗	✗
2	3 0	-1149.03	-1159.49	✓	✓	✗	✗
3	2 0 3	-1151.39	-1168.81	✗	✗	✗	✗
2	0 2	-1159.51	-1168.22	✓	✓	✗	✗
2	0 3	-1160.01	-1170.46	✓	✓	✗	✗
2	0 1	-1160.38	-1167.35	✓	✓	✗	✗
1	1	-1170.08	-1173.56	✓	✓	✓	✓
1	2	-1170.76	-1175.98	✓	✓	✓	✓
1	3	-1171.40	-1178.36	✓	✓	✓	✓
2	0 0	-1269.94	-1275.16	✓	✓	✓	✓
1	0	-1298.84	-1300.58	✓	✓	✓	✓

Bernoulli data distribution: Threshold LEP score = 7

Cohort 1

Number of groups	Model	BIC^2 (n = 67)	BIC^3 (n = 2,294)	Contains at least X students per group, where X =			
				5	*10*	*15*	*20*
3	1 1 3	-1185.98	-1203.64	✓	✗	✗	✗
4	1 3 1 2	-1186.63	-1211.36	✓	✗	✗	✗
3	3 1 2	-1186.82	-1206.26	✓	✗	✗	✗
3	2 1 2	-1186.99	-1204.65	✓	✗	✗	✗
3	3 1 3	-1187.07	-1208.27	✓	✗	✗	✗
3	1 2 2	-1187.31	-1204.97	✓	✗	✗	✗
4	1 1 1 1	-1187.36	-1206.79	✗	✗	✗	✗
4	1 2 1 2	-1187.90	-1210.86	✓	✗	✗	✗
3	1 1 1	-1188.12	-1202.26	✓	✗	✗	✗
3	3 2 2	-1188.40	-1209.61	✓	✗	✗	✗
3	2 2 2	-1188.43	-1207.87	✓	✗	✗	✗
3	1 3 2	-1188.48	-1207.91	✓	✗	✗	✗
4	1 1 3 2	-1188.55	-1213.28	✓	✗	✗	✗
4	1 3 2 2	-1188.62	-1215.12	✓	✗	✗	✗
3	3 2 3	-1188.66	-1211.63	✓	✗	✗	✗
3	1 3 3	-1188.68	-1209.88	✓	✗	✗	✗
4	1 3 1 1	-1188.70	-1211.66	✓	✗	✗	✗
4	1 3 1 0	-1188.71	-1209.91	✗	✗	✗	✗
3	1 0 1	-1188.87	-1201.23	✗	✗	✗	✗
4	1 3 2 3	-1188.89	-1217.15	✓	✗	✗	✗
3	3 1 1	-1189.19	-1206.85	✓	✗	✗	✗
4	1 1 2 0	-1189.31	-1208.75	✗	✗	✗	✗
3	2 1 1	-1189.39	-1205.29	✓	✗	✗	✗
3	3 3 2	-1189.45	-1212.41	✓	✗	✗	✗

3	2 3 2	-1189.49	-1210.69	✓	✗	✗	✗
4	2 3 1 1	-1189.50	-1214.23	✓	✗	✗	✗
4	1 1 1 0	-1189.59	-1207.26	✗	✗	✗	✗
3	3 3 3	-1189.65	-1214.38	✓	✗	✗	✗
4	3 1 3 2	-1189.68	-1217.94	✓	✗	✗	✗
3	2 3 3	-1189.70	-1212.67	✓	✗	✗	✗
3	3 1 0	-1189.85	-1205.75	✗	✗	✗	✗
3	1 2 1	-1189.87	-1205.77	✓	✗	✗	✗
4	3 1 3 3	-1189.88	-1219.92	✓	✗	✗	✗
4	1 0 3 2	-1189.90	-1212.87	✗	✗	✗	✗
3	2 1 0	-1190.15	-1204.29	✗	✗	✗	✗
4	2 3 1 2	-1190.16	-1216.66	✓	✗	✗	✗
4	2 1 3 2	-1190.33	-1216.83	✓	✗	✗	✗
4	3 1 0 1	-1190.41	-1211.61	✗	✗	✗	✗
4	2 1 3 3	-1190.50	-1218.77	✓	✗	✗	✗
4	2 1 2 0	-1190.61	-1211.81	✗	✗	✗	✗
4	2 3 3 3	-1190.72	-1222.52	✓	✗	✗	✗
4	3 1 1 0	-1190.77	-1211.97	✗	✗	✗	✗
3	1 2 0	-1190.88	-1205.01	✓	✗	✗	✗
4	1 2 2 0	-1190.92	-1212.12	✗	✗	✗	✗
3	3 2 1	-1190.95	-1210.38	✓	✗	✗	✗
3	2 2 1	-1191.02	-1208.68	✓	✗	✗	✗
4	1 1 3 1	-1191.06	-1214.03	✓	✗	✗	✗
3	1 3 1	-1191.07	-1208.74	✓	✗	✗	✗
4	1 1 3 0	-1191.13	-1212.33	✗	✗	✗	✗
4	1 2 0 1	-1191.27	-1210.70	✗	✗	✗	✗
4	1 2 1 0	-1191.27	-1210.70	✗	✗	✗	✗
4	1 2 2 1	-1191.32	-1214.29	✗	✗	✗	✗
4	2 3 2 2	-1191.59	-1219.86	✓	✗	✗	✗
4	2 3 2 0	-1191.62	-1216.35	✗	✗	✗	✗

4	0 3 2 2	-1191.71	-1216.44	✗	✗	✗	✗
4	0 1 1 0	-1191.74	-1207.64	✗	✗	✗	✗
3	3 2 0	-1191.88	-1209.54	✗	✗	✗	✗
4	0 3 2 3	-1191.96	-1218.46	✗	✗	✗	✗
4	0 1 3 3	-1191.97	-1216.71	✗	✗	✗	✗
4	1 0 3 3	-1191.97	-1216.71	✗	✗	✗	✗
4	2 2 2 0	-1192.00	-1214.97	✗	✗	✗	✗
3	3 3 1	-1192.03	-1213.23	✓	✗	✗	✗
4	1 3 3 1	-1192.10	-1218.60	✓	✗	✗	✗
3	2 0 2	-1192.11	-1208.01	✓	✗	✗	✗
3	2 2 0	-1192.11	-1208.01	✓	✗	✗	✗
3	2 3 1	-1192.11	-1211.55	✓	✗	✗	✗
3	1 0 3	-1192.15	-1208.05	✓	✗	✗	✗
3	1 3 0	-1192.15	-1208.05	✓	✗	✗	✗
4	0 3 1 1	-1192.15	-1213.35	✗	✗	✗	✗
4	0 2 2 2	-1192.19	-1215.15	✗	✗	✗	✗
4	2 2 1 0	-1192.36	-1213.57	✗	✗	✗	✗
4	1 3 3 0	-1192.37	-1217.10	✗	✗	✗	✗
4	3 1 3 0	-1192.37	-1217.10	✗	✗	✗	✗
4	0 1 3 2	-1192.54	-1215.51	✗	✗	✗	✗
4	1 3 0 2	-1192.54	-1215.51	✗	✗	✗	✗
4	0 3 0 1	-1192.56	-1212.00	✗	✗	✗	✗
4	0 3 1 0	-1192.56	-1212.00	✗	✗	✗	✗
4	3 0 0 1	-1192.56	-1212.00	✗	✗	✗	✗
4	3 0 1 0	-1192.56	-1212.00	✗	✗	✗	✗
4	2 1 2 1	-1192.67	-1215.64	✗	✗	✗	✗
4	3 0 3 2	-1192.73	-1219.23	✗	✗	✗	✗
4	3 1 1 1	-1192.79	-1215.75	✗	✗	✗	✗
4	1 3 0 1	-1192.80	-1214.00	✗	✗	✗	✗
4	2 1 3 1	-1192.90	-1217.64	✓	✗	✗	✗

4	2 3 3 1	-1192.90	-1221.17	✓	✗	✗	✗
4	0 3 3 3	-1192.92	-1221.19	✗	✗	✗	✗
4	3 0 3 3	-1192.92	-1221.19	✗	✗	✗	✗
4	2 1 2 2	-1192.93	-1217.67	✗	✗	✗	✗
4	0 2 1 1	-1192.99	-1212.42	✗	✗	✗	✗
4	3 2 2 2	-1193.02	-1221.29	✗	✗	✗	✗
3	3 3 0	-1193.05	-1212.49	✓	✗	✗	✗
4	1 3 2 0	-1193.09	-1216.06	✗	✗	✗	✗
4	2 1 3 0	-1193.09	-1216.06	✗	✗	✗	✗
4	0 1 2 1	-1193.11	-1212.54	✗	✗	✗	✗
4	1 0 2 1	-1193.11	-1212.54	✗	✗	✗	✗
4	2 3 3 0	-1193.20	-1219.70	✗	✗	✗	✗
4	2 0 3 2	-1193.22	-1217.95	✗	✗	✗	✗
3	2 3 0	-1193.29	-1210.95	✓	✗	✗	✗
4	1 2 2 2	-1193.31	-1218.05	✗	✗	✗	✗
4	3 3 0 2	-1193.54	-1220.04	✗	✗	✗	✗
4	0 2 0 1	-1193.57	-1211.24	✗	✗	✗	✗
4	0 2 1 0	-1193.57	-1211.24	✗	✗	✗	✗
4	2 0 0 1	-1193.57	-1211.24	✗	✗	✗	✗
4	2 3 0 2	-1193.57	-1218.30	✗	✗	✗	✗
4	3 3 2 0	-1193.60	-1220.10	✗	✗	✗	✗
4	2 3 0 1	-1193.81	-1216.78	✗	✗	✗	✗
4	2 3 1 0	-1193.81	-1216.78	✗	✗	✗	✗
4	0 1 2 0	-1193.82	-1211.49	✗	✗	✗	✗
4	1 0 2 0	-1193.82	-1211.49	✗	✗	✗	✗
4	1 2 0 0	-1193.82	-1211.49	✗	✗	✗	✗
4	3 3 0 1	-1193.83	-1218.57	✗	✗	✗	✗
4	3 3 1 0	-1193.83	-1218.57	✗	✗	✗	✗
4	1 3 3 3	-1194.03	-1224.07	✓	✗	✗	✗
4	2 2 2 1	-1194.07	-1218.80	✗	✗	✗	✗

4	0 3 2 1	-1194.15	-1217.11	✗	✗	✗	✗
4	1 0 3 1	-1194.32	-1215.52	✗	✗	✗	✗
4	2 2 1 1	-1194.39	-1217.35	✗	✗	✗	✗
4	2 2 2 2	-1194.40	-1220.90	✗	✗	✗	✗
4	0 3 0 2	-1194.66	-1215.86	✗	✗	✗	✗
4	0 3 2 0	-1194.66	-1215.86	✗	✗	✗	✗
4	3 0 0 2	-1194.66	-1215.86	✗	✗	✗	✗
4	0 2 1 2	-1194.74	-1215.94	✗	✗	✗	✗
4	0 2 2 1	-1194.74	-1215.94	✗	✗	✗	✗
4	2 0 2 1	-1194.74	-1215.94	✗	✗	✗	✗
4	1 3 2 1	-1195.15	-1219.89	✗	✗	✗	✗
4	0 1 3 0	-1195.17	-1214.60	✗	✗	✗	✗
4	1 0 3 0	-1195.17	-1214.60	✗	✗	✗	✗
4	3 1 1 2	-1195.19	-1219.92	✗	✗	✗	✗
4	0 3 3 1	-1195.23	-1219.96	✗	✗	✗	✗
4	3 0 3 1	-1195.23	-1219.96	✗	✗	✗	✗
4	3 3 1 2	-1195.52	-1223.79	✗	✗	✗	✗
4	0 2 2 0	-1195.60	-1215.04	✗	✗	✗	✗
4	2 0 2 0	-1195.60	-1215.04	✗	✗	✗	✗
4	3 3 2 1	-1195.67	-1223.94	✗	✗	✗	✗
4	2 0 3 1	-1195.82	-1218.78	✗	✗	✗	✗
4	0 3 0 3	-1195.93	-1218.89	✗	✗	✗	✗
4	0 3 3 0	-1195.93	-1218.89	✗	✗	✗	✗
4	3 0 0 3	-1195.93	-1218.89	✗	✗	✗	✗
4	3 0 3 0	-1195.93	-1218.89	✗	✗	✗	✗
4	3 3 2 2	-1195.99	-1226.02	✗	✗	✗	✗
4	1 3 0 0	-1196.36	-1215.79	✗	✗	✗	✗
4	2 0 3 0	-1196.83	-1218.03	✗	✗	✗	✗
4	2 3 2 3	-1197.67	-1227.71	✗	✗	✗	✗
4	2 3 1 3	-1199.60	-1227.87	✓	✗	✗	✗

3	0 1 1	-1202.50	-1214.86	✓	✗	✗	✗
3	0 3 1	-1203.53	-1219.43	✓	✗	✗	✗
2	1 1	-1204.23	-1213.07	✓	✓	✓	✓
3	0 2 1	-1204.56	-1218.69	✓	✗	✗	✗
3	2 0 1	-1204.56	-1218.69	✓	✗	✗	✗
3	0 3 2	-1204.61	-1222.27	✓	✗	✗	✗
3	3 0 2	-1204.61	-1222.27	✓	✗	✗	✗
2	3 1	-1204.74	-1217.10	✓	✓	✓	✓
2	2 1	-1205.54	-1216.14	✓	✓	✓	✓
3	0 2 2	-1205.81	-1221.71	✓	✗	✗	✗
2	1 2	-1206.10	-1216.70	✓	✓	✓	✓
3	3 0 3	-1206.37	-1225.80	✓	✗	✗	✗
2	3 2	-1206.53	-1220.66	✓	✓	✓	✓
2	2 2	-1207.49	-1219.85	✓	✓	✓	✓
3	0 1 3	-1207.59	-1223.49	✓	✗	✗	✗
2	1 3	-1208.18	-1220.55	✓	✓	✓	✓
2	3 3	-1208.62	-1224.52	✓	✓	✓	✓
3	0 3 3	-1209.27	-1228.70	✓	✓	✗	✗
4	0 3 1 2	-1209.58	-1232.55	✗	✗	✗	✗
2	2 3	-1209.59	-1223.72	✓	✓	✓	✓
3	0 1 0	-1213.14	-1223.74	✓	✗	✗	✗
3	1 0 0	-1213.14	-1223.74	✓	✗	✗	✗
4	0 0 3 3	-1213.39	-1236.36	✗	✗	✗	✗
3	0 0 2	-1215.22	-1227.58	✓	✗	✗	✗
3	0 2 0	-1215.22	-1227.58	✓	✗	✗	✗
3	2 0 0	-1215.22	-1227.58	✓	✗	✗	✗
3	0 3 0	-1216.00	-1230.13	✓	✗	✗	✗
3	3 0 0	-1216.00	-1230.13	✓	✗	✗	✗
4	0 0 1 0	-1216.87	-1231.01	✓	✗	✗	✗
4	0 1 0 0	-1217.35	-1231.48	✗	✗	✗	✗

4	0 0 0 2	-1218.96	-1234.86	✓	✗	✗	✗
4	0 0 2 0	-1218.96	-1234.86	✓	✗	✗	✗
4	0 0 0 3	-1219.86	-1237.53	✓	✗	✗	✗
4	0 0 3 0	-1219.86	-1237.53	✓	✗	✗	✗
4	3 0 0 0	-1219.86	-1237.53	✓	✗	✗	✗
2	0 1	-1237.43	-1244.50	✓	✓	✓	✓
2	0 2	-1239.16	-1247.99	✓	✓	✓	✓
2	0 3	-1239.43	-1250.03	✓	✓	✓	✓
3	0 0 1	-1240.58	-1251.18	✓	✗	✗	✗
3	0 0 3	-1242.72	-1256.85	✓	✗	✗	✗
2	3 0	-1244.77	-1255.37	✓	✗	✗	✗
2	1 0	-1244.83	-1251.89	✓	✗	✗	✗
2	2 0	-1245.87	-1254.70	✓	✗	✗	✗
1	3	-1285.66	-1292.73	✓	✓	✓	✓
1	1	-1286.08	-1289.62	✓	✓	✓	✓
1	2	-1286.44	-1291.74	✓	✓	✓	✓
4	2 0 3 3	-1315.28	-1341.78	✓	✗	✗	✗
3	0 0 0	-1383.08	-1391.91	✓	✗	✗	✗
4	0 0 0 0	-1385.07	-1397.44	✓	✗	✗	✗
2	0 0	-1395.54	-1400.84	✓	✓	✓	✓
1	0	-1463.19	-1464.95	✓	✓	✓	✓

Cohort 2

Number of groups	Model	BIC[2] (n = 70)	BIC[3] (n = 2,839)	Contains at least X students per group, where X =			
				5	*10*	*15*	*20*
4	1 1 3 2	-1442.14	-1468.06	✓	✓	✗	✗
4	1 1 3 0	-1444.73	-1466.95	✗	✗	✗	✗
4	1 0 3 1	-1445.26	-1467.48	✗	✗	✗	✗
4	1 3 3 1	-1445.76	-1473.53	✗	✗	✗	✗
4	3 1 3 1	-1445.76	-1473.53	✗	✗	✗	✗
4	2 1 3 2	-1445.93	-1473.70	✗	✗	✗	✗
4	1 2 1 1	-1446.38	-1468.60	✓	✓	✗	✗
4	2 1 3 0	-1446.52	-1470.59	✗	✗	✗	✗
4	0 1 3 2	-1446.97	-1471.04	✗	✗	✗	✗
4	1 0 3 2	-1446.97	-1471.04	✗	✗	✗	✗
4	1 1 1 0	-1447.03	-1465.55	✗	✗	✗	✗
4	2 1 2 2	-1447.18	-1473.10	✗	✗	✗	✗
4	2 0 3 1	-1447.36	-1471.42	✗	✗	✗	✗
4	2 3 3 1	-1447.51	-1477.13	✗	✗	✗	✗
4	0 1 3 0	-1447.97	-1468.33	✗	✗	✗	✗
4	1 0 3 0	-1447.97	-1468.33	✗	✗	✗	✗
4	3 1 0 0	-1447.97	-1468.33	✗	✗	✗	✗
4	1 2 0 2	-1448.08	-1470.30	✗	✗	✗	✗
4	1 2 2 0	-1448.08	-1470.30	✗	✗	✗	✗
4	2 1 2 0	-1448.08	-1470.30	✗	✗	✗	✗
4	3 1 1 1	-1448.09	-1472.16	✗	✗	✗	✗
4	1 3 0 3	-1448.22	-1474.13	✗	✗	✗	✗
4	1 3 3 0	-1448.22	-1474.13	✗	✗	✗	✗
4	3 1 3 0	-1448.22	-1474.13	✗	✗	✗	✗
4	1 0 2 2	-1448.37	-1470.59	✗	✗	✗	✗

4	0 3 3 1	-1448.85	-1474.77	✗	✗	✗	✗
4	3 3 3 1	-1449.05	-1480.52	✗	✗	✗	✗
4	2 0 3 2	-1449.06	-1474.98	✗	✗	✗	✗
4	1 2 1 0	-1449.07	-1469.44	✗	✗	✗	✗
4	1 3 0 2	-1449.33	-1473.39	✗	✗	✗	✗
4	1 3 2 0	-1449.33	-1473.39	✗	✗	✗	✗
4	1 0 2 0	-1449.43	-1467.95	✗	✗	✗	✗
4	2 0 2 1	-1449.51	-1471.72	✗	✗	✗	✗
4	2 2 2 0	-1449.90	-1473.97	✗	✗	✗	✗
4	2 3 3 0	-1449.98	-1477.75	✗	✗	✗	✗
4	2 0 3 0	-1450.05	-1472.27	✗	✗	✗	✗
4	0 1 0 1	-1450.28	-1466.94	✗	✗	✗	✗
4	0 1 1 0	-1450.28	-1466.94	✗	✗	✗	✗
4	1 0 0 1	-1450.28	-1466.94	✗	✗	✗	✗
4	1 0 1 0	-1450.28	-1466.94	✗	✗	✗	✗
4	1 3 0 1	-1450.35	-1472.56	✗	✗	✗	✗
4	1 3 1 0	-1450.35	-1472.56	✗	✗	✗	✗
4	3 1 1 0	-1450.35	-1472.56	✗	✗	✗	✗
4	2 3 1 2	-1450.58	-1478.35	✗	✗	✗	✗
4	0 3 2 1	-1450.81	-1474.88	✗	✗	✗	✗
4	2 3 0 2	-1451.09	-1477.01	✗	✗	✗	✗
4	2 3 2 0	-1451.09	-1477.01	✗	✗	✗	✗
4	3 0 1 1	-1451.30	-1473.52	✗	✗	✗	✗
4	0 3 0 3	-1451.48	-1475.55	✗	✗	✗	✗
4	0 3 3 0	-1451.48	-1475.55	✗	✗	✗	✗
4	3 0 3 0	-1451.48	-1475.55	✗	✗	✗	✗
4	3 3 3 0	-1451.51	-1481.14	✗	✗	✗	✗
4	0 2 0 2	-1451.55	-1471.92	✗	✗	✗	✗
4	0 2 2 0	-1451.55	-1471.92	✗	✗	✗	✗
4	2 0 0 2	-1451.55	-1471.92	✗	✗	✗	✗

4	2 0 2 0	-1451.55	-1471.92	✗	✗	✗	✗
4	0 3 1 2	-1452.02	-1476.09	✗	✗	✗	✗
4	3 3 1 2	-1452.11	-1481.74	✗	✗	✗	✗
4	2 3 0 1	-1452.12	-1476.19	✗	✗	✗	✗
4	2 3 1 0	-1452.12	-1476.19	✗	✗	✗	✗
4	0 2 0 1	-1452.36	-1470.87	✗	✗	✗	✗
4	0 2 1 0	-1452.36	-1470.87	✗	✗	✗	✗
4	2 0 0 1	-1452.36	-1470.87	✗	✗	✗	✗
4	2 0 1 0	-1452.36	-1470.87	✗	✗	✗	✗
4	3 3 0 2	-1452.64	-1480.41	✗	✗	✗	✗
4	3 3 2 0	-1452.64	-1480.41	✗	✗	✗	✗
4	0 3 0 2	-1452.83	-1475.05	✗	✗	✗	✗
4	0 3 2 0	-1452.83	-1475.05	✗	✗	✗	✗
4	3 3 0 1	-1453.65	-1479.57	✗	✗	✗	✗
4	3 3 1 0	-1453.65	-1479.57	✗	✗	✗	✗
4	0 3 0 1	-1453.66	-1474.02	✗	✗	✗	✗
4	0 3 1 0	-1453.66	-1474.02	✗	✗	✗	✗
4	3 0 1 0	-1453.66	-1474.02	✗	✗	✗	✗
4	0 3 2 3	-1453.87	-1481.64	✗	✗	✗	✗
3	1 1 1	-1454.38	-1469.19	✓	✓	✗	✗
3	2 1 2	-1454.66	-1473.17	✓	✓	✗	✗
3	1 2 2	-1454.75	-1473.26	✓	✓	✗	✗
3	1 3 2	-1454.80	-1475.16	✓	✓	✗	✗
3	1 3 3	-1455.26	-1477.47	✓	✓	✗	✗
4	1 1 2 3	-1455.47	-1481.39	✗	✗	✗	✗
3	1 3 1	-1456.16	-1474.67	✓	✓	✗	✗
3	3 1 2	-1456.19	-1476.56	✓	✓	✗	✗
3	1 1 2	-1456.23	-1472.89	✓	✓	✗	✗
3	1 2 1	-1456.51	-1473.17	✓	✓	✗	✗
3	2 2 2	-1456.71	-1477.07	✓	✓	✗	✗

3	2 3 2	-1456.71	-1478.92	✓	✓	✗	✗
3	2 3 3	-1457.19	-1481.25	✓	✓	✗	✗
4	2 1 3 3	-1457.29	-1486.91	✗	✗	✗	✗
3	3 1 1	-1457.75	-1476.27	✓	✓	✗	✗
3	2 3 1	-1458.09	-1478.46	✓	✓	✗	✗
3	0 1 1	-1458.14	-1471.10	✓	✗	✗	✗
4	1 3 3 3	-1458.15	-1489.62	✗	✗	✗	✗
3	3 2 2	-1458.23	-1480.45	✓	✓	✗	✗
3	3 3 2	-1458.28	-1482.35	✓	✓	✗	✗
3	2 2 1	-1458.51	-1477.02	✓	✓	✗	✗
4	0 3 3 2	-1458.56	-1486.33	✗	✗	✗	✗
4	1 3 2 3	-1458.64	-1488.26	✗	✗	✗	✗
3	3 3 3	-1458.75	-1484.66	✓	✓	✗	✗
4	3 0 1 2	-1459.26	-1483.32	✗	✗	✗	✗
4	1 2 2 1	-1459.56	-1483.62	✗	✗	✗	✗
3	3 1 3	-1459.65	-1481.87	✓	✓	✗	✗
3	3 3 1	-1459.65	-1481.87	✓	✓	✗	✗
3	0 2 2	-1459.77	-1476.43	✓	✗	✗	✗
3	2 0 2	-1459.77	-1476.43	✓	✗	✗	✗
4	1 1 1 1	-1459.77	-1480.13	✗	✗	✗	✗
4	2 1 3 1	-1459.77	-1485.69	✗	✗	✗	✗
3	3 2 1	-1460.01	-1480.37	✓	✓	✗	✗
3	0 2 1	-1460.20	-1475.01	✓	✗	✗	✗
3	2 0 1	-1460.20	-1475.01	✓	✗	✗	✗
3	0 3 3	-1460.70	-1481.07	✓	✗	✗	✗
3	3 0 3	-1460.70	-1481.07	✓	✗	✗	✗
3	0 3 2	-1460.84	-1479.35	✓	✓	✗	✗
3	3 0 2	-1460.84	-1479.35	✓	✓	✗	✗
4	1 3 1 1	-1460.87	-1484.94	✗	✗	✗	✗
4	2 3 2 1	-1461.30	-1489.07	✗	✗	✗	✗

4	2 2 2 1	-1461.36	-1487.28	✗	✗	✗	✗
3	0 3 1	-1461.37	-1478.03	✓	✗	✗	✗
3	3 0 1	-1461.37	-1478.03	✓	✗	✗	✗
4	2 3 1 3	-1461.44	-1491.06	✗	✗	✗	✗
4	3 3 3 3	-1461.68	-1496.86	✗	✗	✗	✗
4	0 0 2 2	-1462.43	-1482.80	✗	✗	✗	✗
4	0 0 2 1	-1462.97	-1481.49	✗	✗	✗	✗
4	3 3 2 1	-1463.00	-1492.62	✗	✗	✗	✗
4	0 0 3 2	-1463.06	-1485.28	✗	✗	✗	✗
4	0 1 3 1	-1463.43	-1485.65	✗	✗	✗	✗
4	1 0 2 1	-1463.74	-1484.11	✗	✗	✗	✗
4	0 0 3 1	-1463.91	-1484.28	✗	✗	✗	✗
4	3 3 1 1	-1464.36	-1492.13	✗	✗	✗	✗
4	0 2 1 1	-1466.31	-1486.68	✗	✗	✗	✗
4	2 0 1 1	-1466.57	-1486.94	✗	✗	✗	✗
4	3 0 3 1	-1466.79	-1492.71	✗	✗	✗	✗
4	3 3 2 3	-1467.25	-1500.58	✗	✗	✗	✗
4	0 3 1 1	-1467.45	-1489.67	✗	✗	✗	✗
4	2 0 1 2	-1468.44	-1490.65	✗	✗	✗	✗
4	0 3 2 2	-1468.91	-1494.83	✗	✗	✗	✗
4	3 0 3 2	-1468.91	-1496.68	✗	✗	✗	✗
4	0 3 3 3	-1471.29	-1500.91	✗	✗	✗	✗
4	3 1 0 1	-1472.01	-1494.23	✗	✗	✗	✗
3	1 0 3	-1473.00	-1489.67	✗	✗	✗	✗
3	1 3 0	-1473.00	-1489.67	✗	✗	✗	✗
4	2 3 2 2	-1473.37	-1502.99	✗	✗	✗	✗
3	3 2 3	-1473.86	-1497.93	✗	✗	✗	✗
3	2 3 0	-1474.04	-1492.55	✗	✗	✗	✗
3	1 2 0	-1474.76	-1489.57	✓	✗	✗	✗
3	1 1 0	-1475.21	-1488.17	✗	✗	✗	✗

3	3 3 0	-1476.03	-1496.39	✗	✗	✗	✗
3	2 2 0	-1476.17	-1492.83	✗	✗	✗	✗
3	2 1 0	-1476.41	-1491.22	✗	✗	✗	✗
4	2 3 0 0	-1477.54	-1499.76	✗	✗	✗	✗
4	1 2 0 0	-1477.65	-1496.17	✗	✗	✗	✗
3	3 2 0	-1478.10	-1496.62	✗	✗	✗	✗
3	3 1 0	-1478.30	-1494.96	✗	✗	✗	✗
4	3 3 0 0	-1479.33	-1503.40	✗	✗	✗	✗
4	3 2 0 1	-1479.57	-1503.64	✗	✗	✗	✗
4	3 3 0 3	-1480.14	-1509.76	✗	✗	✗	✗
2	1 3	-1484.38	-1497.34	✓	✓	✓	✓
2	1 1	-1485.33	-1494.58	✓	✓	✓	✓
2	2 1	-1486.41	-1497.51	✓	✓	✓	✓
2	2 2	-1486.93	-1499.89	✓	✓	✓	✓
3	0 0 3	-1487.00	-1501.82	✓	✗	✗	✗
3	0 3 0	-1487.00	-1501.82	✓	✗	✗	✗
3	3 0 0	-1487.00	-1501.82	✓	✗	✗	✗
2	3 3	-1487.32	-1503.98	✓	✓	✓	✓
3	0 0 1	-1487.78	-1498.88	✓	✗	✗	✗
3	0 1 0	-1487.78	-1498.88	✓	✗	✗	✗
3	1 0 0	-1487.78	-1498.88	✓	✗	✗	✗
3	0 0 2	-1488.22	-1501.18	✓	✗	✗	✗
3	0 2 0	-1488.22	-1501.18	✓	✗	✗	✗
3	2 0 0	-1488.22	-1501.18	✓	✗	✗	✗
2	3 1	-1488.32	-1501.28	✓	✓	✓	✓
4	0 0 3 0	-1488.64	-1507.15	✗	✗	✗	✗
2	3 2	-1488.85	-1503.66	✓	✓	✓	✓
4	0 0 0 3	-1489.00	-1507.51	✗	✗	✗	✗
4	0 3 0 0	-1489.00	-1507.51	✗	✗	✗	✗
4	3 0 0 0	-1489.00	-1507.51	✗	✗	✗	✗

4	0 1 0 0	-1489.23	-1504.04	✗	✗	✗	✗
4	1 0 0 0	-1489.23	-1504.04	✗	✗	✗	✗
4	0 0 1 0	-1489.26	-1504.07	✗	✗	✗	✗
4	0 0 0 2	-1489.64	-1506.30	✗	✗	✗	✗
4	0 2 0 0	-1489.64	-1506.30	✗	✗	✗	✗
4	2 0 0 0	-1489.64	-1506.30	✗	✗	✗	✗
4	0 0 2 0	-1489.88	-1506.54	✗	✗	✗	✗
4	3 0 0 1	-1492.71	-1513.08	✗	✗	✗	✗
2	0 3	-1515.89	-1526.99	✓	✓	✗	✗
2	0 1	-1516.00	-1523.41	✓	✓	✗	✗
2	0 2	-1517.49	-1526.75	✓	✓	✗	✗
4	0 0 1 1	-1523.01	-1539.67	✗	✗	✗	✗
4	0 0 1 2	-1525.13	-1543.64	✗	✗	✗	✗
3	1 0 1	-1564.93	-1577.89	✗	✗	✗	✗
2	2 0	-1570.36	-1579.62	✓	✓	✗	✗
2	3 0	-1571.43	-1582.54	✓	✓	✗	✗
2	1 0	-1574.33	-1581.74	✓	✓	✗	✗
1	2	-1624.86	-1630.42	✓	✓	✓	✓
1	3	-1625.60	-1633.00	✓	✓	✓	✓
1	1	-1625.88	-1629.59	✓	✓	✓	✓
4	0 0 0 0	-1745.58	-1758.54	✗	✗	✗	✗
3	0 0 0	-1748.57	-1757.83	✓	✓	✗	✗
2	0 0	-1765.53	-1771.08	✓	✓	✓	✓
1	0	-1867.81	-1869.66	✓	✓	✓	✓

Cohort 3

Number of groups	Model	BIC[2] (n = 60)	BIC[3] (n = 1,956)	Contains at least X students per group, where X =			
				5	*10*	*15*	*20*
4	3 1 3 1	-1030.88	-1057.01	✓	✓	✗	✗
4	1 3 0 1	-1031.20	-1052.11	✓	✗	✗	✗
4	3 1 1 0	-1031.20	-1052.11	✓	✗	✗	✗
4	1 3 0 3	-1031.68	-1056.07	✓	✗	✗	✗
4	3 1 0 3	-1031.68	-1056.07	✓	✗	✗	✗
4	3 1 3 0	-1031.68	-1056.07	✓	✗	✗	✗
4	1 0 3 2	-1032.45	-1055.10	✓	✗	✗	✗
4	1 3 0 2	-1032.45	-1055.10	✓	✗	✗	✗
4	3 1 0 2	-1032.45	-1055.10	✓	✗	✗	✗
4	1 2 1 0	-1032.69	-1051.85	✓	✗	✗	✗
4	2 3 3 1	-1032.85	-1060.72	✓	✓	✗	✗
4	2 3 0 1	-1033.25	-1055.90	✓	✗	✗	✗
4	1 2 3 0	-1033.65	-1056.30	✓	✗	✗	✗
4	2 1 3 0	-1033.65	-1056.30	✓	✗	✗	✗
4	2 3 0 3	-1033.73	-1059.86	✓	✗	✗	✗
4	1 1 1 0	-1033.82	-1051.24	✓	✗	✗	✗
4	2 3 1 1	-1033.98	-1058.37	✓	✗	✗	✗
4	1 2 0 2	-1033.99	-1054.90	✓	✗	✗	✗
4	1 2 2 0	-1033.99	-1054.90	✓	✗	✗	✗
4	2 3 0 2	-1034.50	-1058.89	✓	✗	✗	✗
4	1 1 2 0	-1034.70	-1053.86	✓	✗	✗	✗
4	3 3 1 3	-1034.87	-1064.48	✓	✗	✗	✗
4	3 3 3 1	-1034.87	-1064.48	✓	✓	✗	✗
4	3 3 0 1	-1035.23	-1059.62	✓	✗	✗	✗
4	3 3 1 0	-1035.23	-1059.62	✓	✗	✗	✗

3	1 3 1	-1035.31	-1052.73	✓	✓	✓	✗
4	2 2 3 0	-1035.69	-1060.08	✓	✗	✗	✗
4	3 3 0 3	-1035.71	-1063.59	✓	✗	✗	✗
4	3 3 3 0	-1035.71	-1063.59	✓	✗	✗	✗
4	2 3 1 2	-1035.98	-1062.11	✓	✗	✗	✗
4	2 2 0 2	-1036.02	-1058.67	✓	✗	✗	✗
4	2 2 2 0	-1036.02	-1058.67	✓	✗	✗	✗
4	1 1 3 1	-1036.04	-1058.69	✓	✗	✗	✗
4	3 1 1 2	-1036.15	-1060.54	✗	✗	✗	✗
3	1 3 2	-1036.38	-1055.54	✓	✓	✓	✗
4	2 2 1 1	-1036.49	-1059.14	✓	✗	✗	✗
4	1 3 3 1	-1036.50	-1062.63	✗	✗	✗	✗
3	1 3 3	-1036.57	-1057.48	✓	✓	✓	✗
4	3 3 0 2	-1036.60	-1062.74	✓	✗	✗	✗
4	3 3 2 0	-1036.60	-1062.74	✓	✗	✗	✗
4	1 1 2 1	-1036.61	-1057.51	✓	✗	✗	✗
4	2 1 2 0	-1036.74	-1057.65	✓	✗	✗	✗
4	3 3 2 1	-1036.80	-1064.68	✓	✓	✗	✗
3	1 2 1	-1037.20	-1052.88	✓	✓	✓	✗
3	2 3 1	-1037.26	-1056.42	✓	✓	✓	✗
3	1 1 1	-1037.67	-1051.61	✓	✓	✓	✗
4	2 2 2 1	-1037.78	-1062.17	✓	✗	✗	✗
3	1 2 2	-1038.01	-1055.43	✓	✓	✓	✗
4	1 3 1 3	-1038.47	-1064.60	✗	✗	✗	✗
3	2 3 3	-1038.52	-1061.17	✓	✓	✓	✗
4	3 3 3 3	-1038.85	-1071.95	✓	✓	✗	✗
4	1 2 3 1	-1039.05	-1063.44	✗	✗	✗	✗
3	2 2 1	-1039.15	-1056.57	✓	✓	✓	✗
3	2 1 1	-1039.43	-1055.11	✓	✓	✓	✗
4	1 3 2 2	-1039.79	-1065.92	✗	✗	✗	✗

3	2 2 2	-1039.96	-1059.12	✓	✓	✓	✗
4	1 3 2 0	-1040.47	-1063.12	✗	✗	✗	✗
4	1 1 3 3	-1040.64	-1066.78	✗	✗	✗	✗
4	1 3 3 0	-1040.66	-1065.05	✗	✗	✗	✗
3	3 3 1	-1040.71	-1061.61	✓	✓	✗	✗
4	0 1 2 0	-1041.13	-1058.55	✓	✗	✗	✗
4	1 0 2 0	-1041.13	-1058.55	✓	✗	✗	✗
4	0 0 3 1	-1041.15	-1060.32	✓	✗	✗	✗
4	0 3 0 1	-1041.15	-1060.32	✓	✗	✗	✗
4	0 3 1 0	-1041.15	-1060.32	✓	✗	✗	✗
4	3 0 1 0	-1041.15	-1060.32	✓	✗	✗	✗
3	3 2 1	-1041.16	-1060.32	✓	✓	✓	✗
4	0 1 3 0	-1041.17	-1060.34	✓	✗	✗	✗
4	1 0 3 0	-1041.17	-1060.34	✓	✗	✗	✗
4	1 1 3 2	-1041.21	-1065.60	✗	✗	✗	✗
4	0 0 2 1	-1041.22	-1058.64	✓	✗	✗	✗
4	0 2 0 1	-1041.22	-1058.64	✓	✗	✗	✗
4	0 2 1 0	-1041.22	-1058.64	✓	✗	✗	✗
4	0 1 0 1	-1041.37	-1057.05	✓	✗	✗	✗
4	0 1 1 0	-1041.37	-1057.05	✓	✗	✗	✗
4	1 0 1 0	-1041.37	-1057.05	✓	✗	✗	✗
4	0 2 3 2	-1041.59	-1065.98	✓	✗	✗	✗
3	3 3 2	-1041.61	-1064.25	✓	✓	✗	✗
4	2 0 3 0	-1041.92	-1062.83	✓	✗	✗	✗
3	3 1 3	-1042.01	-1062.91	✓	✓	✗	✗
4	0 0 3 3	-1042.06	-1064.71	✓	✗	✗	✗
4	0 3 0 3	-1042.06	-1064.71	✓	✗	✗	✗
4	0 3 3 0	-1042.06	-1064.71	✓	✗	✗	✗
4	3 3 0 0	-1042.06	-1064.71	✓	✗	✗	✗
4	0 2 0 2	-1042.08	-1061.24	✓	✗	✗	✗

4	0 2 2 0	-1042.08	-1061.24	✓	✗	✗	✗
4	2 0 2 0	-1042.08	-1061.24	✓	✗	✗	✗
3	3 1 2	-1042.10	-1061.27	✓	✓	✗	✗
3	3 3 3	-1042.27	-1066.66	✓	✓	✗	✗
3	3 2 2	-1042.30	-1063.21	✓	✓	✗	✗
4	0 0 3 2	-1042.32	-1063.23	✓	✗	✗	✗
4	0 3 0 2	-1042.32	-1063.23	✓	✗	✗	✗
4	0 3 2 0	-1042.32	-1063.23	✓	✗	✗	✗
4	3 0 0 2	-1042.32	-1063.23	✓	✗	✗	✗
4	3 0 2 0	-1042.32	-1063.23	✓	✗	✗	✗
3	3 2 3	-1042.45	-1065.10	✓	✓	✗	✗
4	0 3 1 1	-1042.81	-1063.72	✓	✗	✗	✗
3	0 1 3	-1043.70	-1059.38	✓	✗	✗	✗
3	1 0 3	-1043.70	-1059.38	✓	✗	✗	✗
4	3 0 2 1	-1043.95	-1066.60	✓	✗	✗	✗
3	0 2 1	-1044.05	-1057.98	✓	✗	✗	✗
4	2 0 2 1	-1044.12	-1065.03	✗	✗	✗	✗
3	0 2 2	-1044.29	-1059.97	✓	✗	✗	✗
3	2 0 2	-1044.29	-1059.97	✓	✓	✗	✗
4	2 3 3 2	-1044.48	-1074.10	✗	✗	✗	✗
3	0 3 1	-1044.67	-1060.35	✓	✓	✗	✗
4	3 0 3 3	-1045.02	-1072.90	✓	✗	✗	✗
4	3 3 2 2	-1045.21	-1074.83	✗	✗	✗	✗
3	0 3 2	-1045.27	-1062.69	✓	✓	✗	✗
3	3 0 2	-1045.27	-1062.69	✓	✓	✗	✗
4	0 3 1 3	-1045.54	-1069.93	✓	✗	✗	✗
4	0 3 3 3	-1045.56	-1073.43	✓	✓	✗	✗
3	0 3 3	-1045.83	-1064.99	✓	✓	✗	✗
3	3 0 3	-1045.83	-1064.99	✓	✓	✗	✗
4	0 3 1 2	-1046.00	-1068.64	✗	✗	✗	✗

4	3 0 1 2	-1046.00	-1068.64	✗	✗	✗	✗
4	0 3 3 1	-1046.06	-1070.46	✗	✗	✗	✗
4	0 2 2 2	-1046.15	-1068.80	✗	✗	✗	✗
4	2 0 2 2	-1046.15	-1068.80	✗	✗	✗	✗
4	2 0 3 2	-1047.04	-1071.43	✗	✗	✗	✗
4	3 0 3 2	-1048.10	-1074.23	✗	✗	✗	✗
4	3 1 0 1	-1048.19	-1069.09	✗	✗	✗	✗
3	2 2 0	-1048.69	-1064.37	✓	✗	✗	✗
4	2 2 0 1	-1048.83	-1069.74	✗	✗	✗	✗
3	2 3 0	-1048.84	-1066.26	✓	✗	✗	✗
3	1 0 1	-1049.36	-1061.55	✓	✗	✗	✗
3	1 1 0	-1049.36	-1061.55	✓	✗	✗	✗
2	1 2	-1049.64	-1060.09	✓	✓	✓	✓
3	2 0 1	-1049.69	-1063.62	✓	✗	✗	✗
3	2 1 0	-1049.69	-1063.62	✓	✗	✗	✗
3	3 1 1	-1050.35	-1067.77	✗	✗	✗	✗
2	1 3	-1050.39	-1062.58	✓	✓	✓	✓
2	2 2	-1050.55	-1062.74	✓	✓	✓	✓
3	3 2 0	-1050.74	-1068.16	✓	✗	✗	✗
3	3 3 0	-1050.88	-1070.05	✓	✗	✗	✗
2	2 3	-1051.13	-1065.07	✓	✓	✓	✓
2	1 1	-1051.24	-1059.95	✓	✓	✓	✓
2	2 1	-1051.27	-1061.72	✓	✓	✓	✓
3	2 3 2	-1051.55	-1072.46	✗	✗	✗	✗
3	3 0 1	-1051.73	-1067.41	✓	✗	✗	✗
4	3 1 0 0	-1052.21	-1071.37	✗	✗	✗	✗
2	3 2	-1052.57	-1066.51	✓	✓	✓	✓
2	3 3	-1053.17	-1068.84	✓	✓	✓	✓
2	3 1	-1053.31	-1065.51	✓	✓	✓	✓
3	2 1 2	-1053.41	-1070.83	✓	✗	✗	✗

3	1 2 0	-1054.85	-1068.79	✓	✓	✗	✗
3	1 3 0	-1055.72	-1071.40	✓	✓	✗	✗
4	1 2 0 0	-1056.37	-1073.79	✓	✗	✗	✗
4	1 3 0 0	-1057.04	-1076.21	✓	✗	✗	✗
3	3 1 0	-1059.04	-1074.72	✓	✓	✗	✗
4	2 3 0 0	-1059.08	-1079.98	✓	✗	✗	✗
3	0 0 1	-1064.47	-1074.92	✓	✓	✗	✗
3	0 1 0	-1064.47	-1074.92	✓	✓	✗	✗
3	0 0 2	-1064.52	-1076.72	✓	✓	✗	✗
3	0 2 0	-1064.52	-1076.72	✓	✓	✗	✗
3	2 0 0	-1064.52	-1076.72	✓	✓	✗	✗
3	0 3 0	-1065.35	-1079.28	✓	✓	✗	✗
4	0 0 0 2	-1067.40	-1083.08	✓	✗	✗	✗
4	0 0 2 0	-1067.40	-1083.08	✓	✗	✗	✗
4	0 2 0 0	-1067.40	-1083.08	✓	✗	✗	✗
4	2 0 0 0	-1067.40	-1083.08	✓	✗	✗	✗
4	0 0 1 0	-1067.58	-1081.52	✓	✗	✗	✗
4	0 1 0 0	-1067.58	-1081.52	✓	✗	✗	✗
4	0 0 3 0	-1068.11	-1085.53	✓	✗	✗	✗
4	0 3 0 0	-1068.11	-1085.53	✓	✗	✗	✗
4	3 0 0 0	-1068.11	-1085.53	✓	✗	✗	✗
2	0 2	-1072.78	-1081.49	✓	✓	✓	✗
2	0 3	-1072.81	-1083.27	✓	✓	✓	✗
2	0 1	-1072.89	-1079.86	✓	✓	✓	✗
3	0 0 3	-1076.91	-1090.84	✗	✗	✗	✗
2	2 0	-1082.39	-1091.10	✓	✓	✓	✗
2	3 0	-1084.00	-1094.45	✓	✓	✓	✗
2	1 0	-1084.52	-1091.48	✓	✓	✓	✗
3	3 0 0	-1087.24	-1101.17	✓	✗	✗	✗
3	1 0 0	-1087.89	-1098.34	✓	✗	✗	✗

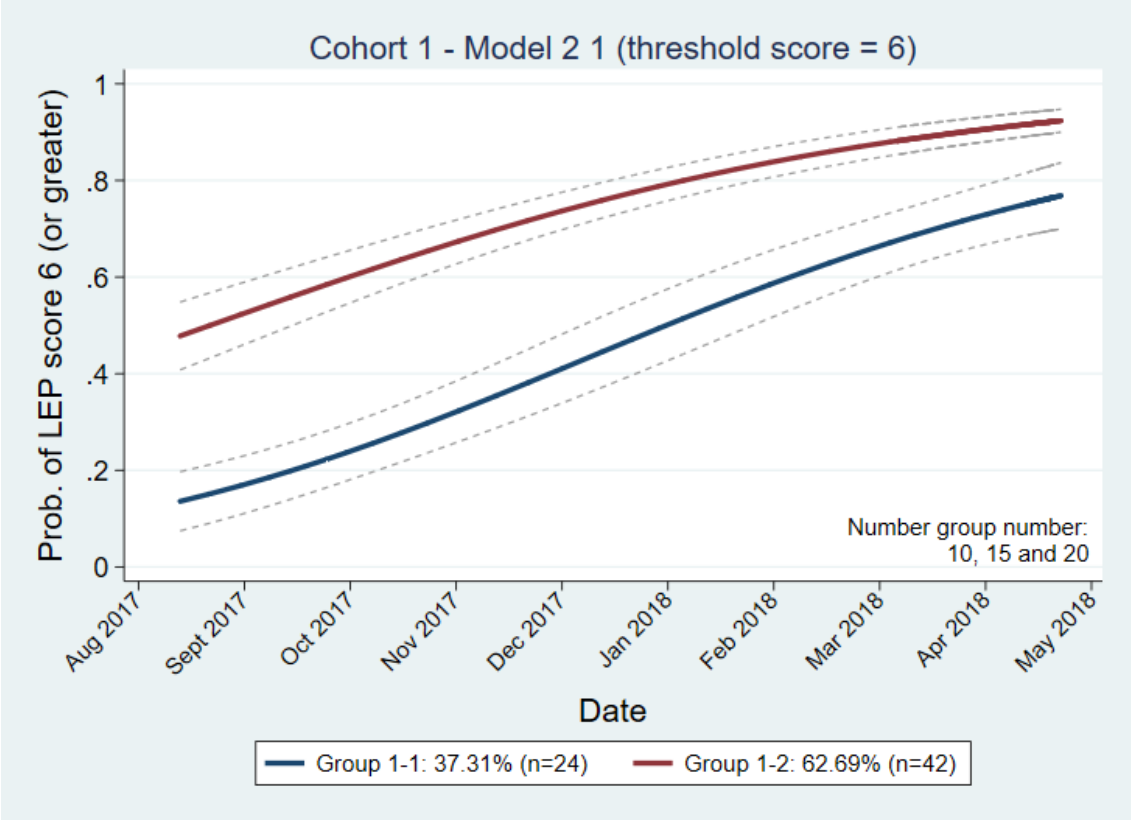
1	2	-1127.48	-1132.71	✓	✓	✓	✓
1	3	-1127.57	-1134.53	✓	✓	✓	✓
1	1	-1129.65	-1133.14	✓	✓	✓	✓
3	0 0 0	-1162.13	-1170.85	✓	✓	✗	✗
2	0 0	-1165.74	-1170.97	✓	✓	✓	✓
4	0 0 0 0	-1165.89	-1178.08	✓	✗	✗	✗
1	0	-1232.74	-1234.49	✓	✓	✓	✓

iii. LEP - alternative models

Threshold score = 6

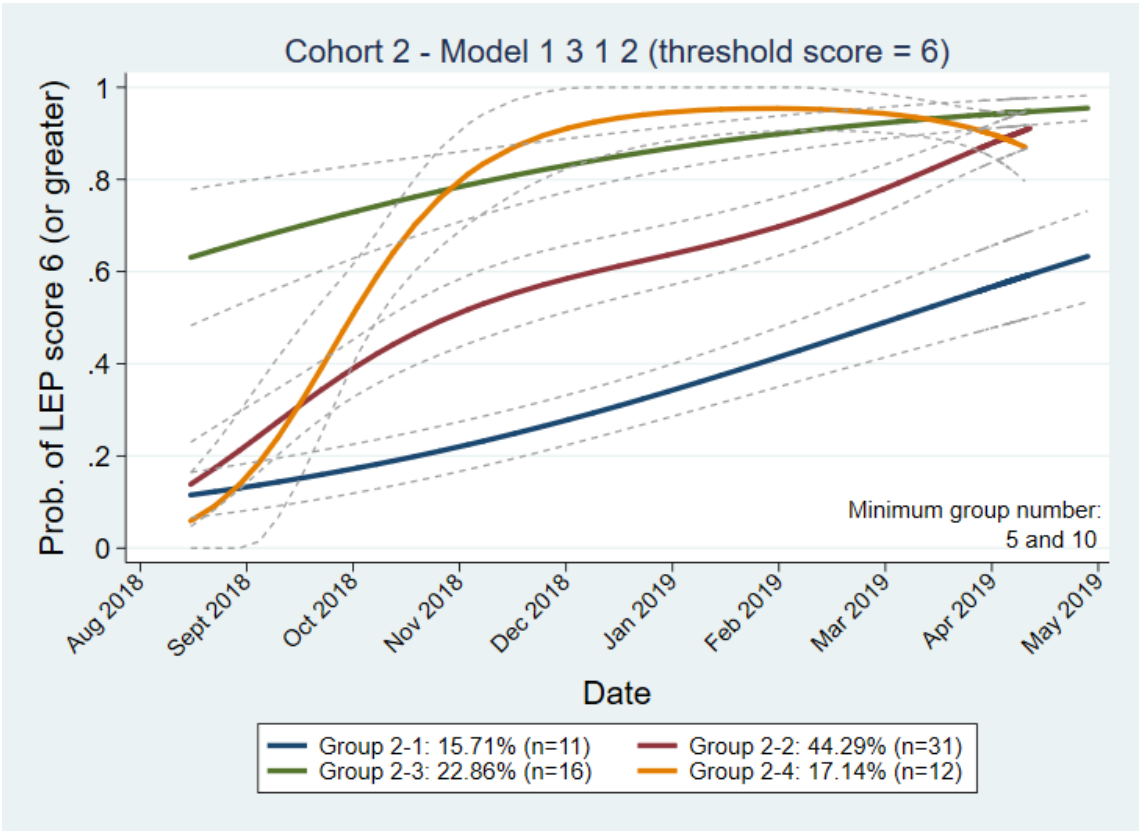
Cohort 1

Model 2 1

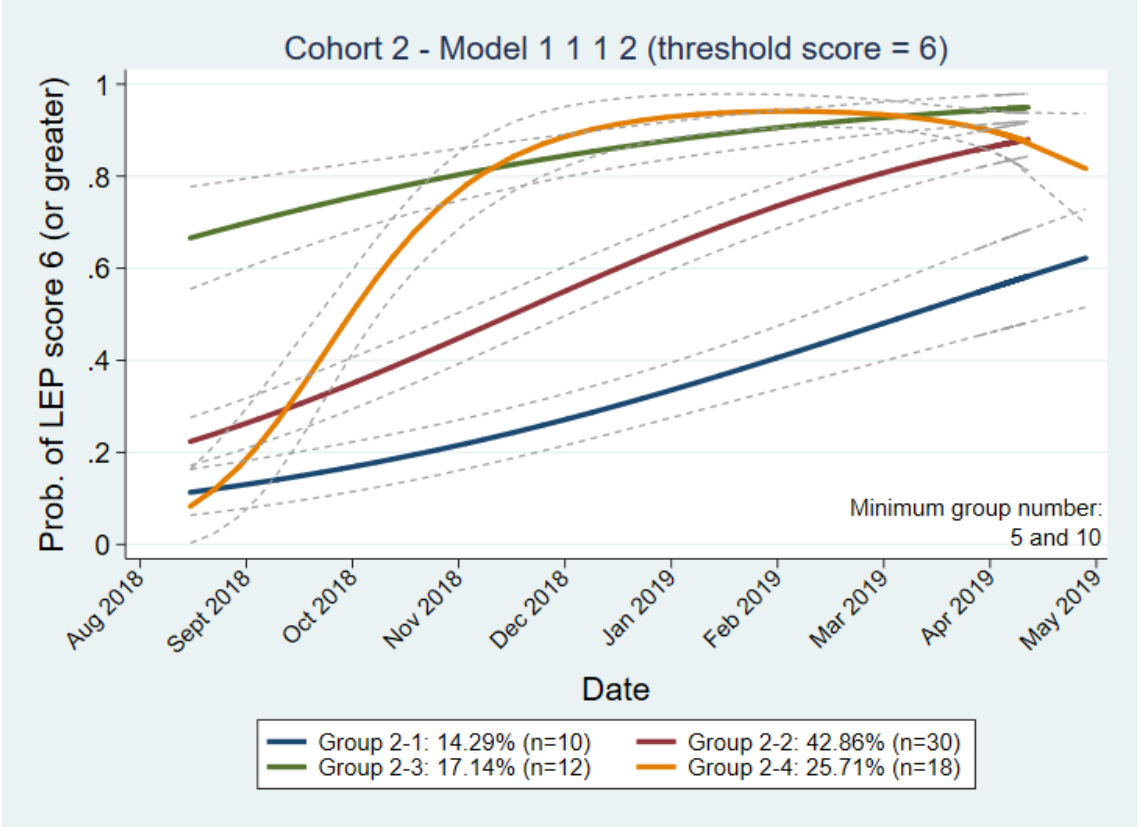


Cohort 2

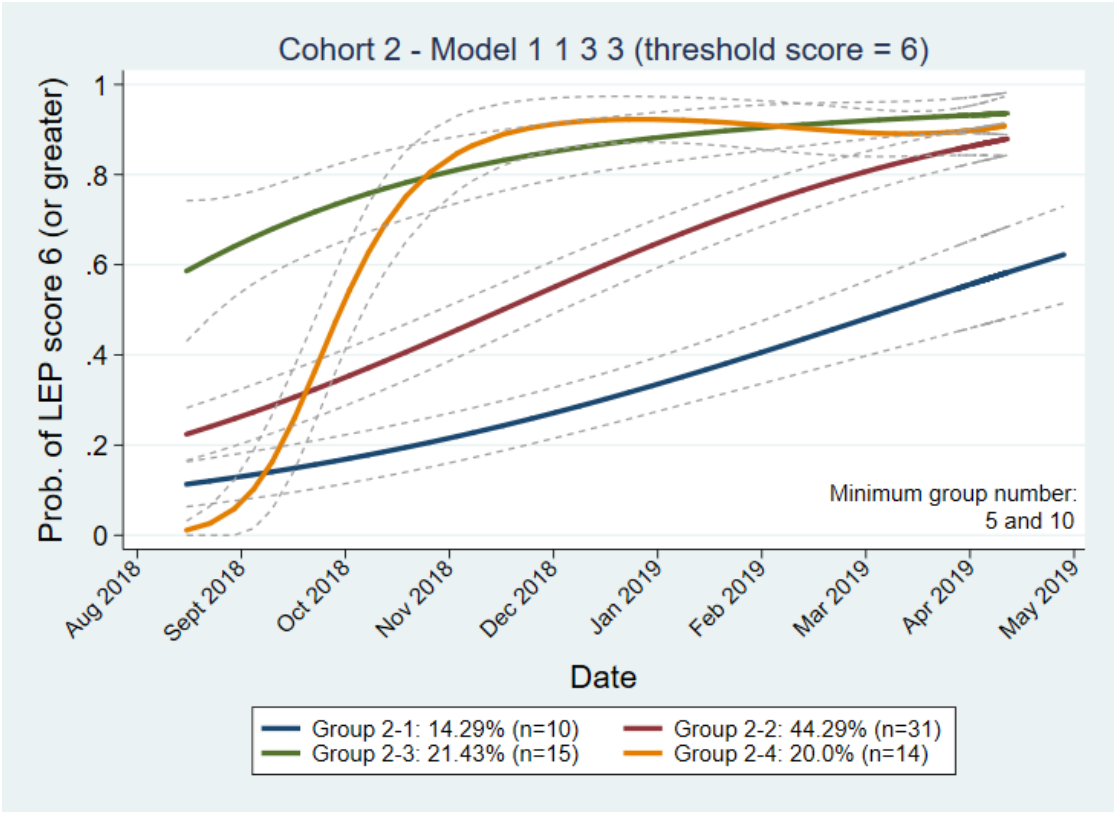
Model 1 3 1 2



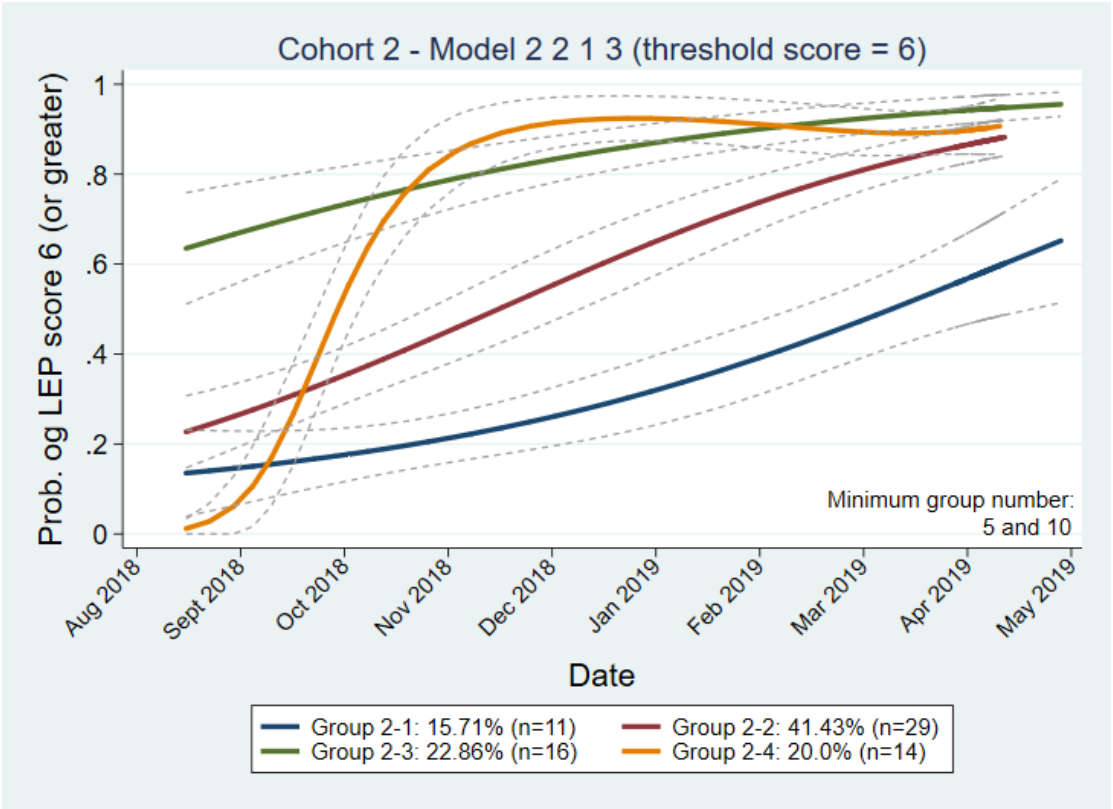
Model 1 1 1 2



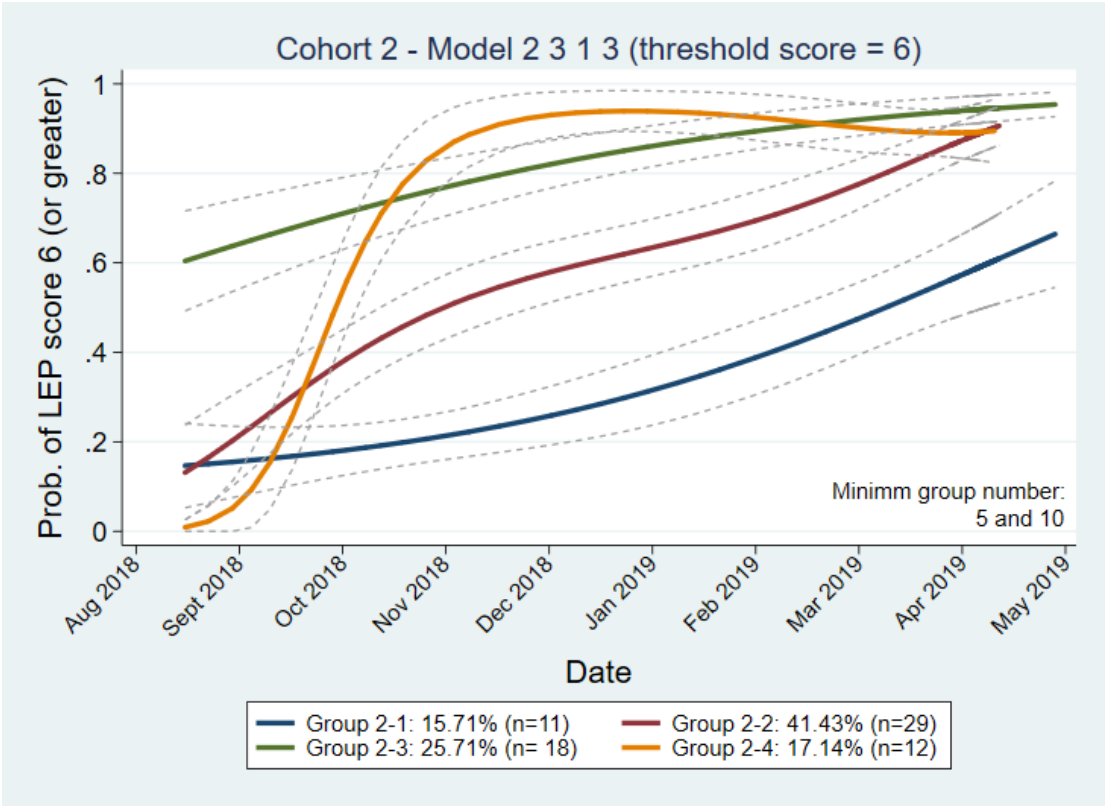
Model 1 1 3 3



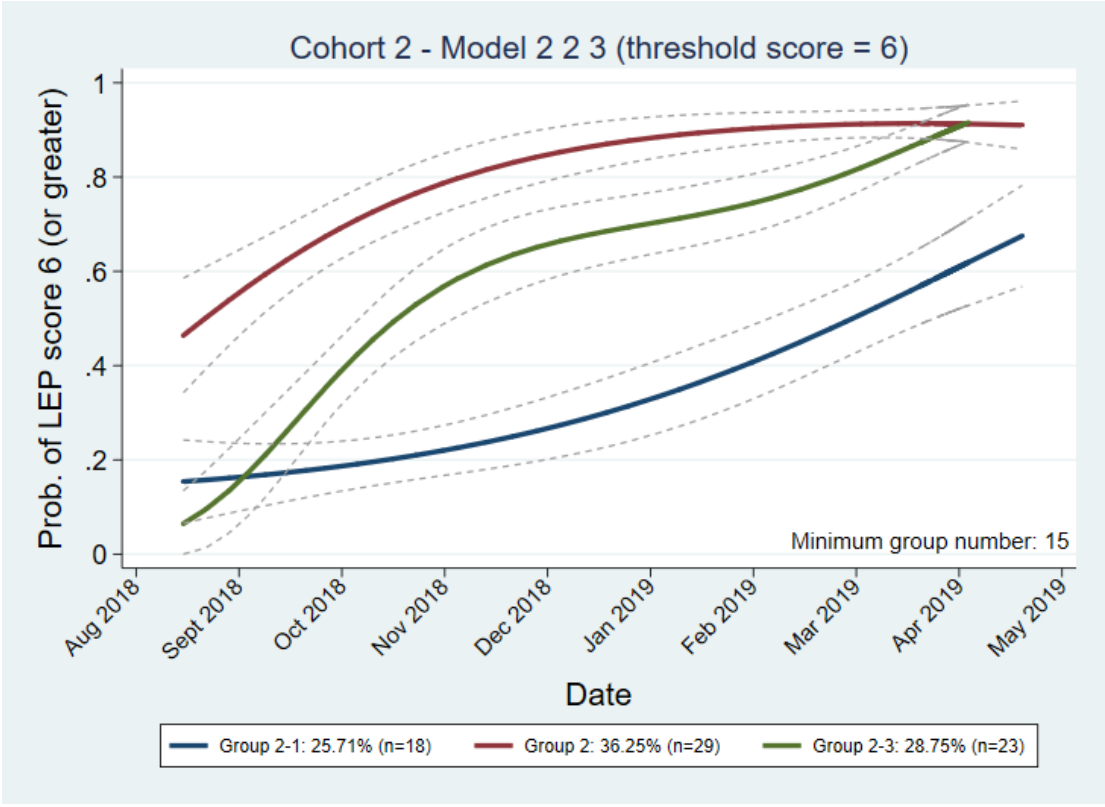
Model 2 2 1 3



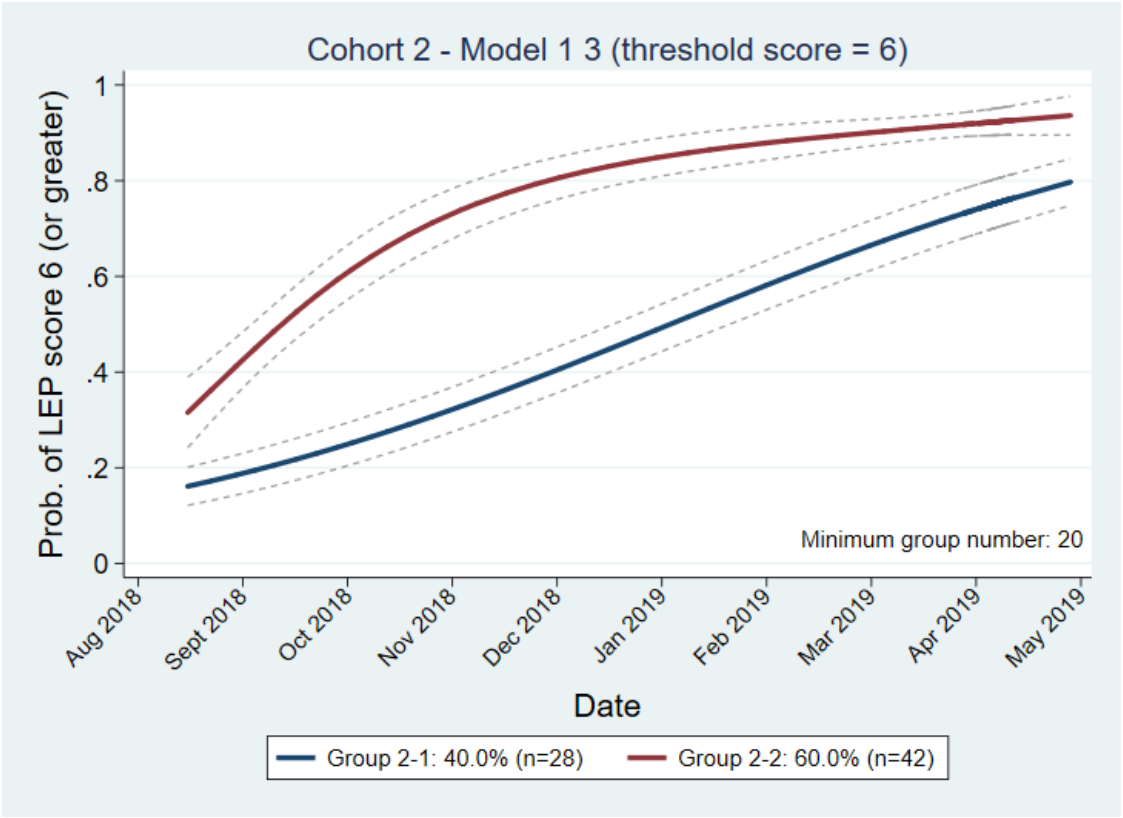
Model 2 3 1 3



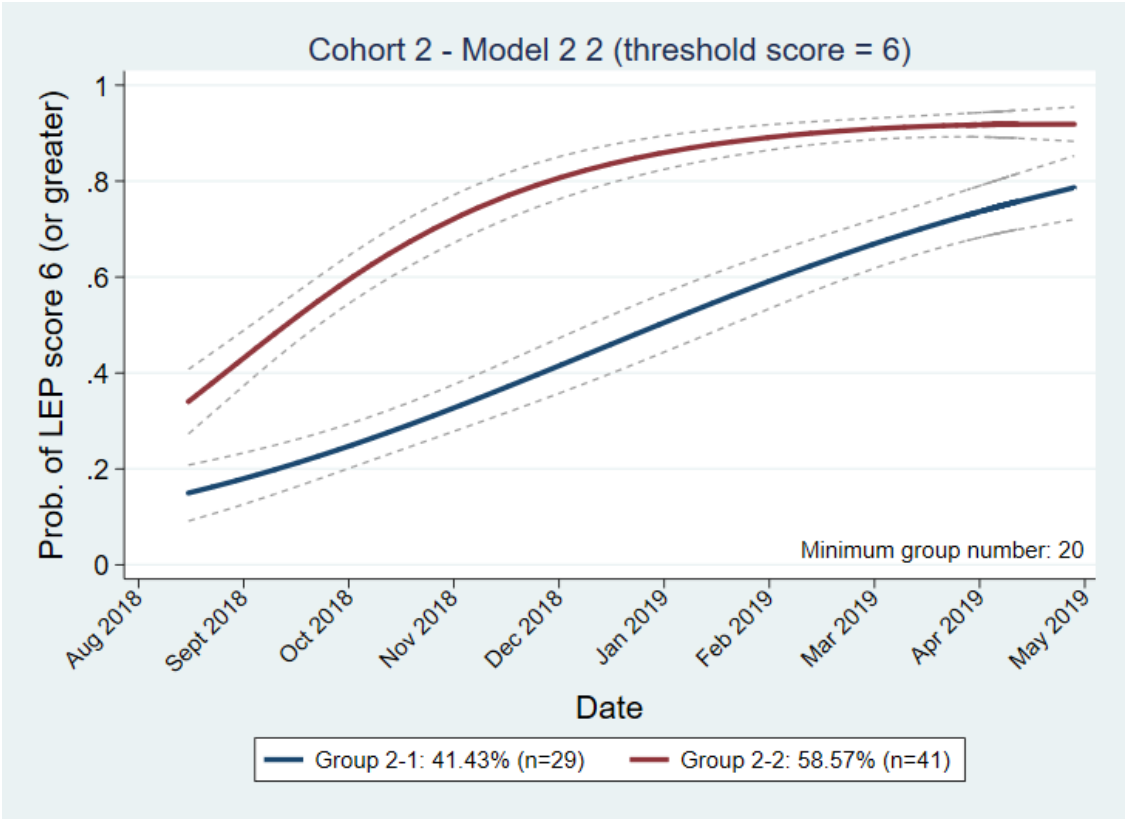
Model 2 2 3



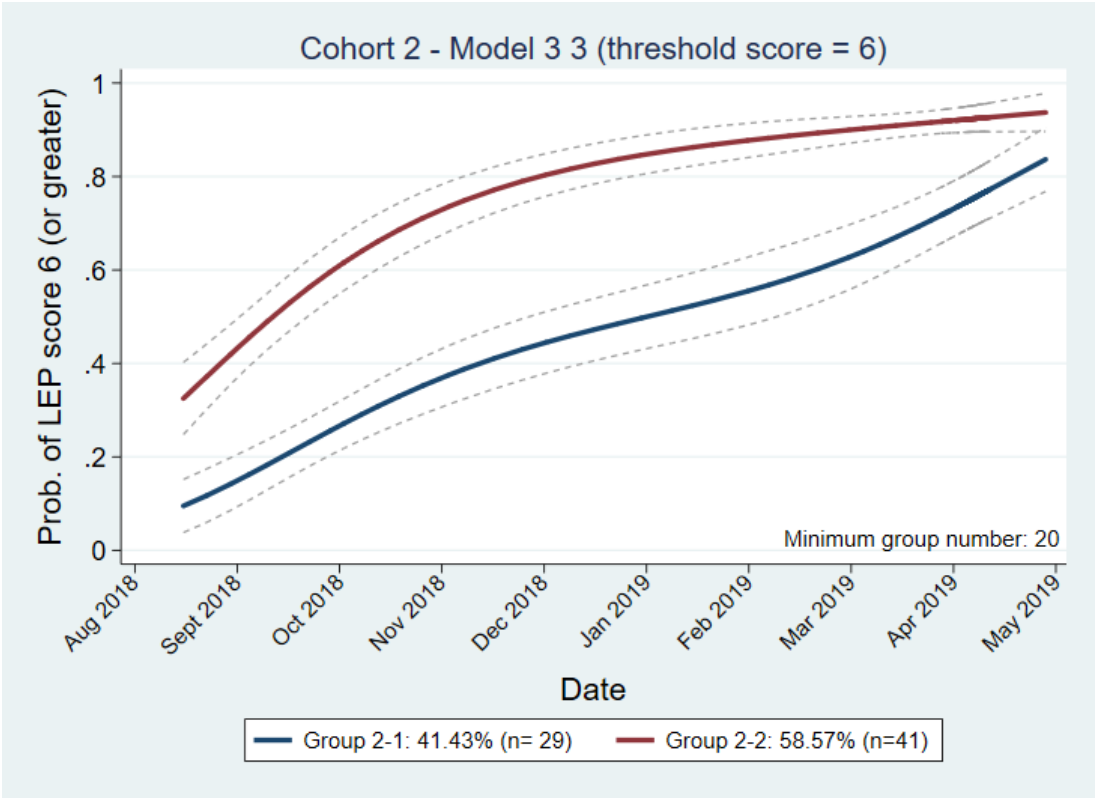
Model 1 3



Model 2 2



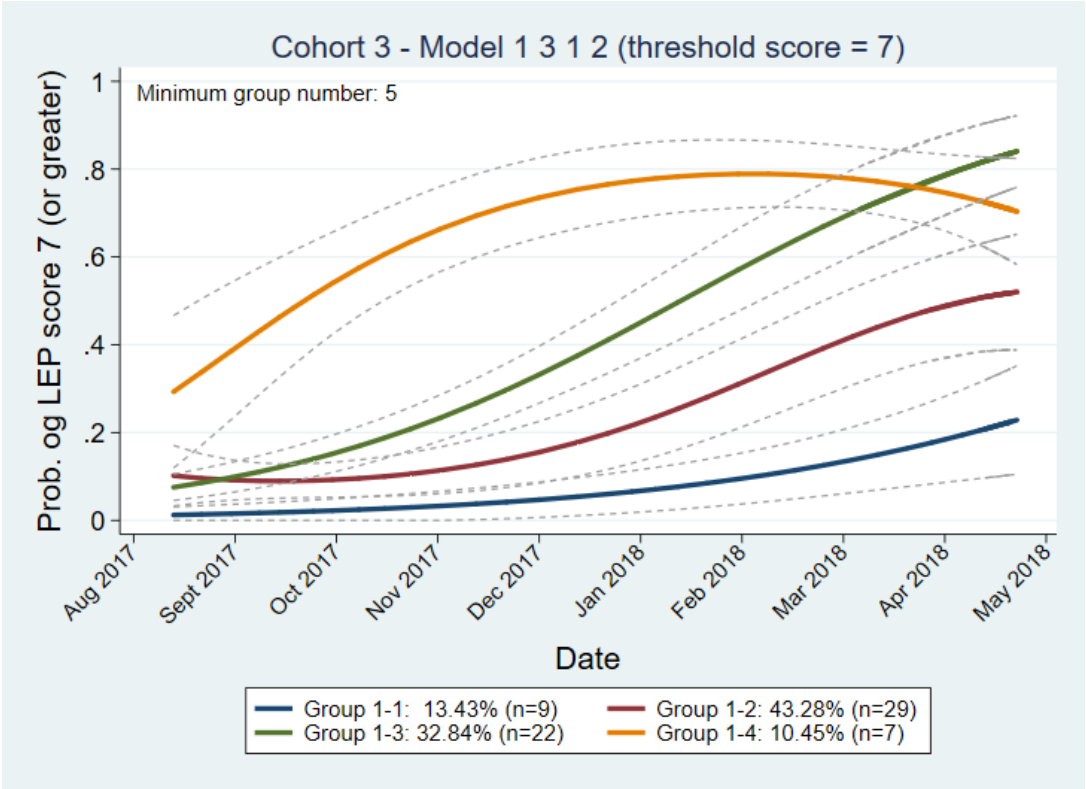
Model 3 3



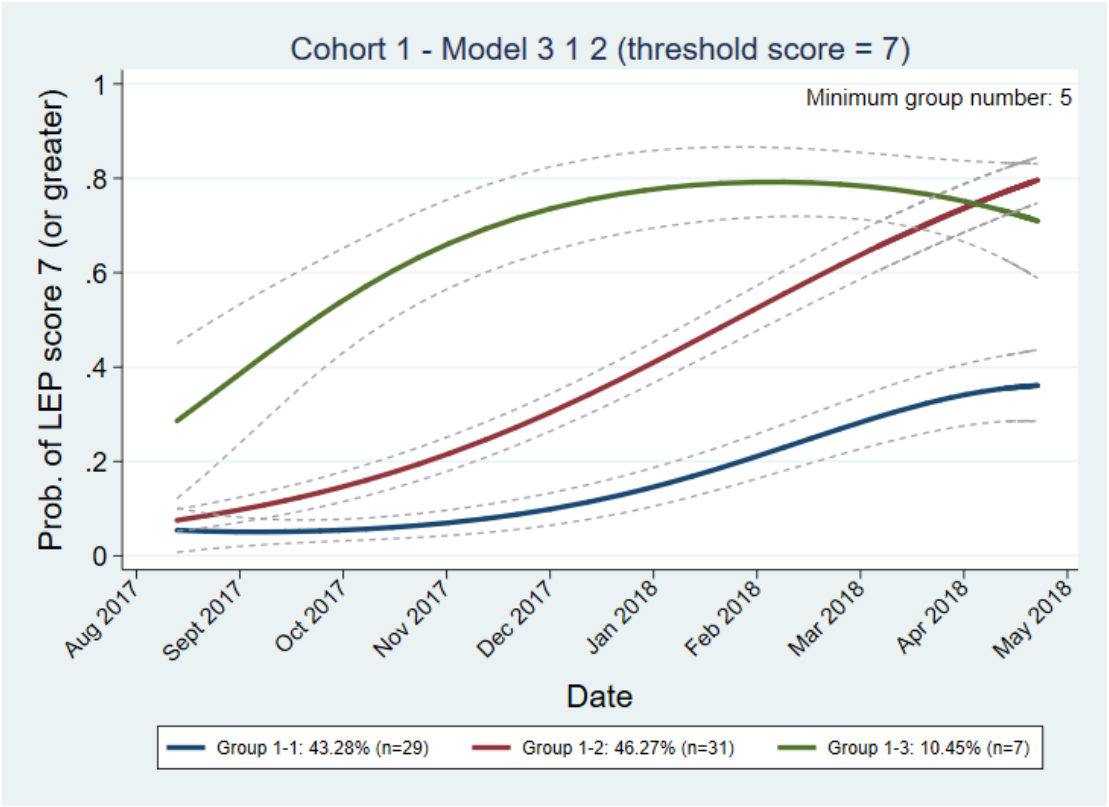
Threshold = 7

Cohort 1

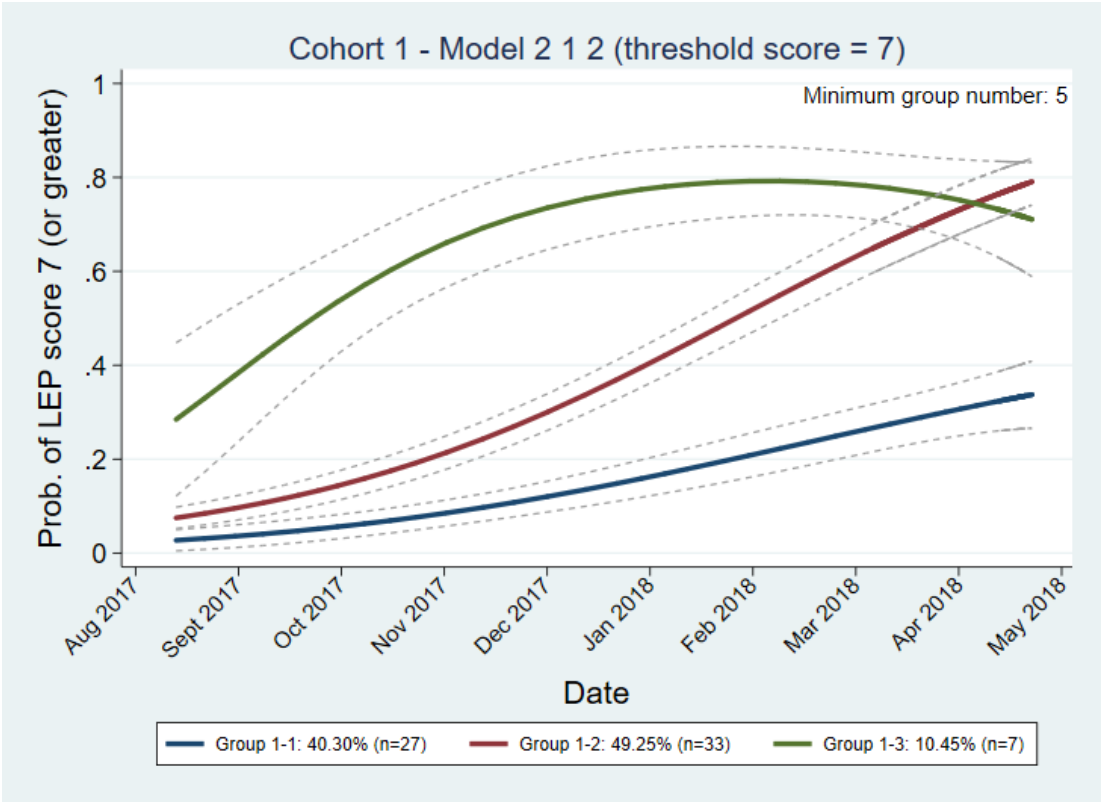
Model 1 3 1 2



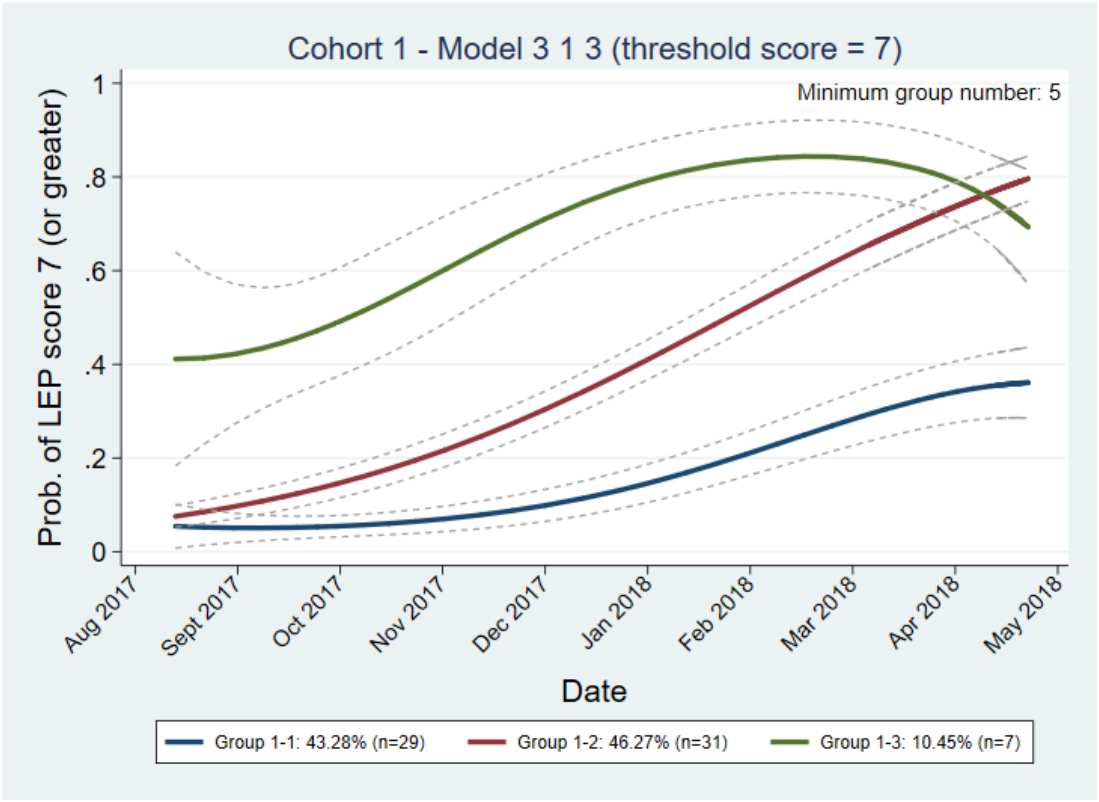
Model 3 1 2



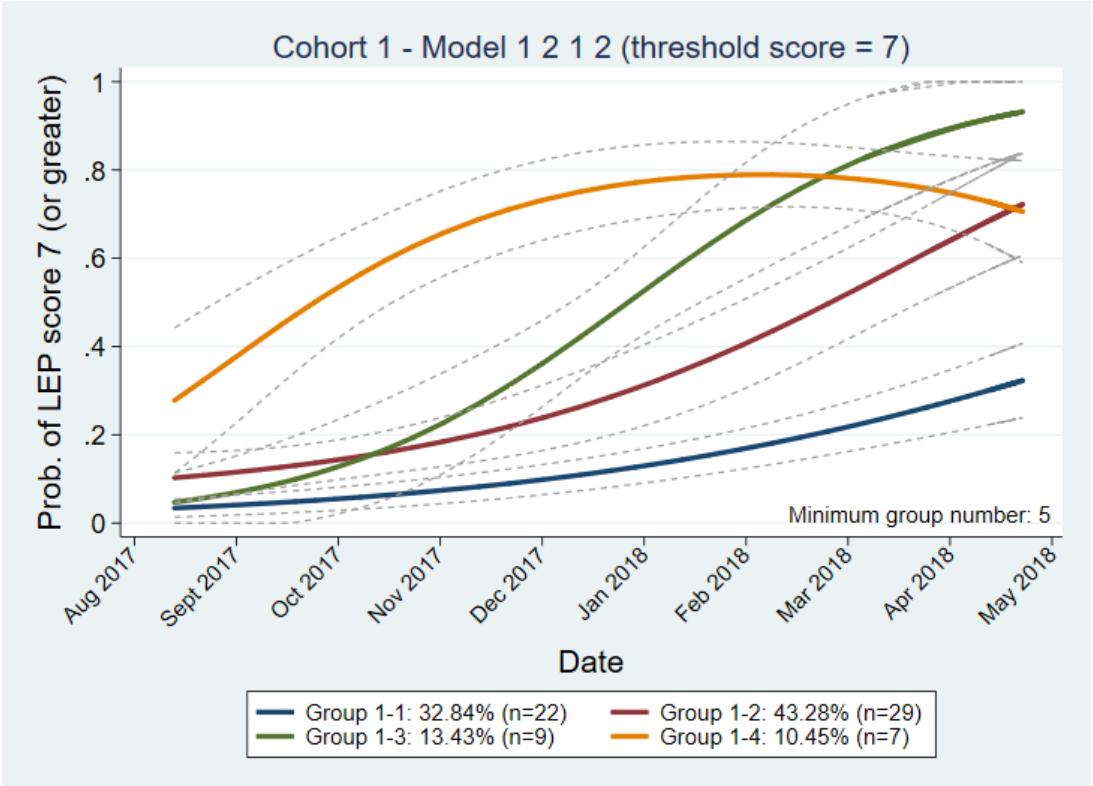
Model 2 1 2



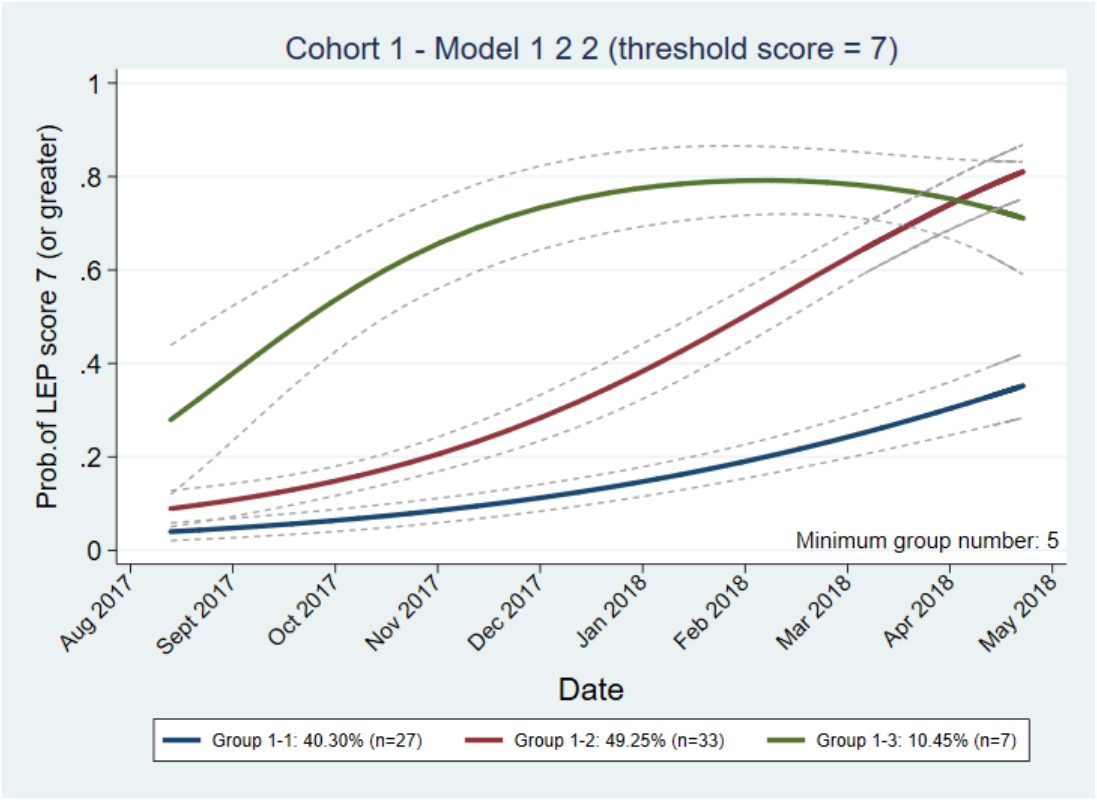
Model 3 1 3



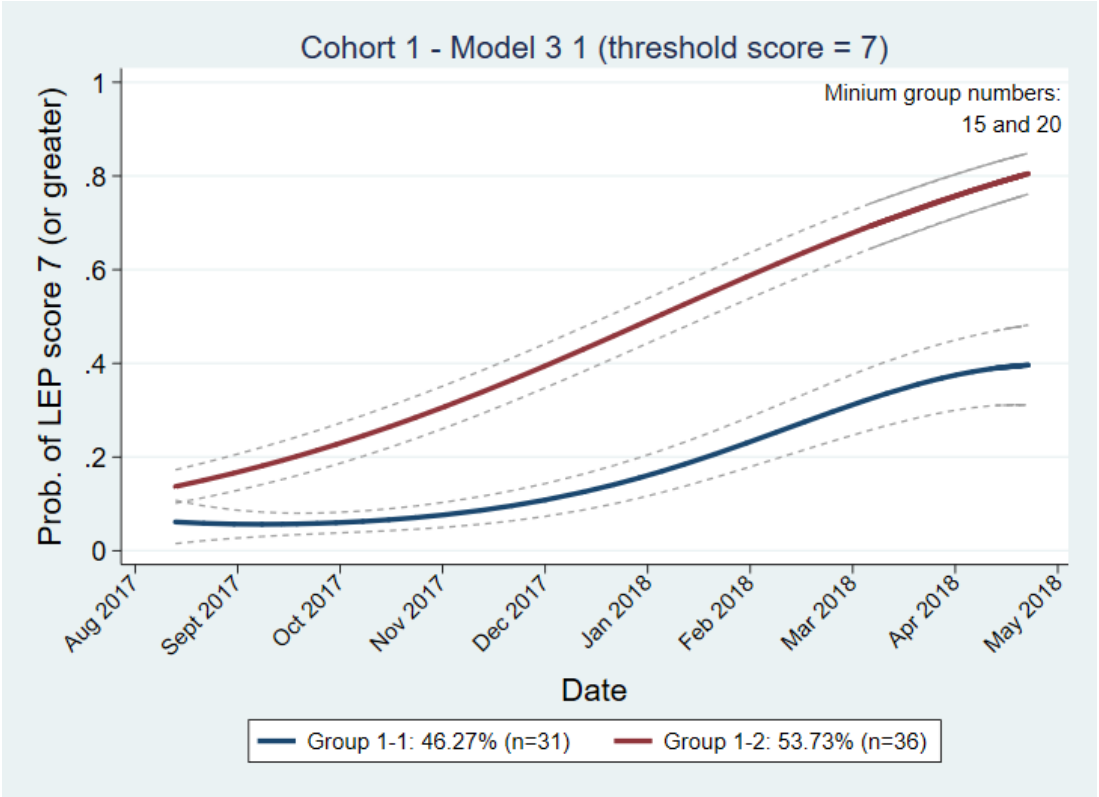
Model 1 2 1 2



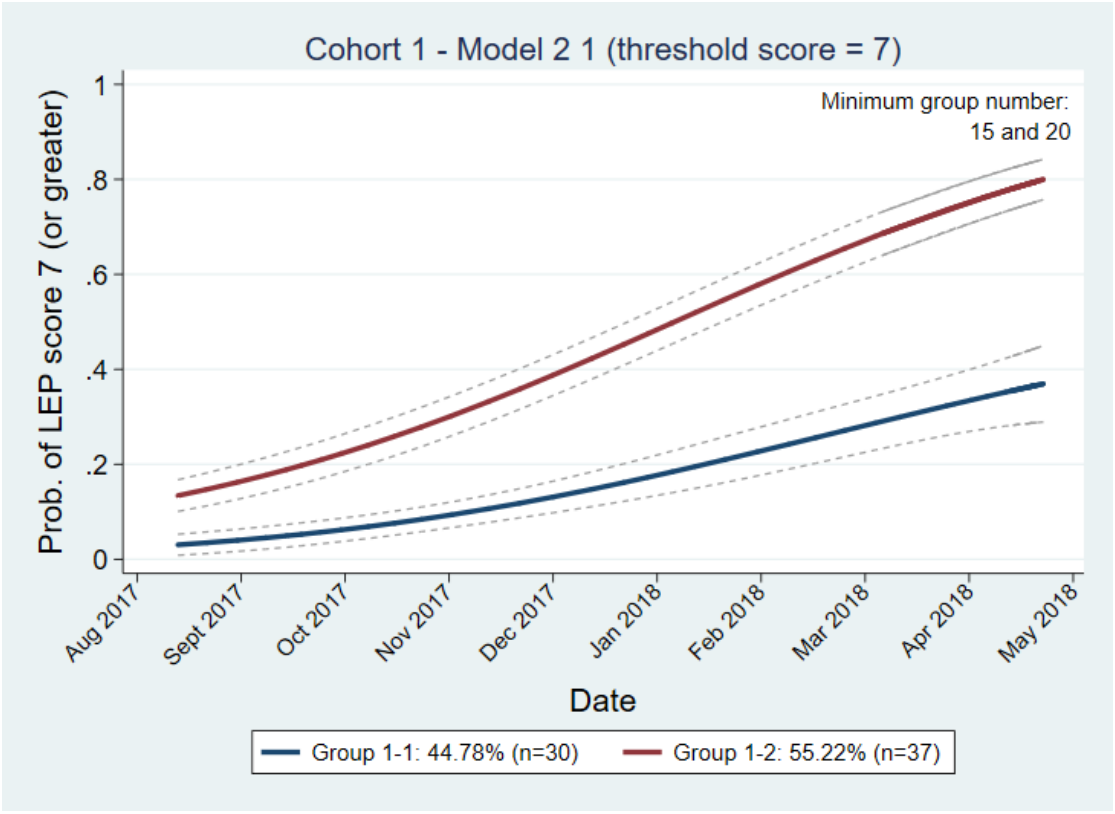
Model 1 2 2



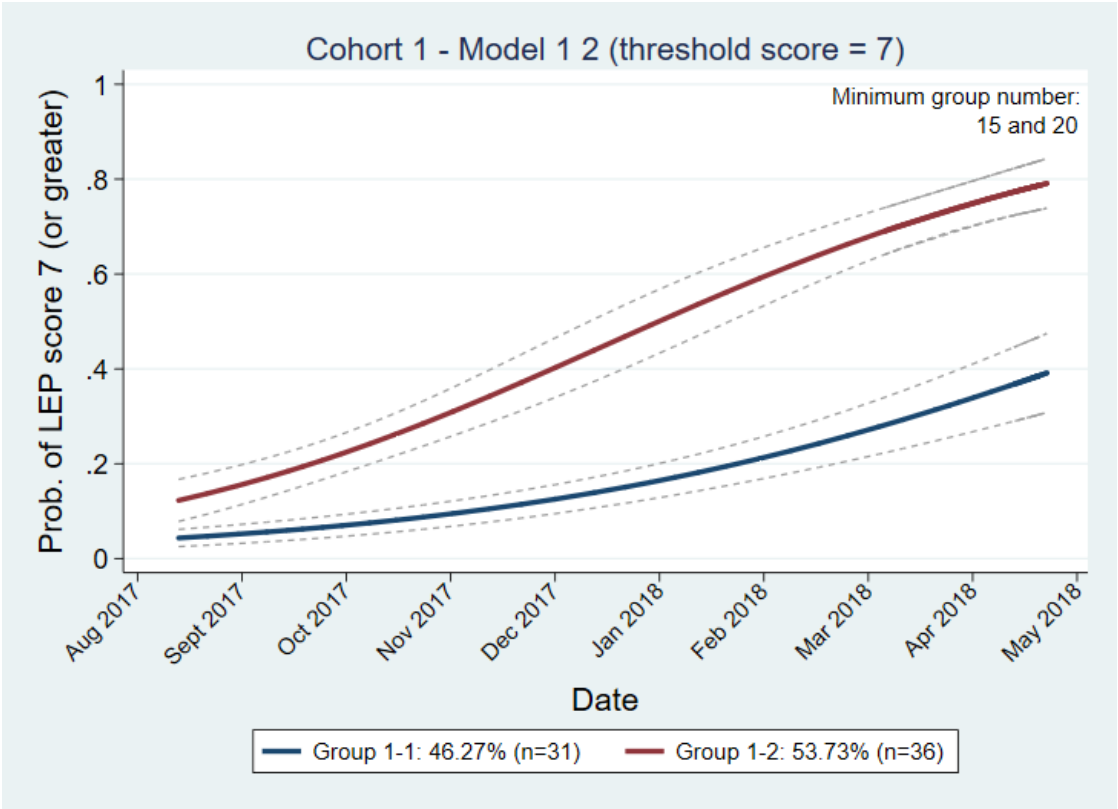
Model 3 1



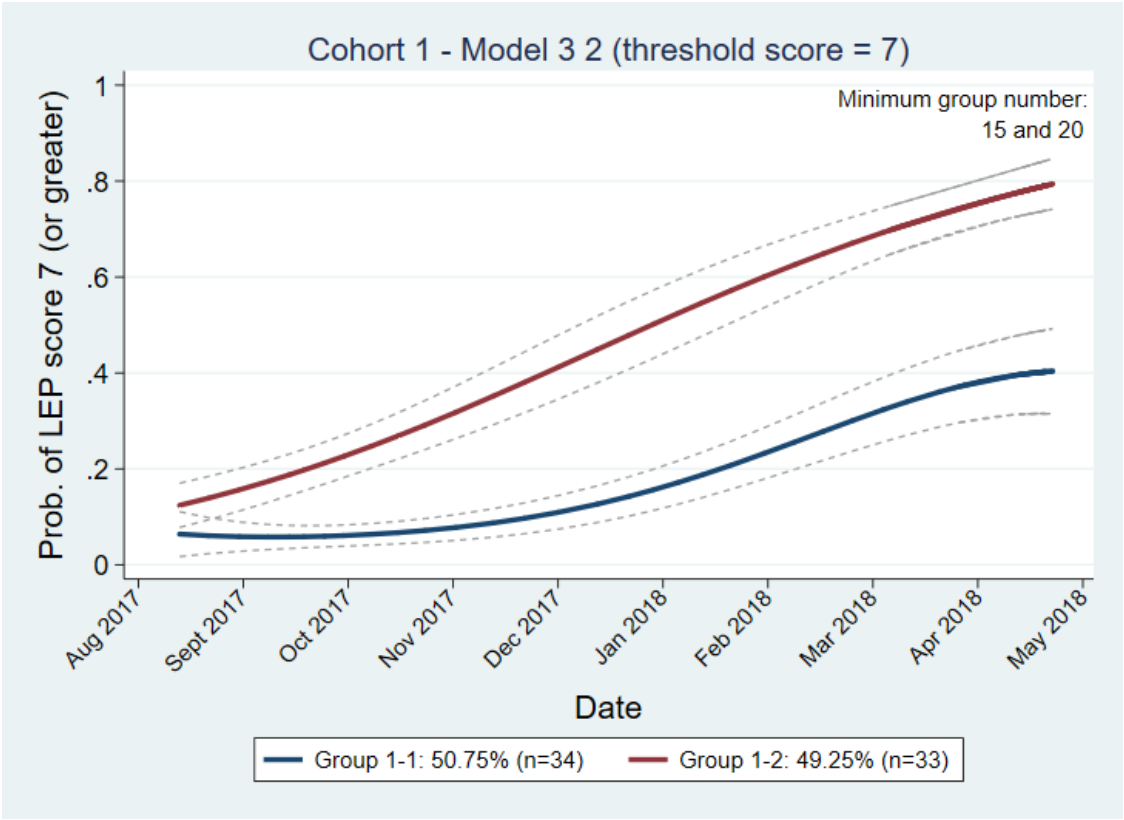
Model 2 1



Model 1 2

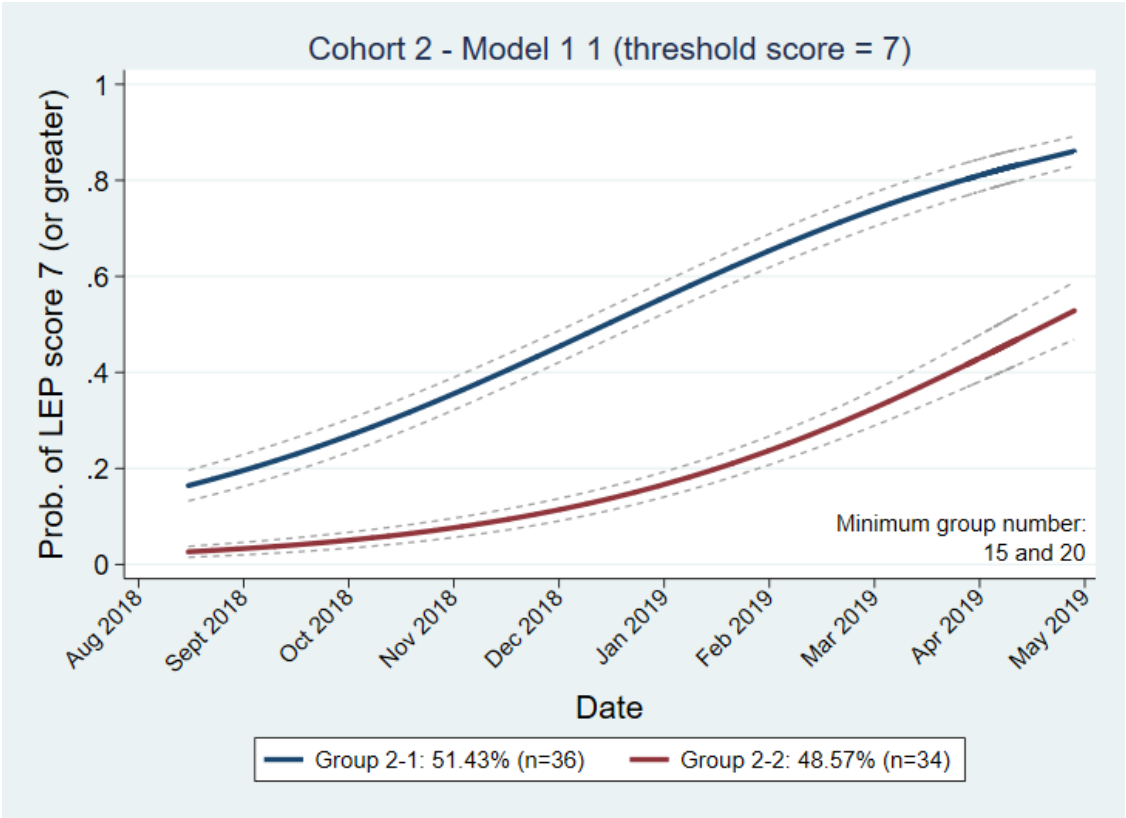


Model 3 2



Cohort 2

Model 1 1



Appendix 11 – Additional cross tabulations between LIFTUPP© and longitudinal evaluations of performance group-based trajectory model memberships

Cohort 1 - LIFTUPP© threshold performance indicator = 5; LEP threshold score = 6

Cross tabulations between the trajectory group memberships for LIFTUPP© model 3 2 and longitudinal evaluation of performance (LEP) model 3 3 3 2. NOTE: The best performing groups for both LIFTUPP© and LEP models are marked with an *.

Cohort 1						
Minimum number participants per group: 5		**LEP model 3 3 3 2 group (threshold score = 6)**				**Fisher's exact (p)**
		1	***2***	***3****	***4***	
LIFTUPP© model 3 2 Group (threshold performance indicator = 5)	***1**** (n = 52)	n = 8 15.38%	n = 22 42.31%	n = 18 34.62%	n = 4 7.69%	0.61
	2 (n = 15)	n = 2 13.33%	n = 6 40.00%	n = 4 26.67%	n = 3 20.00%	

Cross tabulations between the trajectory group memberships for LIFTUPP© model 3 2 and longitudinal evaluation of performance (LEP) model 1 1. NOTE: The best performing groups for both LIFTUPP© and LEP models are marked with an *.

Cohort 1				
Minimum number participants per group: 10 and 15		**LEP model 1 1 group (threshold score = 6)**		**Fisher's exact (p)**
		1	***2****	
LIFTUPP© model 3 2 Group (threshold performance indicator = 5)	***1**** (n = 52)	n = 17 32.69%	n = 35 67.31%	0.37
	2 (n = 15)	n = 7 46.67%	n = 8 53.33%	

Cohort 2 - LIFTUPP© threshold performance indicator = 5; LEP threshold score = 6

Cross tabulations between the trajectory group memberships for LIFTUPP© model 1 3 and longitudinal evaluation of performance (LEP) model 2 3 1 3. NOTE: The best performing groups for both LIFTUPP© and LEP models are marked with an *.

Cohort 2						
Minimum group numbers: 5 and 10		**LEP model 2 3 1 3 (threshold score = 6)**				**Fisher's exact (p)**
		1	***2***	***3****	***4***	
LIFTUPP© model 1 3 group (threshold performance indicator = 5)	**1*** (n = 18)	n = 2 11.11%	n = 7 38.89%	n = 7 38.89%	n = 2 11.11%	0.56
	2 (n = 52)	n = 9 17.31%	n = 22 42.31%	n = 11 21.15%	n = 10 19.23%	

Cohort 3 - LIFTUPP© threshold performance indicator = 5; LEP threshold score = 6

Cross tabulations between the trajectory group memberships for LIFTUPP© model 3 2 and longitudinal evaluation of performance (LEP) model 3 0 3. NOTE: The best performing groups for both LIFTUPP© and LEP models are marked with an *.

Cohort 3					
Minimum group number: 5		**LEP model 3 0 3 (threshold score = 6)**			**Fisher's exact (p)**
		1	***2***	***3****	
LIFTUPP© model 3 2 group (threshold performance indicator = 5)	***1**** (n = 37)	n = 14 37.84%	n = 5 13.51%	n = 18 48.65%	0.30
	2 (n = 23)	n = 13 56.52%	n = 1 4.35%	n = 9 39.13%	

Cohort 1 - LIFTUPP© threshold performance indicator = 5; LEP threshold score = 7

Cross tabulations between the trajectory group memberships for LIFTUPP© model 3 2 and longitudinal evaluation of performance (LEP) model 1 1 3. NOTE: The best performing groups for both LIFTUPP© and LEP models are marked with an *.

Cohort 1					
Minimum group numbers: 5 and 10		**LEP model 1 1 3 group (threshold score = 7)**			**Fisher's exact (p)**
		1	***2***	***3****	
LIFTUPP© model 3 2 group (threshold performance indicator = 5)	***1**** (n = 52)	n = 21 40.38%	n = 26 50.00%	n = 5 9.62%	0.76
	2 (n = 15)	n = 7 46.67%	n = 6 40.00%	n = 2 13.33%	

Cross tabulations between the trajectory group memberships for LIFTUPP© model 3 2 and longitudinal evaluation of performance (LEP) model 1 1. NOTE: The best performing groups for both LIFTUPP© and LEP models are marked with an *.

Cohort 1				
Minimum group number: 15		**LEP model 1 1 group (threshold score = 7)**		**Fisher's exact (p)**
		1	***2****	
LIFTUPP© model 3 2 group (threshold performance indicator = 5)	***1**** (n = 52)	n = 23 44.23%	n = 29 55.77%	0.55
	2 (n = 15)	n = 7 46.67%	n = 8 53.33%	

Cohort 2 - LIFTUPP© threshold performance indicator = 5; LEP threshold score = 7

Cross tabulations between the trajectory group memberships for LIFTUPP© model 1 3 and longitudinal evaluation of performance (LEP) model 1 1 3 2. NOTE: The best performing groups for both LIFTUPP© and LEP models are marked with an *.

Cohort 2						
Minimum group numbers: 5 and 10		**LEP model 1 1 3 2 (threshold score = 7)**				**Fisher's exact (p)**
		1	***2***	***3****	***4****	
LIFTUPP© model 1 3 group (threshold performance indicator = 5)	***1**** (n = 18)	n = 3 16.67%	n = 4 22.22%	n = 4 22.22%	n = 7 38.89%	0.39
	2 (n = 52)	n = 7 13.46%	n = 22 42.31%	n = 6 11.54%	n = 17 32.69%	

Cohort 3 - LIFTUPP© threshold performance indicator = 5; LEP threshold score = 7

Cross tabulations between the trajectory group memberships for LIFTUPP© models 3 2 and longitudinal evaluation of performance (LEP) models 1 3 0 1. NOTE: The best performing groups for both LIFTUPP© and LEP models are marked with an *.

Cohort 3						
Minimum group number: 5		**LEP model 1 3 0 1 (threshold score = 7)**				**Fisher's exact (p)**
		1	***2***	***3***	***4****	
LIFTUPP© model 3 2 group (threshold performance indicator = 5)	***1**** (n = 37)	n = 12 32.42%	n = 9 24.32%	n = 5 13.51%	n = 11 29.73%	0.69
	2 (n = 23)	n = 7 30.43%	n = 9 39.13%	n = 2 8.70%	n = 5 21.74%	

Cross tabulations between the trajectory group memberships for LIFTUPP© model 3 2 and longitudinal evaluation of performance (LEP) model 3 1 3 1. NOTE: The best performing groups for both LIFTUPP© and LEP models are marked with an *.

Cohort 3						
Minimum group number: 10		LEP model 3 1 3 1 (threshold score = 7)				Fisher's exact (p)
		1	***2***	***3****	***4***	
LIFTUPP© model 3 2 group (threshold performance indicator = 5)	***1**** (n = 37)	n = 9 24.32%	n = 13 35.14%	n = 8 21.62%	n = 7 18.92%	0.71
	2 (n = 23)	n = 9 39.13%	n = 7 30.43%	n = 4 17.39%	n = 3 13.04%	

Cross tabulations between the trajectory group memberships for LIFTUPP© model 3 2 and longitudinal evaluation of performance (LEP) model 1 3 1. NOTE: The best performing groups for both LIFTUPP© and LEP models are marked with an *.

Cohort 3					
Minimum group number: 15		LEP model 1 3 1 (threshold score = 7)			Fisher's exact (p)
		1	***2***	***3****	
LIFTUPP© model 3 2 group (threshold performance indicator = 5)	***1**** (n = 37)	n = 13 35.14%	n = 9 24.32%	n = 15 40.54%	0.50
	2 (n = 23)	n = 7 30.43%	n = 9 39.13%	n = 7 30.43%	