Using Quantitative Methods to Analyse Educational Interventions in Undergraduate Physics Teaching

Catriona McAllister

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

School of Physics and Astronomy
College of Science and Engineering
University of Glasgow

July 2022

Abstract

The physicist's approach to understanding the universe is to strive to uncover the basic rules that govern the behaviour of the natural world around us. Physics provides a way to create a model of the world that is testable through quantitative measurements and observation. Using this philosophy as a basis, this manuscript presents a quantitative model of 5 aspects of a particular educational setting. Firstly, a framework to understand how to quantify the efficacy of an assessment is presented using Bloom's Taxonomy and the Force Concept Inventory. The aim of this framework was to identify if there is a key skill or metric to success in physics. The results of this project identified the impact of not only students' conceptual understanding and mathematical skill but the format of the assessment itself. Secondly is presented an analysis of the intrinsic, internal factors which impact student performance. Through this analysis it was shown that students' prior knowledge and metacognitive skill have the most significant impacts on their performance. A two part analysis of group dynamics from a purely quantitative perspective was then undertaken, building on the results from the first two projects. This project posited a quantitative way to define the dynamic within a group so as to show the relationship with student performance. While there is a high level of complexity within group dynamics which can obscure the effect of any one variable, the impact of cognitive diversity can still be seen in overall assessment results. The final project, building on the work of Carl Wieman, analyses the impact of a new style of laboratory teaching on student performance in overall assessment. A significant impact was found in the final assessment results for questions relating to the laboratory material. This is an exciting result as the impact of laboratory teaching has been historically inconclusive. These projects created a quantitative lens through which to understand physics education research, especially for specific research areas which are predominantly considered in a qualitative way. These projects demonstrate a way that physics education research can be integrated into any undergraduate physics course, by applying physics methodologies to educational data available to all physics educators.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank all of my friends and family for their support throughout my PhD experience, especially throughout the pandemic. Thank you to my parents, the only other people who will ever read this whole thing! Thank you to Dr Eric Yao, my supervisor, who has provided support and guidance since my undergraduate days, even if it was mostly over zoom at this point. Finally, thank you to the University of Glasgow who provided the funding for my PhD.

Declaration

University of Glasgow

College of Science and Engineering

Statement of Originality to Accompany Thesis Submission

Name: Catriona McAllister

Registration Number:

I certify that the thesis presented here for examination for a PhD degree of the University of Glasgow is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it) and that the thesis has not been edited by a third party beyond what is permitted by the University's PGR Code of Practice. The copyright of this thesis rests with the author. No quotation from it is permitted without full acknowledgement.

I declare that the thesis does not include work forming part of a thesis presented successfully for another degree [unless explicitly identified and as noted below].

I declare that this thesis has been produced in accordance with the University of Glasgow's Code of Good Practice in Research.

I acknowledge that if any issues are raised regarding good research practice based on review of the thesis, the examination may be postponed pending the outcome of any investigation of the issues.

Signature: Catriona McAllister.

Date: 15/07/2022.

Chapter 1

Introduction

1.1 Introduction

The physicist's approach is to discover the basic rules that govern the behaviour of the natural world around us. Physics provides a way to create a model of the world that is testable through quantitative measurement and observation. This does not seem like something that can been easily applied to people, to education. Yet, there are many physics education researchers who have created a way to take this physics approach and apply it to education.

The work by Richard Hake [Hake, 2002] showed that quantitative research could be used on a broad scale, utilising pre and post tests to understand the efficacy of different educational approaches. His work was broadly effective and showed the power of quantitative work. Many of the approaches in this manuscript build on that work, looking at the broad trends of student performance and using the vast amounts of data at hand to identify what aspects of the educational set up have an influence.

This kind of work was made possible by the introduction of educational frameworks such as the Force Concept Inventory (FCI) [Hestenes et al., 1992, Halloun and Hestenes, 1985] and Bloom's Taxonomy [Bloom et al., 1956]. These tools provided a new way of approaching education research, especially from a quantitative perspective. They provide a baseline of how to analyse and understand students' conceptual knowledge by creating a standard way to assess physics and define assessment respectively. This is fundamental to a broad, consistent and collaborative education research community. Throughout this manuscript the FCI and Bloom's Taxonomy are used as the basis of analysis, especially when considering assessment.

Physics education has also tackled the more human side of education. The work of Eric Mazur has covered many aspects of physics education research from peer instruction [Crouch and Mazur, 2001] to the gender gap in performance [Lorenzo et al., 2006]. His work on peer instruction has fundamentally impacted the way that physics is taught, especially within higher education. The setting for the research within this manuscript uses peer instruction throughout and is a core part of its success. Even with this far reaching impact, Mazur has also highlighted

the gap between physics education research and the teaching of physics as recently as 2014 [Fraser et al., 2014]. There is a growing physics education research community within the UK but engagement needs to reach across the physics teaching community, from primary to tertiary education. The work presented in this manuscript uses data and tools accessible to any physics educator, with these tools deriving from analyses broadly used in physics research. The current approach to education has vastly increased the available data without an explicit research aim. The use of virtual learning environments and a focus on consistent, continuous assessment have widened and deepened the pool of data that can feed into education research without disruption to students or additional effort on the part of educators. This manuscript shows the potential research that could be done in any higher education establishment; how physics education research can be built into the everyday teaching of physics in higher education.

Physics education can also focus on more discreet ideas. Individual interventions that tackle specific issues or introduce a new approach to teaching a specific aspect of physics. Much of the work of the Nobel Prize winner Carl Wieman has focused on how to best develop, utilise and assess physics laboratory teaching. While there have been many advances in physics education, the basic, underlying structure can take much more time to change. The way we teach physics practically did not radically changed throughout the 20th century but we are now seeing new and exciting approaches. The final project in this manuscript is based and developed from the ideas of Carl Wieman with regards to practical physics teaching.

This kind of structural development is also being seen in the rise of the flipped classroom model [Gilboy et al., 2015, Graham et al., 2017, Street et al., 2015, Blair et al., 2016]. This is still a debated idea. Completely changing the emphasis and structure of a lecture, within flipped classroom students are expected to complete pre-reading and perhaps a pre-test and the lecture time is used for discussion and the development of more complex ideas. This is a significant change from the traditional didactic lectures used throughout higher education. This model is building on the work of education researchers like Eric Mazur [Mazur, 1997] but development is still ongoing as to how to understand the impact. The flipped classroom model is part of the backdrop of the research presented in this manuscript. Within that context is shown a potential way of analysing the impact of flipped classroom, especially where that intersects with peer and group dynamics. Especially when analysing the impact of pedagogical shifts such as flipped classroom, it is important to highlight the variables that change inherently with that shift. With a move to flipped classroom there will be a necessary shift to incorporate aspects of peer learning and changes in contact time with subject experts. Considerations must always be made with regards to the intersectional nature of educational pedagogy.

Given this intersectionality, there are a wide variety of approaches that can be taken to education research. At the extremes there are qualitative processes that consider the in depth engagement of a small group of students or quantitative studies that cover 1000 student cohorts. Both provide very different but complementary approaches to education research, providing a

more holistic view of the field. This manuscript presents the results of 5 physics education research projects developed and implemented in a practically idealised educational setting.

1.2 Context and Setting

All experiments in education are influenced by the context and setting of the interventions. The setting for all of the interventions and analyses outlined in this manuscript is the International Physics Summer School at the University of Glasgow (The Summer School). This summer school is uniquely placed as a close to ideal context for education research.

1.2.1 Structure

The structure of the Summer School is unique. It is an 8 week, intensive, flipped classroom course. The material builds throughout the 8 weeks, covering the majority of material from a traditional level 1 and level 2 physics course. The students are in session from 10am until 5pm, Monday to Thursday, predominantly working in groups, and with a sole focus on physics. Much more material can be covered in a short period of time without sacrificing depth of understanding as the use of a flipped classroom model allows for the contact time between staff and students to be used most effectively. Additionally, the fast pace can help students see the connections between concepts as they simply see the concepts over a short period of time, keeping those ideas from previous topics in mind more easily. Complete immersion within a subject also allows for a more natural evolution of student understanding, with connections much easier to establish and build upon. The intensive nature of the course does set it apart from a traditional environment, potentially impacting the long term retention of student understanding. While this limitation should be considered when applying any of the summer school methodologies in a more conventional environment, the impact on learning within the summer school should be minimal.

This intensive environment also allows for much greater control over the educational environment. Potential variables such as student engagement with material or time invested in the subject are no longer a significant concern. This adds a level of stability to the cohort and reduces the noise in the data. If a change is made to the delivery of part of the Summer School, it is much easier to attribute any improvement in performance to that delivery change when the environment is so controlled. This also allows for fair comparisons between years.

The Summer School is delivered and administered by physics education research specialists. This means that any intervention introduced will be delivered by an expert, implementing the idea in the best possible environment. Additionally, as a consequence of the intensive environment and focus on group work, the staff delivering the Summer School have a closer, and therefore often a more positive relationship with the student cohort than in a traditional introduc-

tory physics course. Students feel engaged with the Summer School through these relationships and this is key when introducing potentially unfamiliar educational structures.

Additionally, with the high number of contact hours and physics education research specialists, there is the opportunity for projects including observation to be utilised. To identify trends in behaviour, observations need to be done over multiple sessions. The structure of the Summer School provides the time to observe small groups and identify the individual interactions, but still within the broader context of the Summer School.

1.2.2 Cohort

Unlike a traditional cohort, there is a highly rigorous selection process to join the Summer School, which creates a very high achieving, and uniform cohort both within and between years. These students attend a highly selective college in the United States and arrive at this Summer School already competent and experienced as students in higher education. The majority of these students are majoring in life sciences and are at the same stage of their undergraduate degrees. The cohorts for the Summer School are incredibly low variance, with consistent academic experience and outcomes. No cohort within traditional higher education would be able to practically achieve this level of consistency. This means that when there are variations between students due to interventions, the variations are not lost in the noise and can be clearly attributed. The nature of the cohort construction does mean that there is limited scope for understanding the impact of demographics. This area of research is also limited by the size of the cohorts, which range from 77 to 126 through the four years analysed in this manuscript.

The consequence of the selection process is that, beyond just creating consistency in the cohort, the students demonstrate greater commitment to the Summer School. Greater enthusiasm from the student cohort provides a much better environment to see the best outcome for an intervention. For an intervention to succeed there must be a will, not only from the educator but from the cohort, that it succeeds.

A specific variation which has been maintained throughout all iterations of the Summer School considered in this manuscript is the two stream structure. The vast majority of students are part of the Life Sciences cohort, those pursuing a major in a non-physical science subjects. There is a small cohort of Engineering students, those pursuing a major in physical science subjects. There are differences in the material covered by each cohort, and therefore a difference in the summative assessment completed. For the Engineering cohort this allows for a more significant focus on mathematical and procedural skills. For much of the Summer School both cohorts work together, with students from both cohorts in the majority of peer groups. As there is little variation between these cohorts outside of what has been imposed through the varied learning objectives for each stream, this structure allows for an analysis of the impact of cognitive diversity.

This cohort construction is non-normal. While a traditional introductory physics cohort may

not consist wholly of physics degree track students, they will make up the majority and there is a certain expectation of prior knowledge. Within the summer school this expectation does not exist in the same way. The students may not have any physics background, through they do have a broad scientific background. While this lack of knowledge may be balanced by the greater metacognitive skill of an already higher educated student, this does create a difference in the demographic make up of cohort and as such will subtly change the engagement with and impact of any interventions introduced.

1.2.3 Assessment

The intensive nature of the course also provides the space for more formative and summative assessment.

The formal assessment methods for the summer school are class tests and exams consisting of a mix of multiple choice questions (MCQs) and short answer questions (SAQs). MCQs are well suited for testing conceptual understanding and were the basis of the full class sessions. These sessions gave students a chance to approach conceptual ideas as they would be presented in the assessment. These sessions were not completed as part of the tutorial group, with students interacting with in various groups throughout all of the sessions. In contrast, the questions used during the small group tutorials were designed to emulate the style and difficultly of the SAQs in the formal assessment. As the tutorial questions are always completed only during the group sessions, differences seen in the performance of the SAQs can be ascribed directly to the small group tutorials.

Higher frequency, especially of summative assessment is an incredibly useful tool as it shows a much clearer picture of the trends within a cohort, even in the condensed time-frame of the Summer School. In addition, the use of many individual assessments allows for comparison of question types and styles. With a larger pool of individual questions, specific conceptual areas or taxonomic ranks can be significantly analysed. The use of both multiple choice questions and short answer questions provides a diverse range of question types, providing more opportunities for potential variation in performance to arise.

1.3 Project Outlines

Utilising a physicist's approach and the unique environment of the Summer School, five projects were undertaken. Firstly, an analysis of the assessment tools was undertaken. This is discussed in Chapters 2 and 3. This project aimed to identify what aspects of the assessment are truly necessary; what does the assessment suggest are the key skills to succeed in a physics course. Fully identifying and understanding the limitations of the tools being used must be part of any experiment. A form of calibration was conducted by comparing the assessment with standardised

tests. The analysis considered all parts of the assessment, from broad question type to discrimination power. The large amount of data provided by frequent summative assessment allowed for analysis at a deeper level than is often possible. As the Summer School utilises both multiple choice questions and short answer questions, the analysis will be applicable for the vast majority of introductory physics courses. Though it is important to consider the limited use of problem solving focused questions which may limit the relevance to honours level physics courses.

Following on from the assessment project, a similar approach was taken to identify the impact of individual student characteristics on academic performance, covered in Chapter 4. As this student cohort is more homogeneous than a standard cohort, variances due to gender, prior academic experience, or age may no longer be statistically significant as has been seen in previous studies [Seyranian et al., 2018, Whitcomb and Singh, 2020]. The ability to control for so many individual variables makes this an invaluable opportunity to more accurately analyse the impact of these variables. This is especially necessary when many of these variables may be correlated. Only by controlling for as many variables as possible can the truly significant become apparent.

Projects 3 and 4, covered in Chapters 5 and 6 and building on the individual student analysis, considered the impact of the group setting on student outcomes. Project 3 looked to codify the measurement of group dynamics in a quantitative way. Multiple techniques were used as this is an area which has previously not been explored in a quantitative way. Much of this project is focused on how to integrate a quantitative philosophy and methodology into a traditionally qualitative area. As was highlighted in the introduction, different but complementary approaches to education research can provide new insights. Project 4 was a more practical analysis of group dynamics. By utilising the already very clear delineation between the Life Sciences and Engineer cohort, an analysis of academic diversity in a group setting was possible. While analyses have previously considered the impact of an individual's knowledge and prior experience on their individual outcomes, there is little work considering the impact on a group as a whole. As there is a focus on group work within the Summer School, this is the ideal environment to analyse the impact at group level.

The final project considers, through a new lens, the issue of practical learning in physics. This is discussed in Chapter 7. The role of practical learning has been oft contested, with the impact especially on conceptual learning, not clear. To measure the impact of only practical work requires both a clear categorisation of the assessment and also a way of identifying the specific areas of knowledge covered in the practical sessions. There should not be an expectation of broad improvement, the skills being taught in a practical session are not necessarily applicable in other areas of a course, but the conceptual understanding of a specific concept may be enhanced by the demonstration through experimentation. Again, with the benefit of frequent assessment throughout each cohort, an analysis within the cohort can be done comparing the performance in questions related to the concepts within the practical work, and concepts that

were not covered.

1.4 Summary

These projects are building on and developing ideas from many different areas within the physics education research community. They cover many major aspects of education research, even those that are not traditionally considered from a quantitative perspective.

The physics education research community should be striving to integrate into the standard process of teaching physics in higher education. The methods and results presented in this manuscript are accessible to any physics educator, whether they have a background in education research or not.

The Summer School is a unique educational environment, but the data and analyses presented here can be used by all educators. The implementation of an intervention will be different for every individual instance. Educational settings are highly varied and cohorts can change massively year on year. Identifying whether an intervention has succeeded can be difficult when there is no standard to compare to. All physics educators need to be able to analyse their own cohorts and interventions to fully understand how any changes to their courses have impacted results. The idealised environment of the Summer School should be viewed as an effective gold standard for an intervention, the best possible outcome for an intervention. The environment is so well controlled, the students so enthusiastic and engaged, that the impact of a good, well designed intervention should be clear. This is a model of physics education. It is idealised and does not seek to represent every educational setting. What it provides is a useful view of what can or cannot be successful in education, both from the perspective of intervention, but also from quantitative analyses.

A physics perspective on education is focused on quantitative measurements and observations. This manuscript seeks to identify the limits of that approach within education. Projects 3 and 4 posit a new quantitative approach for understanding groups and group dynamics. The analysis of group dynamics is a topic of much debate within education research as collaborative learning is a key component of the flipped classroom system. A breakthrough is needed in the understanding of the communal educational experience, to understand not just the benefits of collaborative learning, but also the mechanisms through which it supports learning [Stöhr et al., 2020, Doğan et al., 2021]. Project 5, student-led laboratory teaching, considers the impact of peer interactions in a more clearly defined, traditionally quantitative way. The final goal, however, is still to understand how to measure the impact of the sharing of ideas and experience.

This manuscript is a presentation of a model of physics education, with a focus on how to measure the impact of peer-focused and led group work.

Chapter 2

Understanding Assessment Part 1

2.1 Introduction

A standard approach to quantitative education research will be built on assessment tools and data. Assessment is an inherent part of almost any educational setting, and as such is widespread in its use as a research tool. However, any tool or measurement that is taken must be quantified within the context that it is used. A calibration of what is being measured by assessment is necessary to fully understand any data produced by that assessment [Nuttall et al., 1987]. It is important to consider both the reliability of the assessment and the validity of the assessment [Downing, 2003, Downing, 2004]. Assessments should align with the learning outcomes of a course but there is often an assumption that the assessment will provide a fair view of a student's level of knowledge and skill without considering the impact of skills and knowledge outwith the purview of any course and how that can be applied to solve the problems presented within the assessment. This issue has previously been addressed in physics education when considering the notable gender disparity in performance of the FCI [Normandeau et al., 2017, Lorenzo et al., 2006]. However, while there are some other example of physics education considering assessment validity [Dewi et al., 2022, Jandaghi, 2010], it has not been embedded in the overall research pedagogy. As the educational landscape has changed so significantly in the 2020s and adaptations continue to be made to methodologies and approaches to learning and teaching, this kind of reflection becomes more and more important.

The variability of student approaches are not always accounted for in the structure of the assessment. This is both a question of how to predict student performance in terms of success within the assessment, and what approach is taken to problem solving, including consistency of approach and interpretation [Gijbels et al., 2005]. Assessment not only has to accurately measure student knowledge and skill, but also be accessible to a range of students, not limiting success to variables outwith the course being assessed [Watty et al., 2010]. By considering the potential for biases within the assessment, both the questions "Is this assessment a valid tool for measuring interventions?", and "Can assessment be used to predict student performance?" can

be answered.

For the assessment to be a valid tool a consistent, reliable distribution of performance is necessary. Student engagement with the assessment should not be highly influenced by individual interpretation or subject knowledge outwith the purview of the course. However, there should be variation based on the difficulty of the question, with high levels of attainment across the full cohort for bookwork questions, and problem solving questions highly discriminating within a cohort. This can fail when the expectation of the question setters does not match the behaviour of the student cohort. Without an understanding of the background of the students, without considering how a novice physicist would approach a question, then the assessment may not assess in the way expected. The difficulty of a question is relative to the method taken. If a student can memorise a simple calculation or derivation, their approach will not match the one expected, and create what appears to be an inconsistent performance. If a student's approach can be quantified within the trends of performance, not at overall assessment level, but at the level of individual marks, an identification of different approaches can be taken. While individual students cannot be predicted, how sections of a cohort will approach different question types can be predicted. This then creates a much needed baseline from which interventions can be measured.

This manner of measuring assessment is useful in and of itself, as predicting student performance has been a persistent aim in educational research and has been approached from a variety of perspectives. Qualitative approaches have often been used, looking at the impact of learning environments (both physical and digital) [Qu et al., 2019,Sivarajah et al., 2018] on future student performance or using teacher intuition as a predictor of future performance [Foreman and Gubbins, 2015]. Similar approaches have also looked at intrinsic measures of the student to predict outcomes [Bodin and Winberg, 2012]. Given the wealth of data that exists from digital learning platforms like MOOCs, machine based learning is also being integrated with educational data mining to achieve similar aims of prediction [López-Zambrano et al., 2020,Qazdar et al., 2019].

This project seeks to accurately categorise assessment and identify if there are biases within the assessment that make it an unreliable measurement for the impact of interventions. This categorisation will also provide an insight into how novice physicists approach physics assessment and whether that can be used as a predictive tool. This is a much needed exploration of the assessment of physics in higher education and will provide a picture of how students engage with assessment in physics.

2.2 Methodology

The assessments considered in this project are the summative Class Tests and Exams used throughout the Summer School. The format of these assessments are consistent, with 50% of the grade based on multiple choice questions (MCQs) and 50% based on short answer questions (SAQs). The use of two forms of assessment provides an additional avenue of analysis. How-

ever, as there are different skills required for MCQs and SAQs, once categorised, all analyses are only within, and not between, these question types. Within the SAQs, each question part was categorised separately as most SAQs contain some bookwork, application and explanation sections.

2.2.1 Categorisation - Bloom's Taxonomy

To conduct this analysis, it is first necessary to clearly outline how the assessment was categorised. The standard framework for question categorisation is Bloom's Taxonomy [Bloom et al., 1956]. Bloom's taxonomy is a framework which classifies learning objectives into a hierarchy based on the necessary cognitive skills. This can help describe what areas can cause stress at different points throughout a course or assessment. It is therefore imperative to understand how a piece of assessment fits into this framework and how that aligns with the framework within the course. Bloom's taxonomy, however, encompasses all areas of assessment and goes beyond what any introductory level course would require. For this project the first three levels of Bloom's Taxonomy are the most relevant, Knowledge, Comprehension and Application.

Bloom's taxonomy can be applied to any style of assessment so is used throughout this project with regards to both MCQs and SAQs. It is expected that student attainment should be highest for Knowledge questions, with the discrimination power increasing as questions move further up the framework.

2.2.2 Categorisation - Mathematical Skill

As this is a physics course with a component of mathematics involved there is another axis of variation in question writing - mathematical skill. Not all questions in a physics course require mathematical skill but it is a necessary component overall. However, as it is a physics and not a mathematics course, the core of the assessment should never be focused on assessing the mathematical skill. This is also clear in the approach with regards to teaching these skills. Mathematical understanding is not orthogonal to physics understanding but they are not necessarily intertwined, especially in the approach of a novice. Mathematical understanding may provide an additional approach to a question but is not a substitute for conceptual understanding. It can, however, disguise a lack thereof, just like a lack of mathematical ability may suggest a lack of conceptual understanding. It is therefore important to understand if questions with a mathematical focus do not correlate with student performance across the board.

2.2.3 Adapted Bloom's Taxonomy

Within this project an adapted form of Bloom's taxonomy is used, as outlined in Table 2.1 which takes into account the necessary separation of mathematical understanding from general

Bloom's Taxonomic Rank	*SAQ Category*	*MCQ Category*
Knowledge	State	State
Comprehension	Explain	Comprehension
Application	Calculation	Calculation

Table 2.1: Bloom's Taxonomy Categories

This table shows the equivalent categories used within this manuscript to the traditional Bloom's Taxonomy. Categorising calculation question within application is defined within Bloom's Taxonomy, though discussion of comparative complexity with Explain or Comprehension questions will be found at the end of this Chapter. The use of Explain for the SAQs and State for both MCQ and SAQ aligns these categories with the command words from the assessment.

conceptual understanding. As this is already considered within Bloom's Taxonomy, Calculation questions have been categorised as Application questions [Bhaw and Kriek, 2020]. A small number of questions are categorised as Derive or Diagram but there are so few questions in each of these subcategories that they are not included in this analysis.

While the method of recording the answer is different for MCQs and SAQs, the cognitive processes to come to the answer are not necessarily different. When a student is presented with a State question either as an MCQ or an SAQ, it is still a recall question. While there are additional skills that can be used when answering MCQs, the core process still relies on student recall. This is also the case for Calculation questions, where MCQ answering techniques may be less relevant. If all possible answers are within a reasonable estimated range then the student must rely on their conceptual knowledge and mathematical skill. Again the core process is the same, and the answer is likely in the same format for both MCQ and SAQ. For Explain and Comprehension, the format of the answer is different, but the final conclusion is the same. An MCQ that falls into the Comprehension category should require all of the logical steps outlined in the written answer to an Explain SAQ, but the answer given is only the final conclusion. This again is the student using the same cognitive processes, but presenting their solution differently.

The analysis was carried out for MCQs and SAQs separately but given that the Taxonomic categorisations are the same, trends may be identified between these question types.

2.2.4 Analytical Methods

As this project is seeking to demonstrate any potential outliers in student performance based on question type and category, correlations were extensively used. An initial correlation of MCQs and SAQs with overall student performance was conducted. The Pearson correlation coefficient was used and is calculated as:

$$Correlation(X,Y) = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 (y-\bar{y})^2}} \tag{2.1}$$

For this coefficient, correlations beyond 0.7 are considered highly positive correlations.

Student performance in each question category was then correlated for each component of Bloom's taxonomy, with each student represented 3-dimensionally at all times so as to make comparisons as clear as possible. An analysis of the facility and discrimination values of each question type was also conducted to explain any variations seen in the correlation graphs.

The facility index of a question is simply a description of the difficulty of the question, shown by the mean performance. In this manuscript it is shown as the difference from the overall average performance. This was done to provide consistency as all other results are shown after normalisation. The discrimination index describes how well the question differentiates between the highest and lowest performing students and is shown as the difference in average performance between the third highest and lowest percentiles. Item analysis has already been done in this way to evaluate new assessment in physics education [Klein et al., 2017, Day and Bonn, 2011].

Correlation analysis can be considered a broad strokes analysis but with this level of categorisation, any effects, biases or group of students outperforming in question type should be evident. Education analyses must always be conducted at a scale that will still be relevant to student outcomes. When considering the impact on student performance as the most relevant metric for any education analysis, that level of impact should be seen in correlation analysis.

2.3 Results

2.3.1 Context - Summer School

As additional context for the results presented here, the overall distribution of the assessment from 2016 to 2019 is shown in Figure 2.1. There is a remarkable level of consistency in the distribution. While, as was outlined in the Introduction, the Summer School cohorts are as homogeneous as is practicable, this distribution has been consistent even over a 62% increase in cohort size. The correlations between each year are presented in Table 2.2. There is no tailing off beyond the pass mark (40%), as is often seen in traditional cohorts. The lack of outliers and the strong consistency are the first metrics showing that this assessment is reliable as a tool for measuring interventions. The deeper analysis could only be conducted for 2018 and 2019, the data is included for 2017 in this section to provide greater context of the consistency of the Summer School.

2.3.2 SAQ & MCQ Correlation

The analysis was conducted at SAQ and MCQ level before considering any individual question categories. As all further analysis was conducted within the question types, a broad analysis of MCQ and SAQ provides a framework within which to understand the analysis.

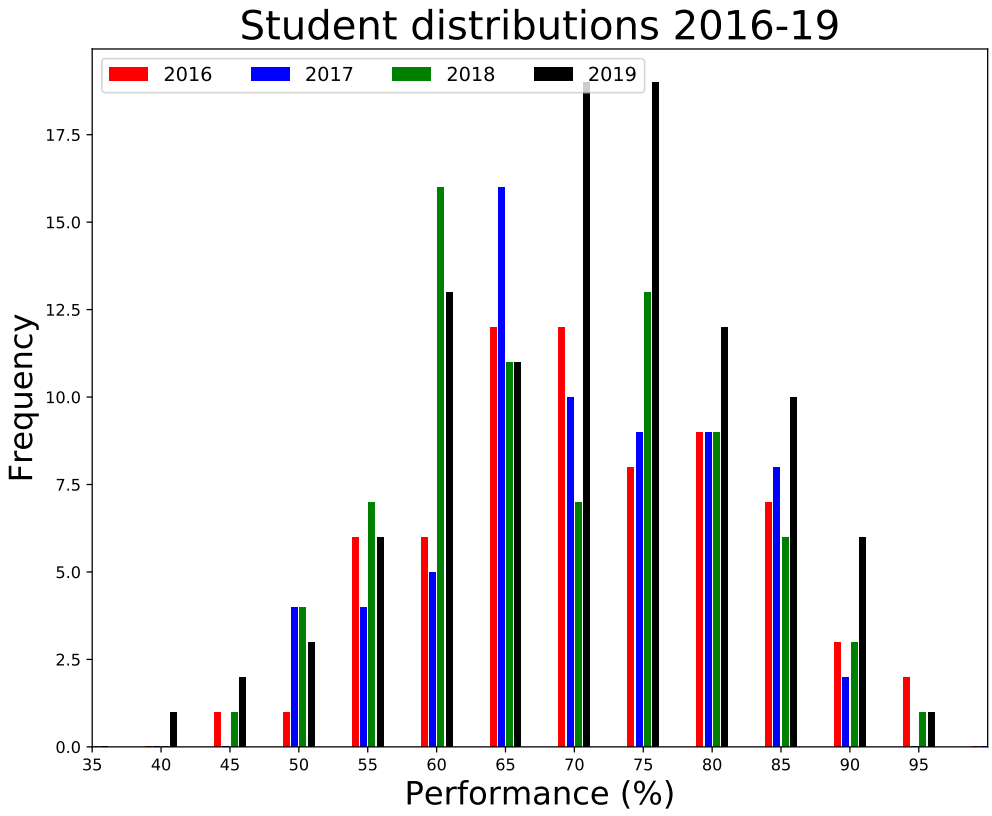


Figure 2.1: Overall distribution of grades 2016-2019

Shown here is the distribution of student grades from 2016 to 2019, encompassing all years that are included in the analysis throughout this manuscript. The overall distribution is similar year on year, even with an increase in students numbers each year, and there is no tailing off below the pass mark (40%). This data shows a high level of consistency in and reliability with the assessment.

Year	***2017***	***2018***	***2019***
2016	0.934	0.755	0.882
2017		0.737	0.794
2018			0.831

Table 2.2: Correlation of overall performance between 2016 and 2019

There is a high correlation in the performance between each year, indicating that there is an overall high level of reliability in the approach used for the Summer School.

Student performance is just as strongly correlated for either question type as seen in table 2.3. This agrees with the current consensus within the physics education field, that while SAQs give a more granular representation of student understanding, there is a qualitative agreement of student ability from both MCQ and SAQ assessment [Lin and Singh, 2013, Scott et al., 2006]. While the small number of students does not allow for more detailed analysis, there is also broad consensus in the analysis of the engineering cohort. This shows that any variation that is seen in

student performance is not driven by one question type dominating the assessment. Additionally, students can be categorised by performance in certain question subtypes within either SAQ or MCQ, as the analysis is still relevant to overall performance.

Year	***Stream***	***Questions Type***	***Correlation with Overall Summative Performance***
2019	LS	MCQ	0.966
2019	LS	SAQ	0.966
2019	ENG	MCQ	0.958
2019	ENG	SAQ	0.956
2018	LS	MCQ	0.913
2018	LS	SAQ	0.922
2018	ENG	MCQ	*0.694*
2018	ENG	SAQ	*0.640*
2017	LS	MCQ	0.941
2017	LS	SAQ	0.912
2017	ENG	MCQ	0.978
2017	ENG	SAQ	0.965

Table 2.3: MCQ and SAQ Correlation with overall performance

The high correlation for both MCQs and SAQs with the overall performance again highlights the consistency and reliability of the assessment as well as signifying the importance of both aspects of the assessment. The agreement with the engineering cohort, a significantly smaller cohort than the life sciences cohort, is shown for 2017 and 2019. The cohort in 2018 was larger but also had a greater diversity of students in terms of major and prior knowledge of physics. This may be the reason for the significantly different correlation coefficients. These students continued to perform well in continuous assessment and so there is a disparity in their performance.

In general, students perform better in SAQs rather than MCQs. While this can be explained by SAQs providing a greater opportunity to demonstrate partial understanding as opposed to the binary of MCQs, it is not a statistically significant difference. However, a small group of students each year do perform marginally better in the MCQs. Treating these students as a group, there are no indications that performing better in MCQs has any noticeable impact on overall performance as can be seen in Figure 2.2. This is likely due to the limited MCQ/SAQ delta values, the variation from the average, in either direction and aligns with the highly correlated nature of both the MCQs and SAQs with the final student scores.

2.3.3 Adapted Bloom's Taxonomy Correlation

As there was little variation in performance between the SAQs and MCQs overall, an initial finer analysis only considering 2019, was conducted by comparing performance between different question types. These correlations can be seen in Figure 2.3. Similar to the MCQ v SAQ graphs there is a very strong correlation between different question types, though it is clear that performance in state questions is consistently higher in relation to both calculation and compre-

hension/explain questions. This aligns with the expected performance as outlined by Bloom's taxonomy. Figure 2.3 does however, show an unexpected effect within the group of high performing students. These students consistently perform better in calculation questions compared to comprehension/explain questions. While this is not a large group, the drop in performance within the comprehension/explain questions does suggest an area for further exploration.

While the strong linear correlation makes this assessment particularly useful to measure the impact of interventions, it suggests that there is an issue with what the assessment is assessing. Broadly, State questions should be accessible to all students as these are recall questions that often do not need a deep understanding of the subject matter. This suggests that there is another factor at play, another skill or approach that is causing students to struggle with more basic questions. A further analysis, considering only MCQ or SAQ for each question type was conducted to clarify if this linear correlation was consistent.

2.3.4 Modified Bloom's Taxonomy Correlation within MCQ

To further investigate this linear correlation the analysis was conducted again, considering the MCQs and SAQs as separate data sets. This analysis was conducted for 2018 and 2019.

Considering the MCQs alone showed that performance in MCQs varies very little between questions when considering their taxonomic rank. Figures 2.4 and 2.5 show a similar story for both 2018 and 2019 with high correlations between all types. The tailing off of performance within the comprehension questions for high level students, seen in the overall modified Bloom's Taxonomy correlation, is non-existent within the MCQ analysis.

The strongly linear correlation in the state questions with both calculation and comprehension questions is even clearer when considering only the MCQs. This suggests that additional skills necessary to answering MCQs are obscuring the differences based on the taxonomic rank. Understanding a student's approach to these questions, as there is no working provided by the student, can also be very difficult. It is possible that the categorisation of these questions does not align with the student experience of them. Research within the STEM education field [Zaidi et al., 2018] has suggested that the disconnect between examiners and students with regards to how to approach a question may mean that the ranking of questions within structures such as Bloom's taxonomy are inaccurate or missing nuances, especially when considering the intersection of mathematical and conceptual understanding.

The use of simple mathematical tools such as proportionality or geometry are intrinsic in the approach that a physicist takes to solving a problem. A novice physicist may not think to use these tools, or does not have the knowledge or prior skill to implement them. This can vastly change how an individual engages with a question. What the strongly linear correlations for all MCQs suggests is that the skills necessary to answer these questions, be that mathematical or cognitive, in the approach to answering an MCQ, are far more significant than the content of the question. Conceptual understanding is still necessary, but the trends in the data do not provide

a measure of the understanding of the students, they provide a measure of how well the student can answer MCQs.

This can still potentially be used as a tool for predicting student performance but it is measuring a variable which is correlated with student understanding, not directly measuring it. A consideration of how students will engage with assessment must always be done when designing any course. Especially within physics, it is important to consider how you wish students to demonstrate their conceptual understanding in contrast to the mathematical skills required to apply that conceptual understanding. MCQs may not provide that distinction as students cannot provide their working in addition to their answers.

2.3.5 Modified Bloom's Taxonomy Correlation within SAQ

A similar method was used for the SAQs. The correlation between state, calculation and explain are much less linear than for the MCQs, seen in Figures 2.6 and 2.7. There is a shift towards more positive performance for state questions, and this is likely where similar shift in the data for the overall modified Bloom's taxonomy correlations originates. Comparing the State question data with that for the MCQs, there is support for the idea that skills necessary for answering MCQs are causing the linear trend, rather than a lack of conceptual understanding. Some consideration must also be given to the fact that marks can be given for partial answers in SAQs - students can more easily demonstrate some, if not complete, understanding. This effect is not so significant for calculation or explain questions.

There is a somewhat linear relationship between calculation and explain but it has a gradient less than 1, showing a more clear version of the correlation seen Figure 2.3. Explain questions are done disproportionately well by students in the lowest 30% of overall performance and vice versa with calculation questions and students in the top 30% of overall performance. A similar trend is seen in both 2018 and 2019. While there is a potential to predict student performance based on this trend it can also be explained by variation in facility and discrimination indices for the question types.

In Table 2.4 the facility and discrimination values were calculated for 2019 and 2018. The general trends between the different subtypes are similar for both years. The key values are for calculation and explain as these values show where the trends seen in Figures 2.6 and 2.7 originate. As expected the facility values are similar for calculation and explain questions. This similarity is not seen with the discrimination index however. For both 2018 and 2019, there is a more than 10% difference in the discrimination, with the calculation values differentiating high performing students much more successfully than explain questions. This highlights that there are some limitations to taxonomic ranking, as within Bloom's calculation questions would be considered more complex than explanation questions but does not indicate whether this discrimination is from a conceptual physics aspect, or from a practical mathematical skill aspect.

Question Subtype	*2019*		*2018*	
	Facility (%)	*Discrimination (%)*	*Facility (%)*	*Discrimination (%)*
State	8.81	22.60	9.77	7.43
Calculation	-2.17	33.23	-3.90	36.76
Explain	-3.67	21.63	0.78	19.64
Diagram	9.69	19.41	14.36	14.99
Derive	-10.21	36.92	12.41	31.05

Table 2.4: Facility and Discrimination Indices (2019 & 2018)

The facility and discrimination indices were found for each question subtype for the SAQs in both 2019 & 2018. The facility index is show as distance from the overall average performance. The discrimination index is the difference between the top and bottom 1/3 percentiles. The trends identified are similar for each year. While the facility values are similar for both calculation and explain questions, particularly in 2019, there is a much a clearer difference in the discrimination values. While the facility values would align explain and calculation questions in the same taxonomic category, the discrimination values show that student engagement with the questions is varied.

2.4 Discussion

When considering how to interpret the results of assessment there is a wealth of data tied into the format of the assessment. How a student interprets a question is incredibly important and how a question is formatted has a significant impact on that. The results here have shown that MCQs, no matter where they fall within Bloom's taxonomy, will create the same ranking within a cohort. This suggests that there is an impediment to student performance within the MCQ structure. A student should not be able to perform just as well in a question that requires problem solving as one that only requires recall, unless the barrier to success is within the structure of the question. Student performance in MCQs and SAQs are still highly correlated however so there is either a common skill to answering both styles of question or there is some measure of physics knowledge that is tempered by the skill of answering MCQs.

The barrier to entry when considering SAQs may not be different but the evaluation of performance in these questions allows for student working to be shown. A student has the opportunity to demonstrate some understanding without requiring a complete knowledge. These questions provide a more nuanced view of student understanding, highlighted in the results by the significantly higher levels of discrimination for calculation questions versus explain questions. These question types are taxonomically similar, have very similar average performances and assess the same kinds of conceptual understanding. The reason for the significantly different discrimination values can only be caused by another skill being necessary to complete the questions. In this case, it is clearly mathematical skill that is the discriminating factor.

However, there are mathematically focused questions within the MCQs and there is no discrimination. This may be due to the lack of range, the difference between 0 and 1 is not as great as the difference between 0 and 4. There is also the potential that the categorisation of a question as calculation or explain/comprehend is different for a novice versus an expert.

While the results provided in this chapter have highlighted various skills required to suc-

cessfully answer a question, these are not the only skills that a student can use. The skill of answering an MCQ is not one single process, it incorporates problem solving, textual analysis, logic and many others. What these results demonstrate is that understanding how a student performs in an assessment goes far beyond the understanding of the course material.

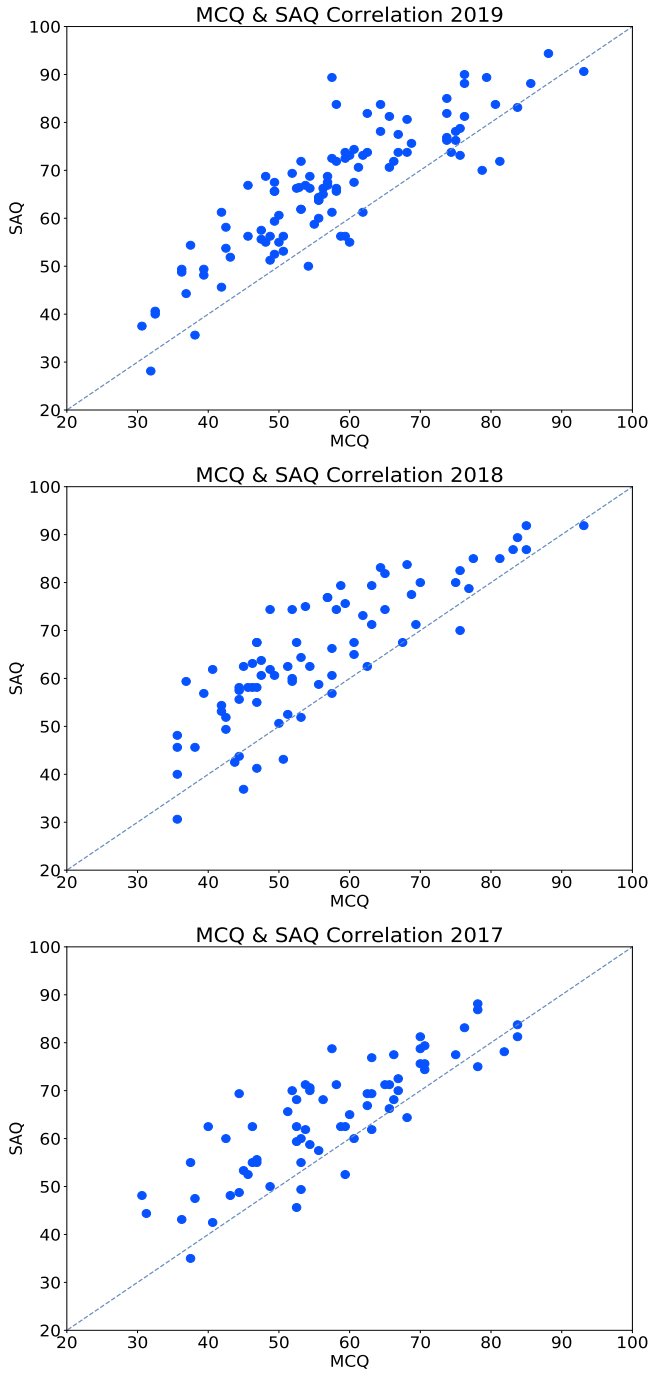


Figure 2.2: Correlations of Student Performance in MCQs and SAQs 2017-19

This figure shows the correlation between the MCQs and SAQs for all students within the life sciences cohort for 2017 to 2019. The data is very similar year on year, with students showing comparatively high attainment in SAQs, with less than 15% of any cohort showing comparatively higher achievement in MCQs. However, the delta value between MCQ and SAQ performance within the group that performed better in the MCQs is small and is evenly spread amongst overall high and low attainment students.

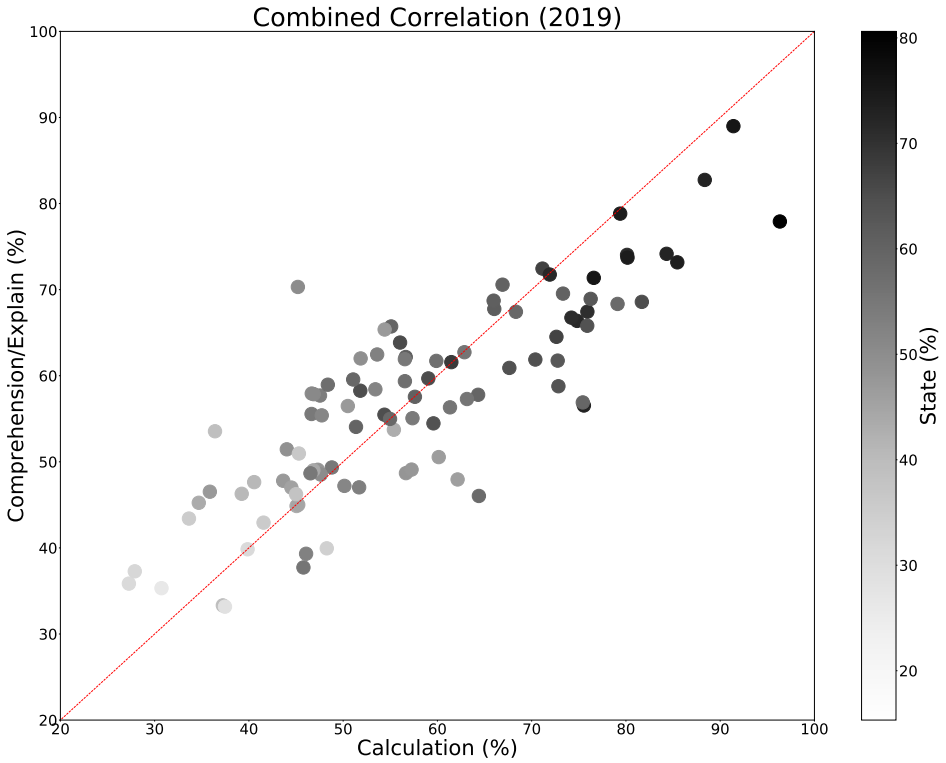


Figure 2.3: Correlations of Student Performance in question types (2019)

This figure shows the correlation between the different question types for the 2019 cohort. There is, similar to the correlation graphs for MCQs v SAQs, a general linear trend, with a similar shift to higher attainment in State questions, similar to SAQs.

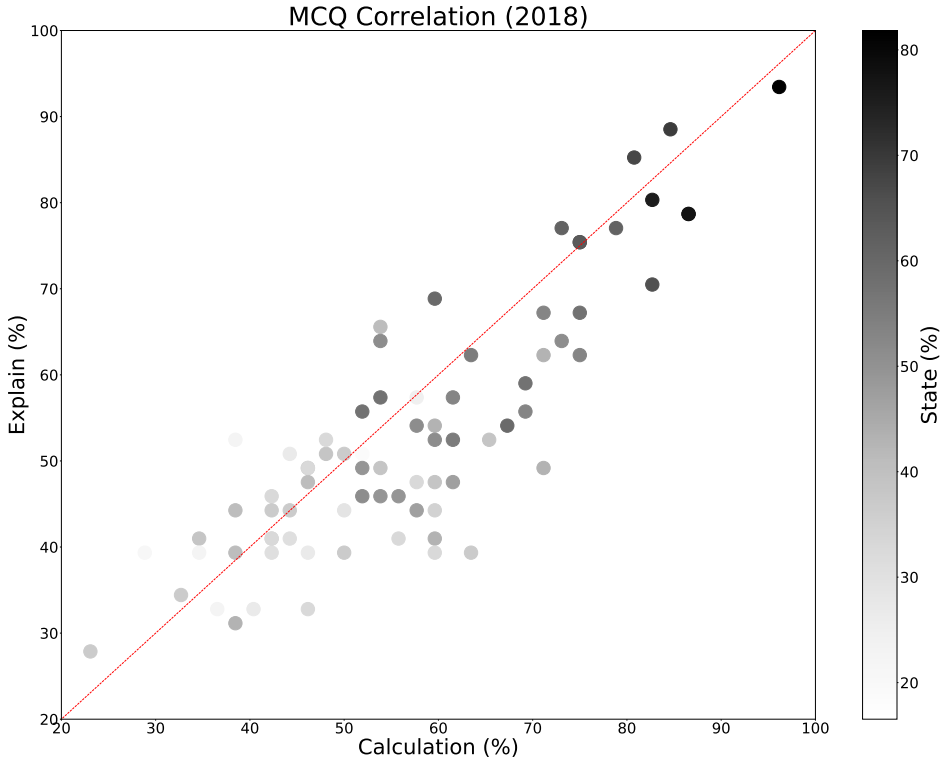


Figure 2.4: Correlations of Student Performance in categorised MCQs (2018)

This figure illustrates the correlation of the different MCQ question types for the 2018 cohort: State, Calculation, and Comprehension. The correlations are strong and consistent, as was expected given the similar averages shown in the table 2.1. The correlation values, 0.810, 0.809, and 0.854 respectively, indicate that there is little variation in student performance within the MCQs, though a higher correlation between Comprehension and Calculation questions may be explained by partially occupying a similar rank within Bloom's taxonomy.

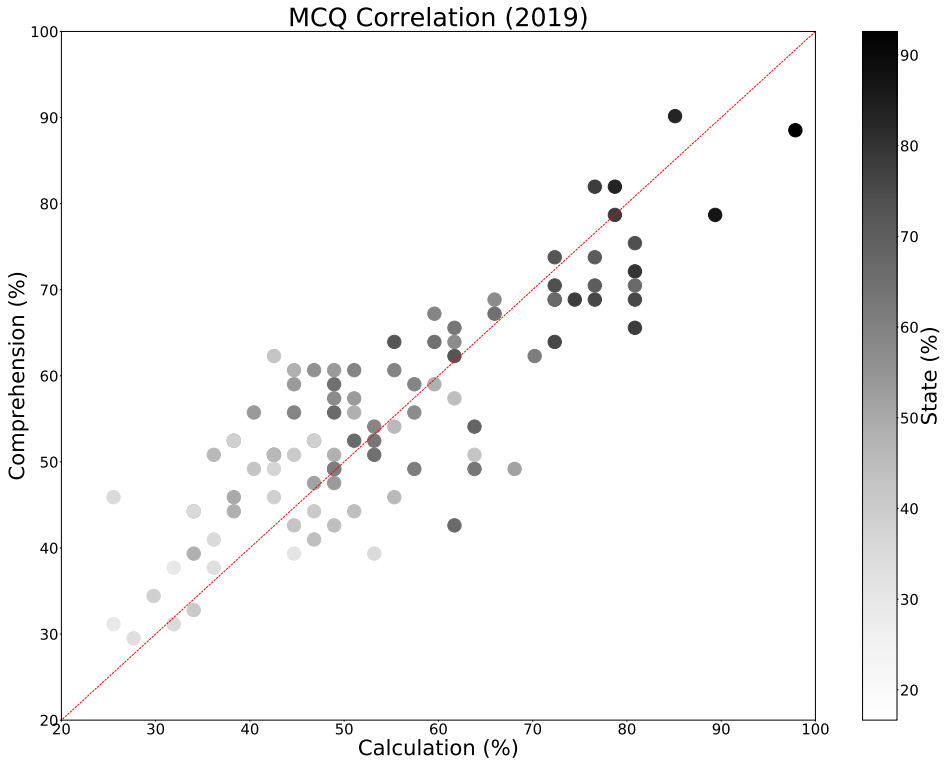


Figure 2.5: Correlations of Student Performance in categorised MCQs (2019)

This figure illustrates the correlation of the different MCQ question types for the 2019 cohort: State, Calculation, and Comprehension. The correlations are strong and consistent, as was expected given the similar averages shown in the table 2.1. The correlation values, 0.852, 0.850, and 0.850 respectively, indicate that there is little variation in student performance within the MCQs.

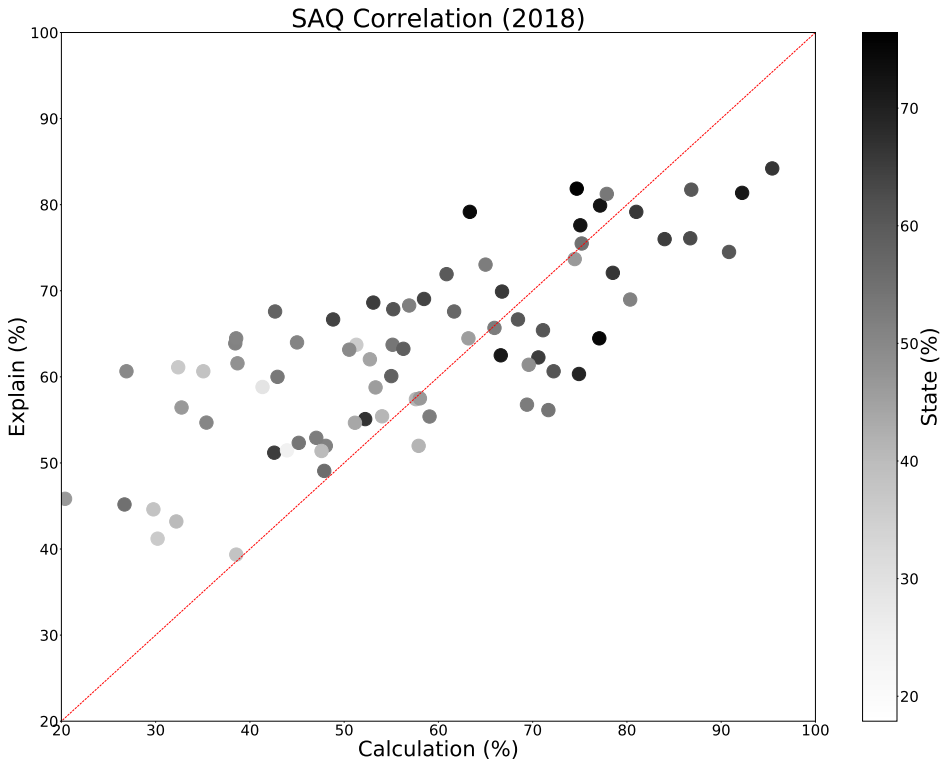


Figure 2.6: Correlations of Student Performance in categorised SAQs (2018)

In contrast to the MCQ graphs seen in Figures 2.4 & 2.5, there is a much weaker correlation for all question subtypes, State, Calculation and Explain, within the SAQs for 2018. There is an obvious attainment difference between State and both Calculation and Explain, aligning with Bloom's taxonomy, and expected. Considering the correlation between Calculation and Explain suggests a potential area of interest as the data has a gradient less than 1.

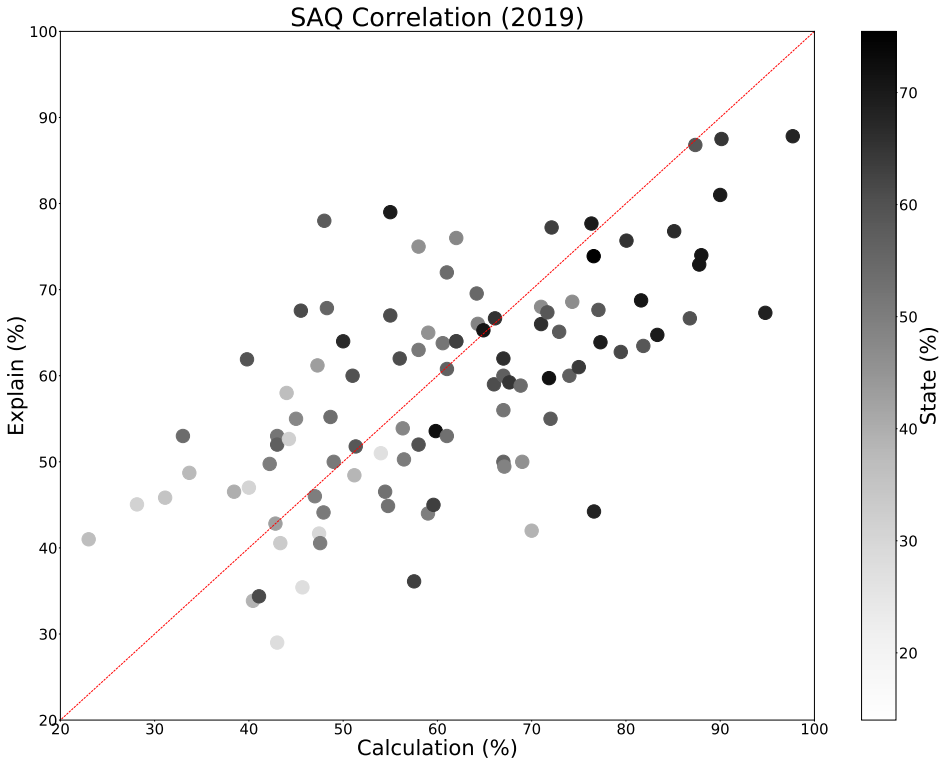


Figure 2.7: Correlations of Student Performance in categorised SAQs (2019)

The graph for the 2019 SAQs shows a very similar picture to that of 2018 as seen in Figure 2.6, again showing a similar gradient for the correlation between Calculation and Explain. As with many of the results discussed in this chapter, there is a high level of consistency within the data.

Chapter 3

Understanding Assessment Part 2

3.1 Introduction

In Chapter 2, Understanding Assessment Part 1, it was identified that when considering assessment through the lens of Bloom's Taxonomy, the student performance in the Summer School was not as expected. Performance in the MCQs was consistent across all taxonomic ranks. Whereas, in the SAQs the performance varied between mathematically focused and conceptually focused questions. This suggests that the question format has a large influence on student performance, especially for MCQs. When considering assessment as a tool to analyse the impact of education interventions, it is important to understand what the tool is measuring. To fully understand the intersection of conceptual understanding and question structure on student performance and the use of assessment as a tool, a way to measure conceptual understanding is needed. For this assessment this means a measure of both physics knowledge and mathematical skill.

The Force Concept Inventory (FCI) [Hestenes et al., 1992] is a standard set of questions used throughout physics education to assess student understanding of the basic concepts that underpin much of classical physics. The questions are designed to be accessible to all and avoid using jargon or unfamiliar settings. The FCI is often used as a pre and post test to assess the impact of a course on students' conceptual knowledge. It has been used in this way for the Summer School since its inception. If an assessment is designed to only require physics knowledge to solve, then the FCI (either pre or post test) performance should correlate well as they should be measuring the same variable.

Similarly, for the Summer School a Maths Skills Test (MST) was introduced in 2018. This test, like the FCI, has a set of standard questions which cover all aspects of mathematics used during the Summer School but removed from the physics setting. If an assessment can be solved mathematically and without a strong physics understanding, then the MST performance should correlate strongly.

With these standard measures, a comparison of each question type can be done.

3.2 Method

In Chapter 2, each question from the Summer School assessment was categorised within Bloom's Taxonomy. Within this structure these questions were already separated along conceptual and mathematical lines. This allowed for a simple correlation to be calculated for each question type within each year, similar to those calculated in Chapter 2. The initial hypothesis suggests that the MST should correlate with Calculation questions, and that Explain questions for SAQs and Comprehension questions for MCQs should be better correlated with the FCI.

The categories of derive and diagram have been included in this analysis as they are representative of the extremes of mathematically or conceptually focused questions respectively. Derivations can be approached from a purely mathematical perspective and do not necessarily rely on an understanding of the physical context. Diagram questions require a deeper understanding as a student must demonstrate their understanding through a clear model. This cannot be simply regurgitated from a textbook. They also often incorporate multiple concepts and the intersection of different ideas is where complexity arises.

In 2017 the MST had not been introduced and so correlations could only be calculated for the FCI. In 2019, no FCI test was completed by the students and so a correlation can only be calculated for the MST.

3.3 Results

The results have been divided between SAQs and MCQs. The categories used are different to accommodate the different ways in which the questions are answered. State and Knowledge, and Comprehension and Explain are equivalent Bloom's Taxonomy levels. The same correlation analysis was conducted for both question types.

3.3.1 Short Answer Questions

2017

The FCI pre-test was given to the whole cohort during the first laboratory session on the second day of the course. The post-test was given to the cohort during the final laboratory session. The data for 2017 is shown in table 3.1. The question types are ordered as they are within Bloom's Taxonomy, from lowest (State) to highest (Derive) levels of complexity. None of the correlations are extremely significant, however there is a large range of correlations, with a minimum of 0.09 and a maximum of 0.64.

It is unsurprising that the derive questions have the lowest correlation with either the pre or post-test. Derive questions can be categorised as skill-only questions, a variable which is not necessarily related to any aspect of the FCI measurement. In a similar way, both State and

	State	***Calculation***	***Explain***	***Diagram***	***Derive***
Number of questions	7	36	20	8	3
FCI (pre test)	0.35	0.64	0.59	0.33	0.11
FCI (post test)	0.40	0.57	0.61	0.28	0.09

Table 3.1: 2017 SAQ correlation values

FCI correlation values for SAQ question subcategories as outlined in Bloom's Taxonomy. FCI pre-test was completed at the start of the Summer School. FCI post-test was completed during the final laboratory session of the Summer School.

Diagram questions have low correlations with the FCI.

Diagram questions, while requiring strong conceptual understanding, require additional skills to complete, particularly spacial and visual awareness. It is impossible to say what influence these skills have on student performance without a standardised measure. This may account for the low correlation.

For the State questions, the low correlation is likely due to the low discriminatory power of these questions, as shown in Chapter 2. Within the Summer School, the results of the FCI are highly varied. State questions are designed to be accessible to all students, relying on recall only, with no further application of the knowledge. As such these questions are assessing different aspects of student knowledge and a low correlation is to be expected.

As expected there is a high correlation with Explain questions for both the pre and post test FCI. While these correlations are not close to 100%, when comparing with the very weak correlations for Derive, Diagram and State questions, it does indicate that there may be similarity in how the students are being assessed. Though this may also be driven by the small sample sizes for Derive, Diagram and State.

The Calculation correlations did provide a surprising result, with the highest correlation with the pre test FCI. This is especially interesting as this is higher than the correlation with explain questions. Considering the context of the students in the course, this can potentially be explained by the suggestion that students with a broader mathematical background are more likely to have engaged with physics to a higher level before starting the course.

2018

In 2018 the students again completed the FCI pre and post tests. The students also completed the MST at the start of the course. This was done to highlight to the students what kind of maths skills would be necessary for the summer school and as such was not posited as assessment. However, it is still useful as a metric of student knowledge.

Overall, the correlations with the MST results are much lower than for the FCI, as shown in Table 3.2, and this is to be expected as these questions are not physics focused and should only be correlated with the Calculation and Derive questions. However, there is only a limited variation between the different question types. While the Calculation questions are the most

correlated with the MST, it is in no way conclusive.

While the FCI values are generally similar to those of 2017, the values for Diagram and Derive questions are significantly higher in 2018. This is likely due to there being few questions in these categories each year and therefore performance can vary quite significantly. The most significant difference overall is the correlation values increasing across the board when looking at the FCI post test. This is encouraging as performance in the post test is consistently higher. Calculation continues to be the highest correlated question type.

	State	***Calculation***	***Explain***	***Diagram***	***Derive***
Number of questions	15	45	27	9	4
FCI (pre test)	0.27	0.57	0.45	0.53	0.25
FCI (post test)	0.45	0.72	0.62	0.62	0.45
MST	0.35	0.39	0.28	0.21	0.32

Table 3.2: 2018 SAQ Correlation Values

FCI and MST correlation values with SAQ question subcategories for 2018. The FCI correlations for State, Calculation and Explain questions align with those for 2017. The MST correlations across all categories align with those for 2019.

2019

The results from 2019 only include a comparison with the MST as no FCI was completed this year, as shown in Table 3.3. The SAQ results align with what was seen in 2018 though there is a much clearer outlier in the Calculation question correlation. This may be down to slight changes made to the selection of questions used in the MST, removing those that are not as relevant to the skills necessary for the course for example, and therefore creating a better analytical tool. No changes were made to the wording of any questions. This would also explain why the correlation with State and Explain questions has decreased in 2019.

	State	***Calculation***	***Explain***	***Diagram***	***Derive***
Number of questions	19	43	21	12	6
MST	0.17	0.46	0.19	0.28	0.27

Table 3.3: 2019 SAQ Correlation Values

MST correlation values with SAQ question subcategories for 2019. The correlation for Calculation questions is much higher than all others.

Summary

Considering the strong discriminatory power of the FCI it is perhaps not surprising that the Calculation questions consistently have the highest correlation with the FCI. This performance is consistent across all years and aligns with the evidence presented in Chapter 2. The nature of

Calculation SAQs allows these questions to analyse physics knowledge is an all encompassing way. A lack of conceptual knowledge or a lack of mathematical skill will impede a student. However, if the Calculation questions are correlated with the FCI then the impediment may not be the mathematical understanding but how to apply the conceptual knowledge in an unfamiliar set up. This may be the root cause for the higher discrimination value for the Calculation questions.

3.3.2 Multiple Choice Questions

An analysis of the MCQs could only be conducted for 2018 and 2019. The Diagram questions are considered as a separate category for MCQs when interpretation of a graph or experimental set up was provided. This is a slightly different interpretation to the SAQs, where a diagram question generally required a diagram to be provided as the answer.

2018

The correlations with regards to the MST align with what was seen in the SAQs with the calculation questions providing the strongest correlation, shown in Table 3.4. This is in opposition to the FCI correlations which, like the correlations shown in Chapter 2, are very consistent across all question types, other than Diagram questions. The Diagram questions require a very clearly differentiated skill, spacial and visual awareness. This likely explains the overall lower correlation with the FCI.

While there is little variation between the FCI correlations for each question category, the correlation values are very similar to those for the SAQs. This supports the conclusion from Chapter 2 that SAQs and MCQs are equally accurate measures of student understanding.

	Knowledge	***Comprehension***	***Calculation***	***Diagram***
Number of questions	49	61	54	6
FCI (pre test)	0.60	0.56	0.58	0.35
FCI (post test)	0.71	0.71	0.71	0.41
MST	0.30	0.34	0.46	0.29

Table 3.4: 2018 MCQ Correlation Values

FCI and MST correlation values with MCQ question subcategories for 2018. The Knowledge, Comprehension and Calculation correlations are consistent for each FCI test, reflecting the similar results from Chapter 2. The correlations are not as consistent for the MST.

2019

For 2019 only the MST was used. There is limited variation between the different question types, with the same very low correlation for Diagram questions as seen for both the FCI and MST in

2018, shown in Table 3.5. For this analysis the Diagram question correlation was negative, but this is due to small numbers in this category and is not representative of the category as a whole.

	Knowledge	***Comprehension***	***Calculation***	***Diagram***
Number of questions	55	62	49	6
MST	0.55	0.46	0.51	-0.11

Table 3.5: 2019 MCQ Correlation Values

MST correlation values with MCQ question subcategories for 2019. The MST correlations across all categories are more consistent than those in 2018. The Diagram question correlation is the only instance of negative correlation and is due to the small number of questions within this category.

3.4 Discussion

What is clear from these results is that using external tests like the FCI and MST can show that there is variation in what is required to successfully complete different aspects of assessment. While these tests cannot identify every factor that impacts a student's performance, they can highlight where there is variation within an assessment. The calculation questions consistently have a higher correlation than other question types with the MST and the pre and post test FCI. This may be due to the broader nature of Calculation questions. However this may also be related to other issues that have been raised with regards to the FCI. The validity of the FCI as a tool has been challenged, with performance often divided along gendered lines [Normandeau et al., 2017, Lorenzo et al., 2006]. This has been related to the context that questions are couched within, often using examples that are gendered within society, i.e. sports examples. There may also be a similar effect at work with regards to mathematical language, with the barrier to entry for questions lowered if one has a better understanding of mathematical language. These questions require not only conceptual understanding and mathematical skill, but also the ability to apply that conceptual understanding and mathematical skill in a new context. These high correlations with the FCI for the calculation questions may also indicate that while mathematical skill may be an impediment to many students, the application of the conceptual knowledge may also be the cause of the high levels of discrimination. There may be many other skills that are somewhat correlated with mathematical skill and application of knowledge that are underlying in this data as well. However, what has been identified in these results are the broadly applicable skills.

The correlations for MCQs were also consistent for each test, mirroring the results seen in Chapter 2. This does not include the diagram questions as these correlations were significant outliers. Both the small number of questions in this category and the different skills required to answer these questions contributed to this. This supports the suggestion that there are additional skills that are potentially correlated with these key skills.

The correlations for both SAQ and MCQ with both the FCI and MST are consistent within the years and question types and indicate that both assessment types can be used interchangeably for future intervention analysis when considering physics knowledge. However, it is important to consider that the assessment is, in many cases, measuring a variable that is dependent on physics knowledge but is likely also influenced by other knowledge and skills that a student possesses. This does not take away from the reliability of the assessment, which is consistent year on year. Further analysis of the validity of the assessment is needed to fully understand what variable is being measured.

There are many other variables that can impact student performance, and these will be explored in Chapter 4, but this project has demonstrated that the simple use of correlations and discrimination analysis can identify which variables are the most significant.

Chapter 4

Individual Student Analysis

4.1 Introduction

Considering everything that may impact student performance from the perspective of the assessment itself naturally leads to a consideration of the interaction of the student with the assessment. Much of the research in this area considers this through the lens of student perception [Lizzio et al., 2002, James, Alisa R;Griffin, Linda L;France, 2005]. However, there are a variety of factors, intrinsic to the individual, that may impact their performance and their perceptions of both their performance and the assessment itself. In this case only factors which are linked to easily quantifiable variables derived from academic experience are considered. There is no consideration of socio-economic factors. While this is a limitation, the fact that these students are coming from the same university means that broadly the cultural and socio-economic background of these students will be similar enough.

This analysis, as with all analyses in this chapter unless otherwise stated, was only carried out for the LS students. There is a broad point to be made, that the engineering students almost always perform better on average and therefore clearly prior experience is the most important factor. While this is certainly true, the variation in background is more broad than just greater prior experience in physics. These students potentially are using physics in their major day to day, almost certainly use maths in a similar context day to day, and likely are more motivated to engage with the course as their degree more closely aligns with the material. Conversely, the motivation for many of the Life Sciences students is based on their goals beyond their current majors. As the the Life Sciences cohort are predominantly pre-medical students, an American term which has evolved due to the need to have a degree before applying to medical school, they require physics to succeed in the Medical College Admissions Test (MCAT), the standardised medical school entrance exam. However, there is a difference in motivation based on a requirement to succeed in a chosen field, and studying in an adjacent, but similar field. These significant differences between the cohorts make comparisons between engineering and Life Sciences students impossible as there are too many differentiating factors.

4.2 Methodology

The variables being considered in this analysis are gender, age, year of study, degree major, grade point average (GPA), prior experience in physics (completion of AP Physics, the equivalent of A-level or Advance Higher physics in the UK), campus (within the University of California). Considering all of the variables, a Mann-Whitney-U analysis was conducted to compare the performance of the individuals in each of these categories. This is a non-parametric analytical method often used in educational research settings, exemplified in Wieman's laboratory focused research [Wieman, 2015]. The test is designed to identify variations in distributions using a ranked sum method. This is well suited to analysis on test scores as the ranking also functions as a normalisation process. This test also does not require data sets of similar sizes like many analytical tests, which is particularly suited to this project as group sizes can vary.

Some of the variables are very simple to categorise, gender and year of study for example have easily defined categories (in this case gender is defined as only male or female) and as such are respectively categorical and ordinal. Others, such as GPA or major, were amalgamated into broad ranges with similar numbers of students, to constitute a fair comparison. However, any categorisation for GPA and major that would provide a reasonable amount of detail contains too few individuals and is therefore not useful for this kind of analysis. A general measure of the correlation for GPA was calculated as it is the only variable with a purely mathematical composition. For each variable a comparison was made of the performance within each category.

This analysis is deliberately broad and is looking for significant, general trends. When considering individual variables such as these it is important to consider that random fluctuations based on sampling can influence correlations and analyses. Taking a very broad strokes, a somewhat holistic view, ignores the small fluctuations and only considers the overall picture.

4.3 Results

In general, this analysis yielded minimal significant results. However, the statistically significant results do have some value.

4.3.1 Variable - Major/Campus

The analysis presented with regards to choice of major, as shown in Figure 4.1, is a very high level overview of the influence of choice of major. The categorisation of students via major requires too many categories to maintain the integrity of the analysis and so little can be inferred from the data. Additionally, the difference in number of students for each major is very significant, as shown in Table 4.1. While there does appear to be variation in performance between different major choices, the individual courses that a student will undertake over the course of their degree mean that assuming a standard knowledge base from major choice is not reflective

of student experience. While significant and clear delineations in performance between majors may have indicated a potential for a similarity of experience, this does not appear to be the case.

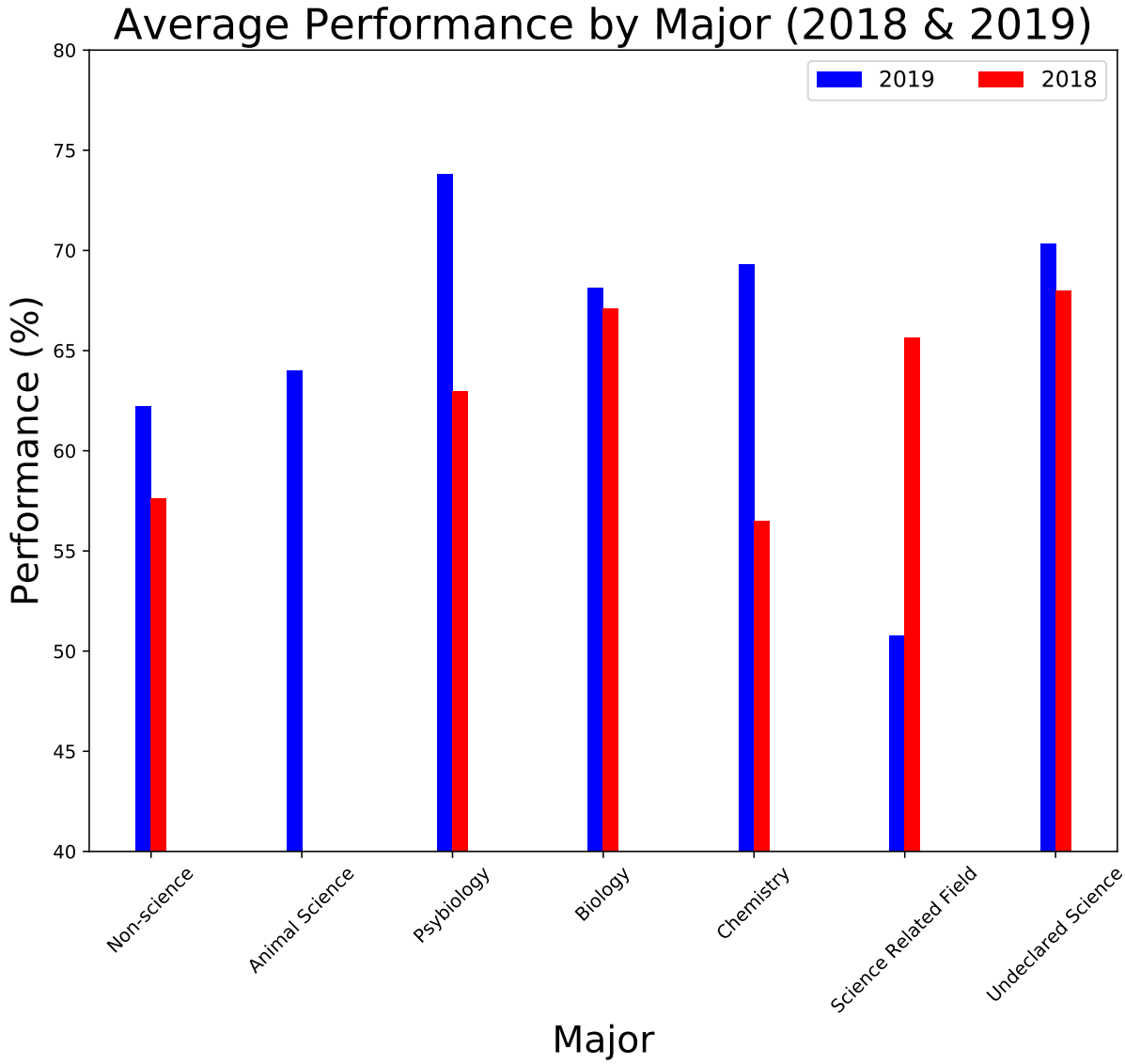


Figure 4.1: Average performance based on major choice (2018 & 2019)

The performance is similar within each major, considering the worse overall performance in 2018. There is no clear skew to a particular degree. The variation between the years is too significant to draw any strong conclusions

In a similar vein, there are some statistically significant variations in student performance based on campus, shown in Figures 4.2 and 4.3. While all of the Summer School students (barring at maximum of 2 individuals per year, who come to the Summer School as independents) are University of California students, the cohort contains individuals from each undergraduate campus. Specialities between different campuses do mean that these students come in with more varied prior knowledge than would perhaps be expected from one university. For example, the majority of students from UC Davis are animal science students as it is a world leader for veterinary medicine. The data suggests that these students tend to struggle more than the average.

Major	*2018*	*2019*
Non-science	1	5
Animal Science	N/A	10
Psybiology	2	15
Biology	57	43
Chemistry	4	6
Science Related Field	5	10
Undeclared Science	20	32

Table 4.1: Frequency of Majors (2018 & 2019)

This table shows the frequency of each major for 2018 and 2019. This data demonstrates the difficulty in drawing comparisons as the number of students in each major varies significantly both within and between years.

While there is a potential to use this as a predictor of performance, either campus or major, the reason why these students struggle is unclear. This is especially true when statistically comparing performance from all campuses. There is variation between years that is not consistent, with average performance changing by almost 10% between years in some cases. This effect can be seen in the shift of the median performance for campus 5 from 50% in 2018 to 70% in 2019 in Figures 4.2 and 4.3. Both major and campus may be measuring a secondary effect but offer little clarity with regards to understanding the actual mechanisms that affect performance. This is exacerbated by the small numbers attending from many of the different campuses. As one campus dominates it is difficult to draw any strong conclusions. Given the variation year on year

4.3.2 Variable - GPA

Both major and campus are very complex variables which intersect with other variables recorded for the cohort. GPA is measure that should be equivalent across all majors and identify if variation that is seen between the various majors is due to variation in average GPA amongst those students or another factor that has not been captured by this data.

Broadly, GPA correlates with student performance, shown in Figures 4.5 and 4.4. However, this correlation has varied year on year. For 2018 and 2019 the correlation has increased from 0.48 to 0.65, though this is still not a strong correlation. This is in line with a significant increase in performance between those two years. GPA is only a good predictor of student performance if the subject areas are similar. Good performance in English will never be a good predictor of performance in physics. However, as GPA is both a measure of knowledge within a subject area and a measure of an individual's ability to learn (and to be assessed), there will always be some correlation seen between GPA and performance in any course. Poor performance for certain majors or campuses may be more readily identified as poor learning skills as provided by those courses/institutions.

This is highlighted in the comparison of correlations between GPA and performance for

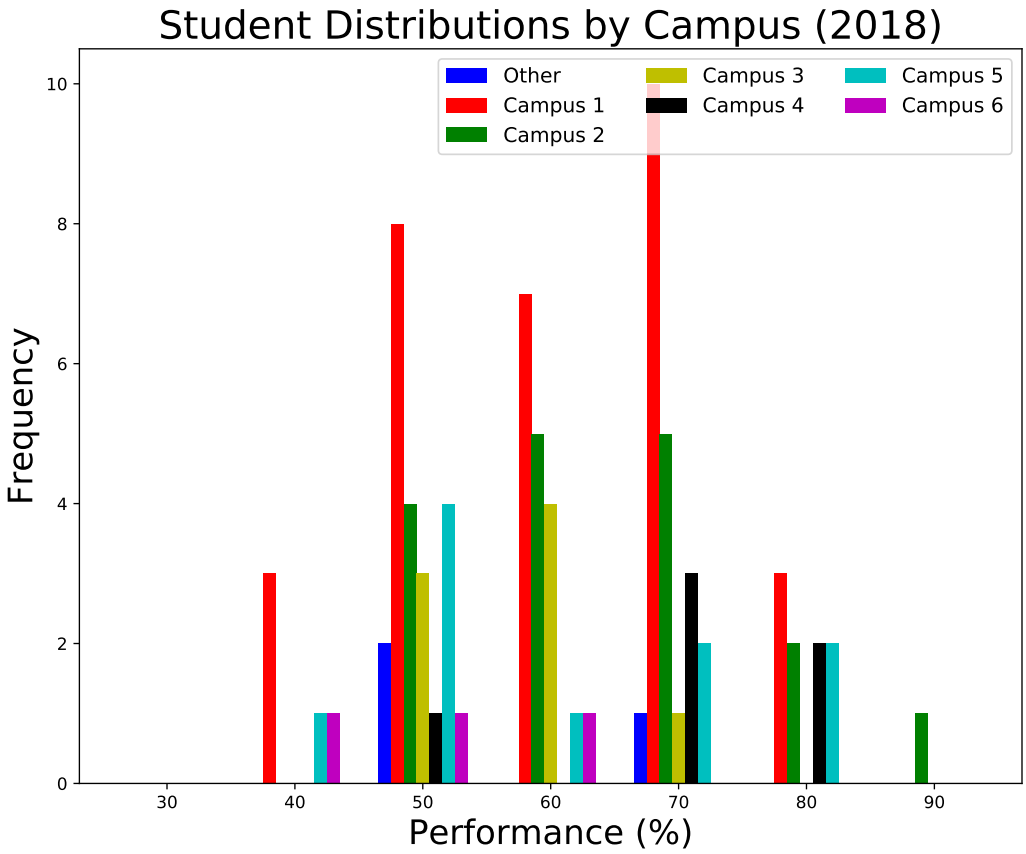


Figure 4.2: Average performance based on campus (2018)

There are significant variations between campuses. However, this variation is not consistent year on year. For 2018, there is a relatively narrow variation in performance between campuses. With the lower overall performance in 2018 there is less obvious variation between campuses. The higher the average the clearer these variations become.

the engineering students. These students tend to perform better overall and tend to come in with a higher GPA. However, the correlation between these variables is significantly lower for engineering students versus life sciences. The overall performance for the engineering students varies over a much smaller range than their GPA and suggests that GPA is measuring knowledge or skills outwith those necessary for the summer school course, and therefore is only of limited use.

The higher correlation within the LS cohort is likely down a lower barrier to entry within the LS course. There is a much higher maths skill requirement for the engineering course and is that is likely not captured within the GPA measure.

There is also potentially something to be said for the fact that the correlation is generally higher for the LS cohort when the overall performance is better. While there is something relevant being measured it is clearly not the only factor.

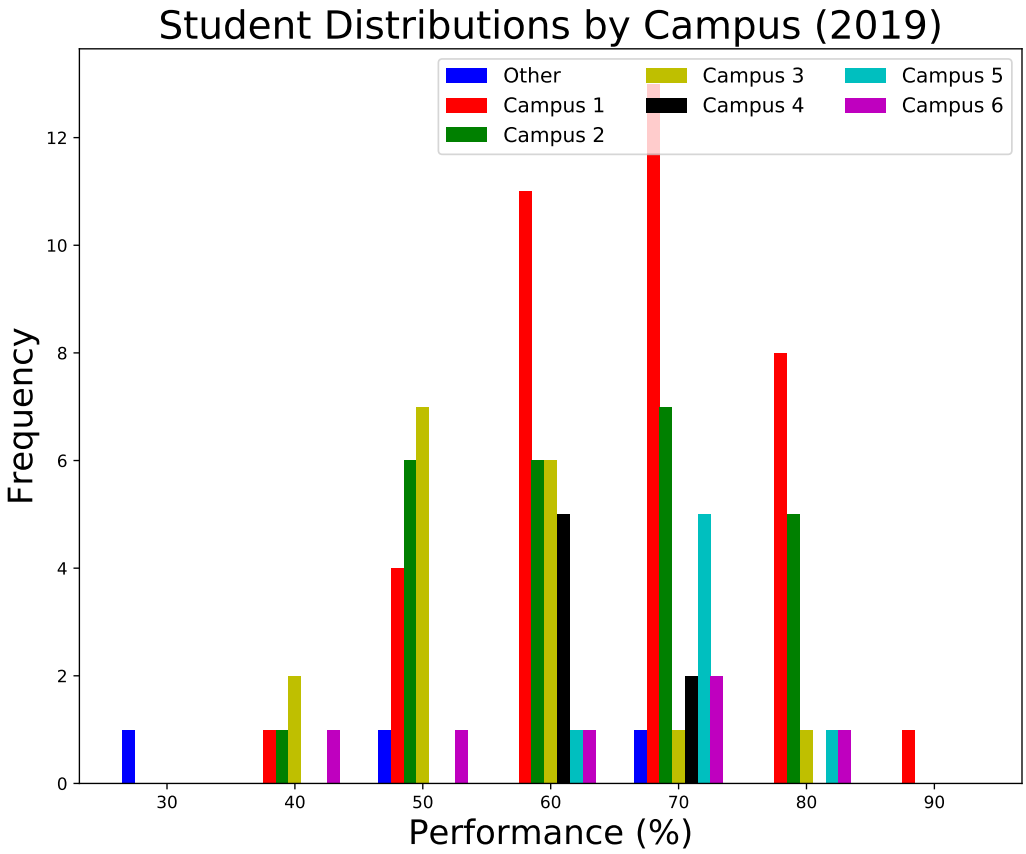


Figure 4.3: Average performance based on campus (2019)

There are significant variations between campuses. However, this variation is not consistent year on year. For 2019, there is a 20% different between the median for 2 campuses, from 60% to 80%. Variations in the way pre-requisite courses are delivered or what courses are available to students may have an impact that cannot be identified from the data produced by the Summer School.

4.3.3 Gender/Year of Study

The previous variables that have been considered are strictly academic but can be influenced by the individual. Gender and year of study are variables that are intrinsic and consistent. These variables should not show significant differences (this is the case for year of study as well as this is an introductory course with only first year mathematics as a requirement). However, there is a consistent difference in the performance between male and female students (Figure 4.6, as well as second year students consistently performing better, especially when compared to third or fourth year students (Figures 4.8 and 4.7).

Differences in the performance of students in physics based on gender has been reported time and time again [Simmons and Heckler, 2020, Seyranian et al., 2018]. These differences are often associated and correlated with differences in feelings of belonging and identity within physics [Eddy and Brownell, 2016]. The data from the summer school is not comparable with

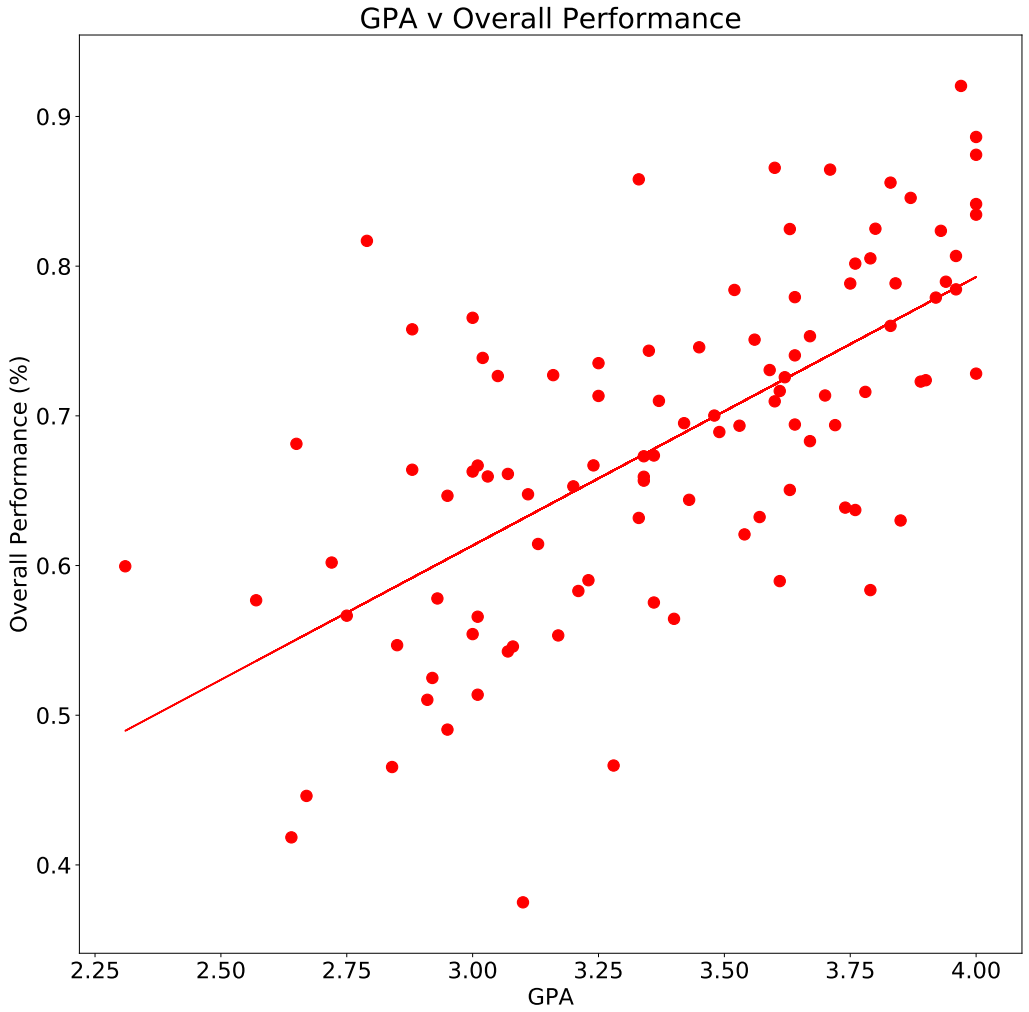


Figure 4.4: Correlation of GPA with overall performance (2019)

There is some suggestion of a correlation Life Sciences students. There is more of a spread for the lower performing students, likely deriving from the higher incoming GPA non-science students. The engineering students are not included in this analysis as they are a small group and it is difficult to derive any strong evidence of any correlation.

the vast majority of studies into gender based performance in physics as the cohort is not only predominantly non-physicists but also majority female. This cohort is unlikely to be concerned with identifying with physics or wishing to be part of the physics community, and additionally is not afflicted by a common issue raised for women in physics, that "there are no others like them" in their cohort. Though it must also be considered that the teaching staff was still majority male every year the Summer School ran. However, this cohort will be impacted by the limited uptake of physics by girls in high school [Riegle-Crumb and Moore, 2014]. The longitudinal effects of a limited uptake of physics by girls in high school is difficult to ascribe but is likely to have an impact on our summer school cohorts. There will be a lower average of prior knowledge in the cohorts as a whole and alongside that a lack of confidence. Considering this, it is perhaps

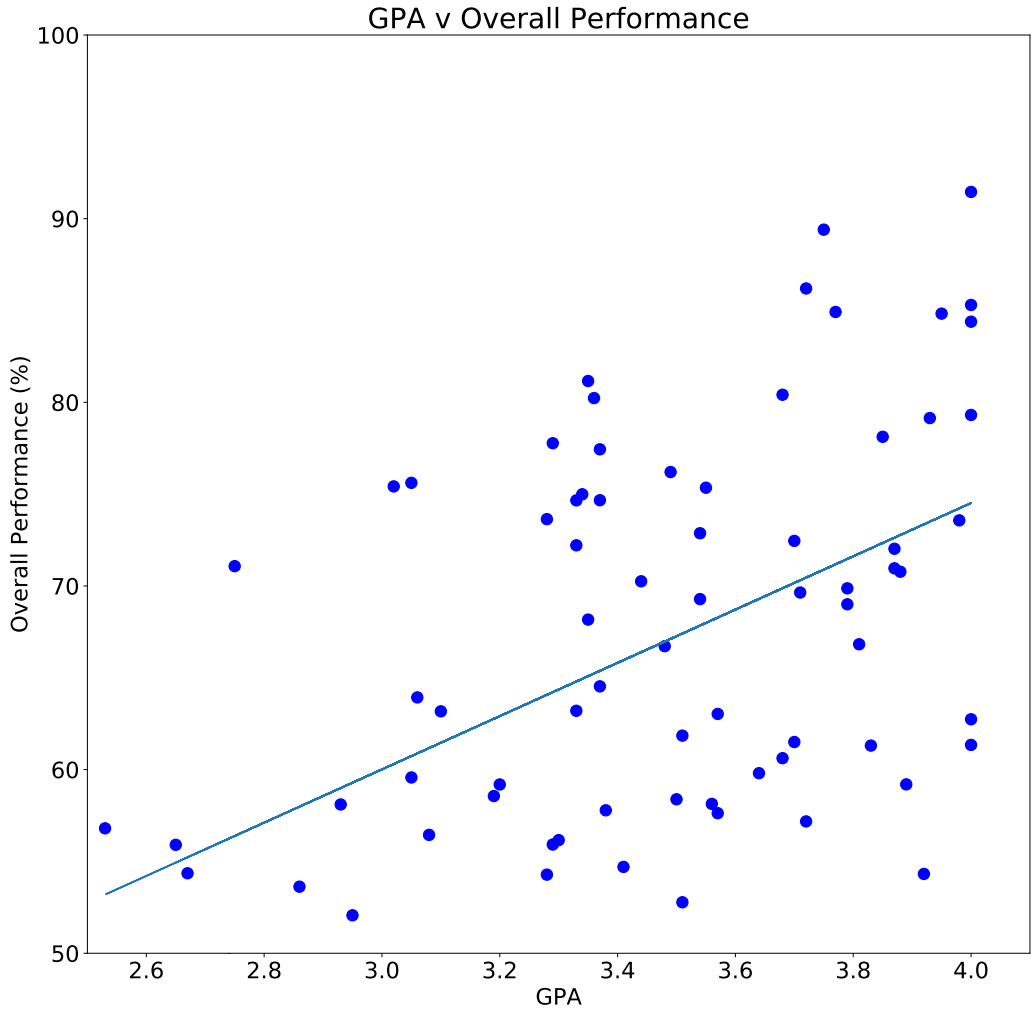


Figure 4.5: Correlation of GPA with overall performance (2018)

Performance overall was lower in 2018 and this is reflected in the much lower correlation with incoming GPA. While there is still appears to be a correlation at the lower performance and GPA, the high incoming GPA students are incredibly varied. This again may be linked to the background of the students. A high GPA in predominantly non-science subjects will not necessarily predict strong performance in the Summer School.

unsurprising that there is a performance difference based on gender. This does suggest that, rather than gender, it is the prior experience that is blame. Within this context though, it is impossible to fully distinguish the root cause of the gender disparity.

Similarly, it is difficult to identify to mechanism that causes such a distinct difference in the performance between difference years of study. There is a consistent difference between 1st and 2nd year students when compared to 3rd and 4th year students. It is potentially a case of motivation. It is clear that motivation has an impact on how well students perform [Steinmayr et al., 2019, Kusurkar et al., 2013] but understanding the mechanisms of motivation are still unclear. Anecdotally, within the summer school, earlier year students are generally more motivated and

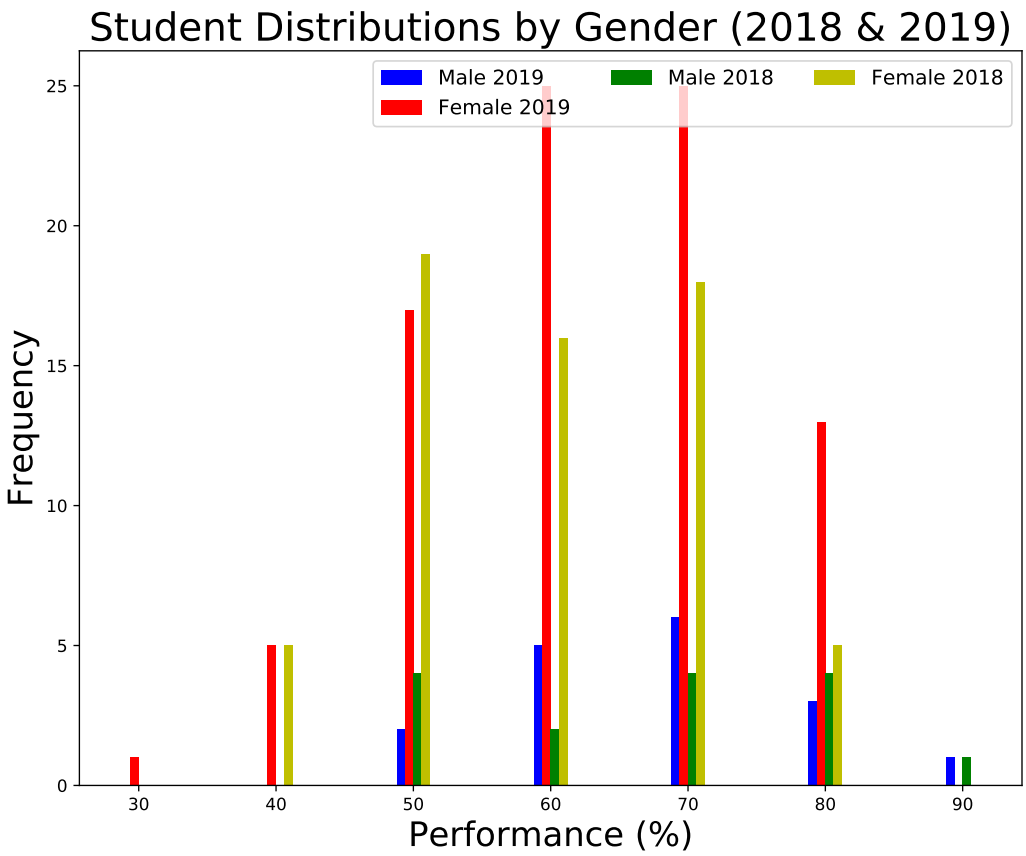


Figure 4.6: Distribution of performance based on gender (2018 & 2019)

The most clear variation between male and female students is the peak and tail of the distributions. There is a not insignificant tail to the distribution for female students that does not exist for male students.

engaged, as evidenced by their enthusiasm within the course but also by their mere presence on the course. If a 1st year student is attending the summer school they are likely planning to graduate ahead of schedule and so want to complete their physics requirement as soon as possible. This demonstrates a level of dedicate and studiousness that is lacking in students in later years. While it is difficult to fully quantify the influence of these factors it is an area that could be explored in future study. This is an area that could be explored most fully in a qualitative way, utilising focus groups and more in depth conversations with the cohort.

There is an additional variable within year of study which may also have an influence. In 2019, data was also provided with regards to the prior experience of the cohort in physics. This was defined by completion of AP Physics. This data is provided in Table 4.2. There does appear to be a small effect, seen most clearly in the result for year 2. This may be the most significant difference as this is by far the largest group within the cohort and so it is easier to seen any impact. The effect of the small sample size may explain why the average for year 1 students with prior experience in physics is lower. Prior experience in any subject should

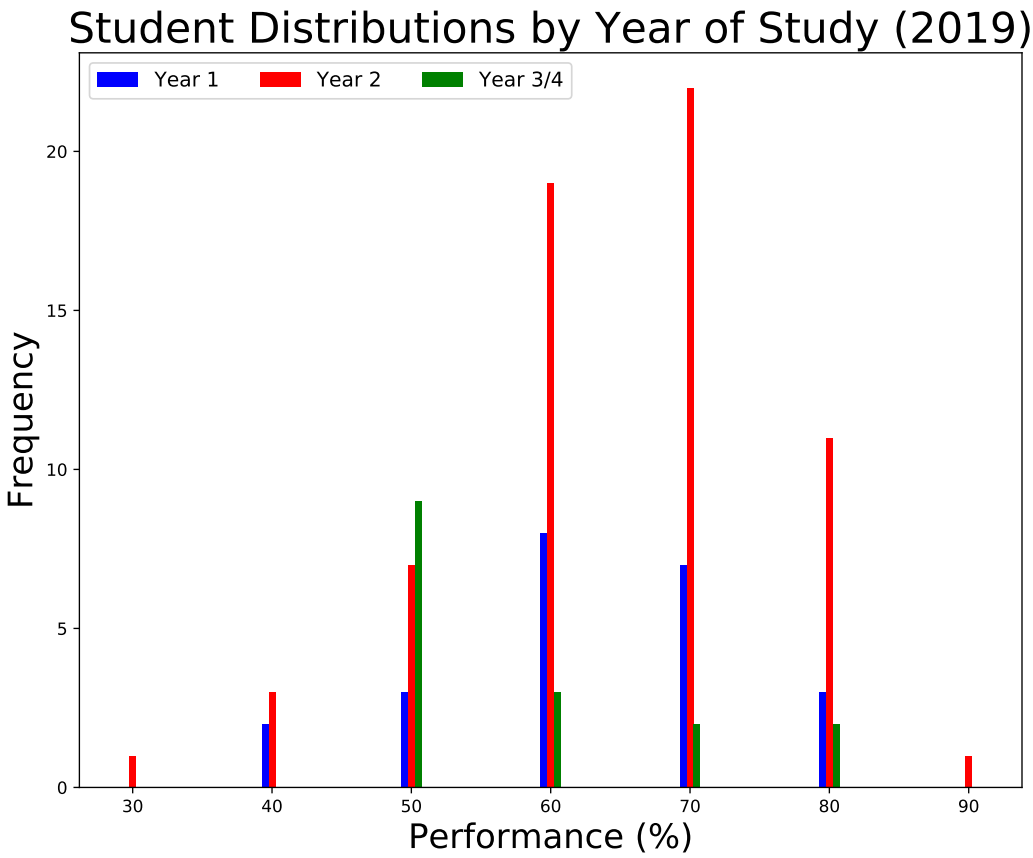


Figure 4.7: Distribution of performance based on year of study (2019)

The median performance for 3rd/4th year students is clearly and significantly lower than that for 2nd year students. 1st year students also tend to perform better than 3rd/4th year students. Though it must be noted that the 2nd year students make up the majority of the cohort and as such does present a more spread distribution. This data is just to illustrate overall trends. A very similar trend is also seen in 2018.

translate to better performance, however, as discussed earlier in this chapter, student motivation and engagement can have a massive impact on performance as well. The data provided did not include any detail on how well students performed in their AP class. Considering this data in conjunction with student performance on the FCI may provide a more detailed understanding of students' prior knowledge. The 8.2% difference for year 2 students does indicate that this is a potential area for further research.

4.4 Discussion

All of these significant differences can be linked to two categories, GPA and prior experience. While GPA is not a categorical predictor of student performance within the summer school it does have a correlation. While the vast majority of students on this course have not completed

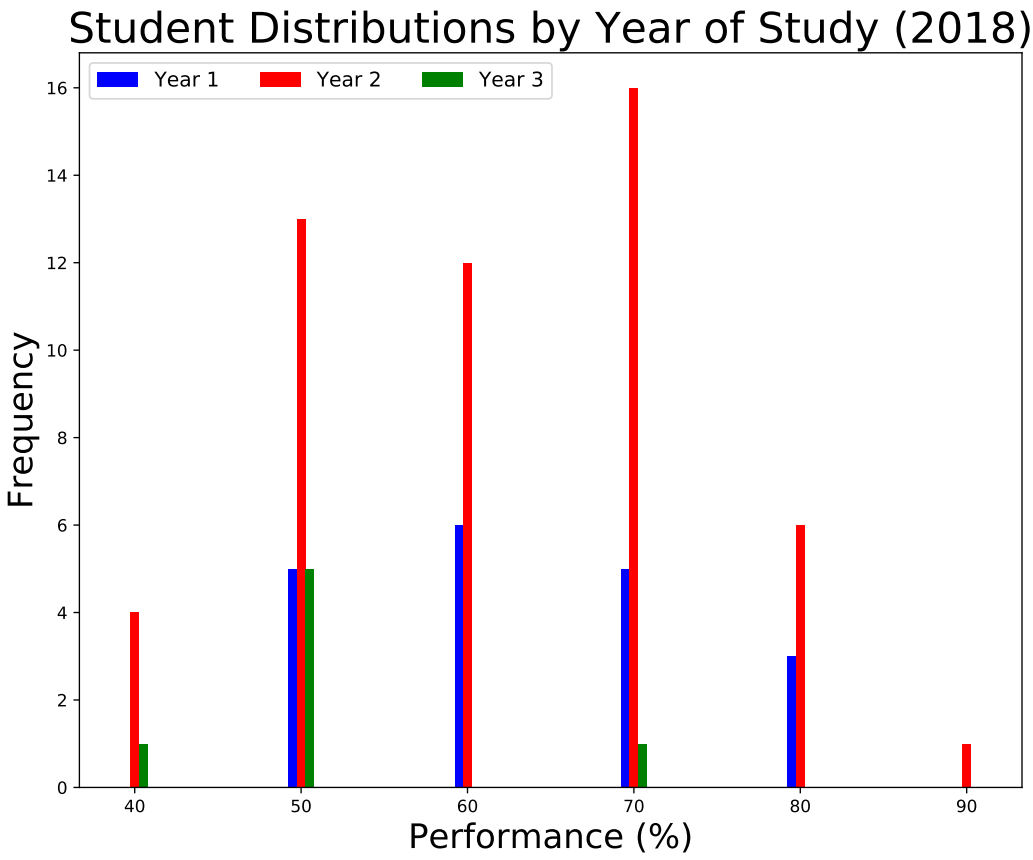


Figure 4.8: Distribution of performance based on year of study (2018)

The median performance for 3rd year students is clearly and significantly lower than that for 2nd year students. A similar trend also seen in 2019.

Year	*No Prior Experience*	*Prior Experience*
Year 1	70.4%	68.5%
Year 2	64.5%	72.7%
Year3 & 4	60.7%	64.4%

Table 4.2: Impact of Prior Experience on Performance

This table compares performance of life sciences students when considering prior experience. There may be a small impact due to prior experience in physics, however, with limited information with regards to the content of the prior experience it is difficult to ascribe any significance to the results.

any physics based courses at university, GPA can still identify students with a strong ability to learn, something that is transferable to any course. Additionally, for 2019 information regarding students' experience of physics in high school (defined by having completed the AP physics course) was included. Taking this prior experience into account could potentially explain statistically significant results that are not directly linked to GPA.

There are a variety of different approaches to understanding what influences student per-

formance, but this data suggests that the impact of metacognition and the ability to integrate that understanding of learning with prior knowledge lowering the barrier to entry are very significant. Considering this in the context of the analysis from Chapter 2, it is clear that while an understanding of the material is there, the ability to demonstrate that can be impeded by a lack of understanding in other areas, be that related academic areas (i.e. mathematics) or in a broader metacognitive sense such as correct application of knowledge. It is also important to consider the impact of a shared vocabulary that can arise from prior experience. Application of knowledge is much easier when the only impediment is the situation in which the knowledge is to be applied. Adding the additional barrier of unfamiliar language can be especially taxing as it amplifies all other barriers to performance. Additionally, the ability to adapt understanding to new situations and make connection between different areas of knowledge is a key aspect linked to application of knowledge. The ability to build knowledge as a cohesive unit rather than as discrete knowledge centres is why measures such a GPA still have such a strong correlation with performance.

Chapter 5

Student Group Analysis

5.1 Introduction

Given that so much of the student learning during the summer school takes place within the group setting it is key to understand whether there are group dynamics at play which cause the variation in performance between groups. Analysing group dynamics has always been under the purview of qualitative analyses. Group learning is a high dimensional system and qualitative data can create a high fidelity facsimile, maintaining the large amounts of information. However, this comes at the cost of limited scope and time intensive data collection. This data collection can also be influenced by outside factors such as individual and sampling bias [Barker, 1980, Cotton et al., 2010]. As discussed in the previous chapter the individual characteristics of the students - be they academic or personally intrinsic - can have a significant impact on the performance of the student, and therefore the group [Chatman et al., 2008]. Because of these impacts, identifying the dynamics within a group can be difficult, especially from a quantitative perspective as within such high complex systems as student cohorts. Specifically isolating the variable which is the cause of the effect can be difficult.

There is also little evidence of what specific group dynamics are linked to academic success, or even always a clear consensus on what is meant by a group [Lorge et al., 1958]. A group that is working well together within an educational context will be highly discursive and encourage the asking of questions [Rusk and Rønning, 2020, Lindblom-Ylänne et al., 2003]. However, this measure is purely about the success of the group "as a group", it has no inherent link to the academic success of the group. Many analyses of groups do not consider the output of the group as a significant variable, focusing more on the experience of individual members of the group [Gapp and Fisher, 2012]. While this is a necessary avenue to explore, it does not show whether group work has a positive impact of individual success. This may be due to the issue of comparison, as there is no way to accurately measure the impact on an individual's performance working both as an individual and as part of a group [Lamm and Trommsdorff, 1973]. This is a limitation that cannot be avoided, and so an assumption that group work has the potential to

have a beneficial impact on individual performance has been taken.

The aim of this project was to create a quantitative measure of group dynamics. While there may be a loss of fidelity the advantages of quantitative analyses, namely, easier and more thorough data collection, more broadly applicable results and more direct comparisons with other variables, outweigh the potential limiting of accuracy.

The initial hypothesis made some assumptions to allow the use of a quantitative approach. This is partially due to the use of correlation analysis as the basis for the quantitative approach. This was used as it is potentially analogous to the cohesion of a group working well together. A group that is working well will slowly integrate over time and perform more similarly, becoming more cohesive, with the mixing of ideas and approaches. Therefore, this hypothesis assumes that the correlation of academic results will reflect positive group dynamics. It is important to highlight that measure of success being used within this project is positive group dynamics and cohesion, and not overall academic success. A cohesive group may not contain the strongest students but will support those student more and allow them to perform better than they would have otherwise.

To understand if this quantitative measure is accurate as a description of group dynamics a comparison was made with observation These are variables which can be measured through observation. While this is a relatively limited scope, it allows for higher levels of accuracy for the variable measurements during the observations.

5.2 Observation Experiment

5.2.1 Introduction

Observations were made during group work sessions (tutorials). The questions were provided in advance and students were expected to attempt before the session - aligning with the flipped classroom structure of the summer school. The aim of the sessions was to give students the opportunity to discuss issues with their group. Weaker students would get targeted support (ideally from other students in the group) as they would have specific questions and stronger students had the opportunity to solidify their understanding by supporting the weaker students. The mentor is there primarily to facilitate the group discussions and only step in when the students are not making any significant progress towards the solutions. Ideally the mentor is there as a scaffold for the students to then create a self sufficient group.

A self sufficient group is highly discursive and functions as one unit - discussions involved the whole group and do not allow for wildly different levels of understanding. For this to be possible the members of the group cannot have a vastly different base line understanding, while still maintaining a diversity of approaches. A group that is working well together should see the weaker students improve - at least in terms of ranking - and a narrowing of the performance gap

within the group.

This is something that can be measured but identifying if it is determined by the group dynamics requires a specific measure of the group. In this case the groups were observed over the course of weeks 2 to 6 of the summer school for 20-25 minute intervals during the tutorials. These observations were then compared with some quantitative measures of group consistency. However, to be clear, these comparisons were not done to look at whether groups that are working well together are more academically successful, but rather to understanding if quantitative measures of group cohesion can identify the same overall trends as qualitative observations.

5.2.2 Method

The observations made during the tutorial sessions were looking at the nature of the discussions within the groups. To homogenise the results from the observations, the data that was collected focused on the use of questions within the groups discussions. The results collected identified whether the mentor or the students (divided into LS and Eng) were asking or answering the questions. An ideal situation would find the students asking and answer all questions - with the mentor taking a back seat through most of the discussion. Broader notes were also taken to clarify the nature of the discussion within the groups and provide context for numerical results gathered.

These observations were then compared to quantitative analysis of the group performance. A high correlation between the observation data and the data representing the performance gap within the groups would indicate that the two sets of data are measuring the same variables and can be used in conjunction with each other. The observations made during this experiment did not cover the entire time the students were in their group sessions and do not necessarily provide a full picture of the group dynamics. Ideally quantitative data can fill these gaps - but only if it can be shown that the measurements are of the same variables.

The quantitative analysis was deliberately used in a broad strokes manner. The aim was to identify if the observations were consistent with quantitative measures of group performance and diversity, rather than linking group performance to any academic consequence. If a broad correlation can be found between the qualitative observations and the quantitative data, then both methodologies can be used in conjunction.

Converting the qualitative results into data that could be usefully compared to any quantitative measures was done in two ways: compiling the question answering data and categorising the written notes from each observation. While the question data is more robust (there is little observer interpretation that influenced the gathering of that data), it cannot necessarily capture the nature of a free-ranging discussion in all contexts. However, the observation notes had the capacity to take into account variations in group behaviour that would not be picked up in the other data collection process. To categorise these notes, each individual statement within the observations was coded as either a positive or negative attribute of the group. An example of

positive observations would be free-form discussions within the group, especially if the group mentor is sitting back and allowing the students to lead. A negative example would be the mentor taking a "lecturing" role, or an individual student dominating. A ratio was then found for each group based on this coding. As each group was generally only observed once or twice in a week this ratio is for the summer school as a whole, and does not allow for a comparison from week to week.

5.2.3 Results

The results from the observations have been separated into the purely quantitative and purely qualitative data. The quantitative data is the question asking and answering data shown in Table 5.1. The qualitative data is the summary of the group interactions from the observations 5.2. During the data collection process it became clear that quantifying the discussions during the observations (measuring who was asking and answering questions) was not capturing enough information. When a group was successfully working together the conversations were more free ranging and were not bound by a question and answer format. As can be seen in Table 5.1 there are some groups with no questions. While this result could suggest that those groups were working poorly considering all data available it is more likely that this measure is too restrictive and does not capture the full picture.

The qualitative data collected during the observation process was much more fruitful however, and gave a much more nuanced view of the group dynamics at play. So as to compare this data with potential quantitative measures, the individual comments were considered data points and categorised as positive or negative. This was then represented as a ratio of positive to negative comments and is shown in column 3 of Table 5.2. Only 5 of 12 groups had a ratio of 1 or greater. This is likely due to negative aspects of the group dynamic standing out more during observation. However, since the impact of the observer was consistent between all groups the data is still representative of the ranking of the groups by group dynamics.

As the observation data is the only direct measure of the groups dynamics, if there is any correlation with performance then it should be seen with this data. As has been used throughout this manuscript, the correlation was calculated a shown below. This calculation is used only to find a possible relationship, and would not indicate the strength of any relationship found between positive group dynamics and performance in assessment.

$$Correlation(X,Y) = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 (y-\bar{y})^2}} \tag{5.1}$$

A comparison of the observation data with the average group performance in Table 5.4 shows little correlation and suggests that a purely qualitative measure of group dynamics is not a strong predictor of performance. This also provides a framework through which to understand subsequent correlation calculations, that to justify the use of a quantitative measure of group dynamics

does not need to demonstrate a relationship with performance in assessment.

5.3 Design - Correlation Analysis

The initial design for a quantitative method of measuring group dynamics considered that a group which has strong positive dynamics will work to improve the performance of the lowest ranked students and will maintain the high performance of the highest ranked students. Working as part of a group allows strong students to gain a deeper understanding of a topic through presenting ideas to the rest of the group and provides an opportunity for the weaker students to approach ideas from a new perspective. Therefore a group with strong positive dynamics should become more cohesive over the course of the summer school. To measure cohesion, the overall correlation of the group performance was measured.

A program was written to calculate the correlation coefficient for each student as compared to all students within their group, using the summative assessment results (i.e. class tests and exams). A value was calculated for each week of the summer school to highlight changes within the group. The hypothesis suggests that groups with positive dynamics should have an increasing average correlation coefficient. The assessment results were normalised to remove week to week variations based on material difficulty.

Both the idea of correlation being indicative of positive group dynamics and whether this is a useful measure when considering student performance were tested with comparisons to the observation data and overall group performance.

5.3.1 Results

The coefficient values vary quite significantly, especially in module 1 as there are fewer data available for the analysis. By the end of the summer school all correlation values are below 25%. Considering the groups contain 7-10 students (the engineering students are not included in this analysis as they complete different assessment), a low correlation is unsurprising. Another consideration is whether there is a positive or negative trend in the correlations. This varies from group to group and there is no consistent behaviour.

One of the assumptions previously highlighted was that correlation would not necessarily follow the overall academic performance and has been demonstrated in Figure 5.1. Additionally, Figure 5.1 demonstrates the vast inconsistency in results from the analysis.

Groups are high dimensional systems and such inconsistency may be expected. To test if this analysis is identifying similar metrics to qualitative approaches of group dynamics, a comparison was made with data gathered through observation.

A comparison was then made between the average performance, the observation ratio, and the final correlation values. These values were calculated for each group and a correlation calculated between each measurement. As this was an exploration of the potential use of internal

			Asking Questions			Answering Questions		
Group	**No. of Engineers**	**Time**	**Life Sciences**	**Engineer**	**Mentor**	**Life Sciences**	**Engineer**	**Mentor**
1a	1	20 mins	2	1	2	2	0	3
1b	0	17 mins	0	1	1	0	0	
2a	0	20 mins	4	N/A	1	4	N/A	1
2b	3	20 mins	0	0	2	2	0	0
3a								
3b	2	13 mins	5	0	1	6	0	0
4a								
4b	2	20 mins	6	0	4	7	1	2
5a	0	20 mins	6	N/A	10	9	N/A	7
5b	2	20 mins	6	1	3	5	1	4
6a	3	20 mins	0	0	0	0	0	0
6b	3	20 mins	8	3	1	2	5	5

Table 5.1: Example of data collected during observations

This is an example of quantitative data collected during observations of group interactions. This quantitative data did not garner useful results as gathering these results consistently was difficult given the nature of conversations within some groups. Groups that focused on more structured discussion provided clear opportunities to gather data. Groups that tended towards multiple small group discussions were more difficult to observe accurately.

Group	Notes (Example from one session)	Numerical Representation (total of all sessions)
1a	Eng involved in summary. No discussion during summary but engagement. Most work done in small groups - but since group is engaging and talkative this could be improved.	0.66
1b	New technique for group does not encourage discussion	0.33
2a	Overall easy tutorial and evident in mentor student interactions. Everyone understands what's happening	2.50
2b	Lots of discussion in group but not really question based - more back and forth	1.55
3a	Maths revision from previous questions. Lots of discussion in group without mentor but impossible to follow in terms of data collection.	0.50
3b	Only one eng and other eng is ill - student on board very strong and leading tutorial rather than mentor. They ask questions to check understanding - like the mentor should be	0.50
4a	Eng explaining to rest of group - very competent. Mentor needs to take more of a back seat - get students to ask each other rather than through the mentor. More dictation than discussion.	0.33
4b	One eng missing - weaker eng still there. Good discussion in that students are asking questions - weaker students still too quiet. Good use of the textbook	1.40
5a	Lots of questions from mentor - v. quiet students. Explanation of the background on board from mentor. Lots of targeting students with specific questions. More focused on the board vs 5b.	1.00
5b	Discussion within group not necessarily about question. Somewhat distracting. Summary from student on board - not a lot of engagement from group. V. low energy. One eng cause of much of the disruption. Still progressing even with the distraction. Other eng never speaks. Q5 by 11 is too fast.	0.37
6a	Students working in groups on Q6 tutorial already done by the time I got there	2.00
6b	Eng on board for Q. Long set up. Group just copying the answer from board. Eng explaining on board and students and students engaging quite well. Discussion good across groups - questions finished on board - broken into smaller group discussions. Mentor challenged students after answers to make them think. Longer discussion after question as eng challenging mentor when they shouldn't	0.90

Table 5.2: Example of notes from observation data

These are an example of notes taken during observations of the group work sessions. These notes were coded as positive or negative to produce a numerical representation of the data, shown in column 3.

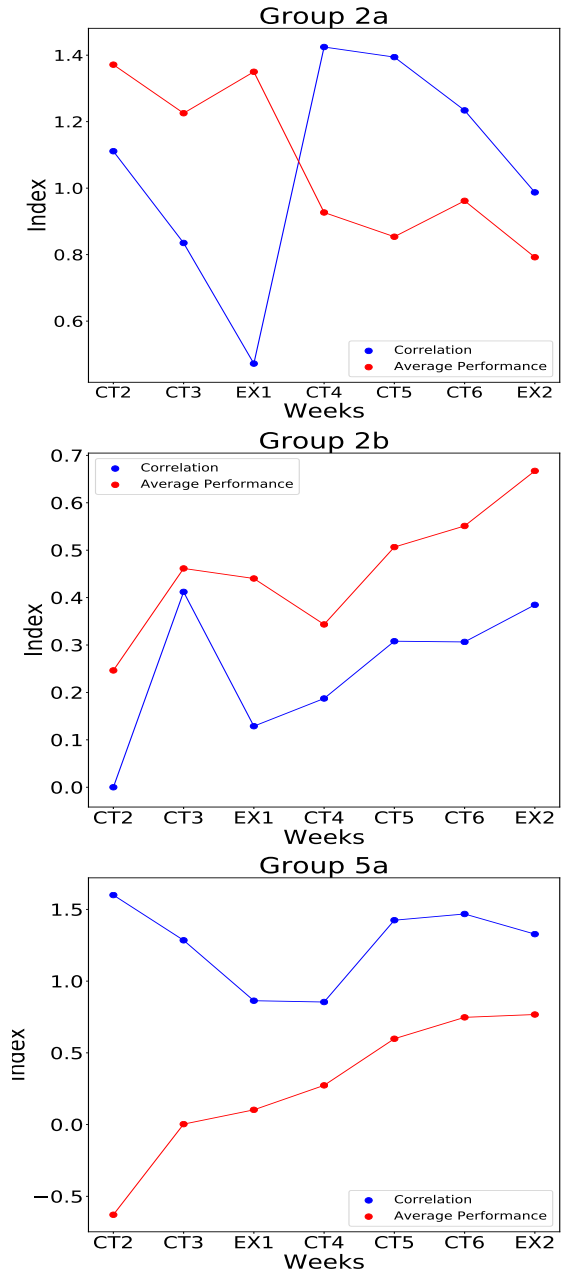


Figure 5.1: Example of correlation and assessment comparisons for 3 groups (2019)

To compare the trends the correlation has been scaled up by 10. While there may be individual groups where the trends in correlation and performance match, there is no consistency. Additionally, with the requirement to scale up the correlation values these fluctuations are exaggerated in this figure.

group correlation as a measure of group cohesion, a full programme view of the data was taken, rather than a more granular apporach. As such, the overall average performance across the full programme, the observation ratio considering all observation data, and the internal group cor-

relations of the group by the end of the programme were used respectively. The correlation, following the same process as before, between each of these values is poor and suggests that the overall correlation analysis does not identify the same variables as the observations.

This is potentially due to limitations of correlation analyses. As the groups are fairly large, 7-10 students, and were designed to be as diverse as possible, the final correlations are very poor. Small values mean that the inherent noise within the system will have a large impact on the final results and allow for limited differentiation between the groups.

5.3.2 New Approach - Standard Deviation

The standard deviation was considered as an alternative to the correlation. It can be used in a similar manner as it also considers how cohesive the group performance is, but has less stringent constraints. A value can also be calculated week by week, like the correlation values. Additionally, this standard deviation considered the ranking of the students within the class rather than absolute assessment results. Not only does this normalise the results, it provides a better way of differentiating the students as the ranking provides a broader categorisation than the standard assessment. For a standard undergraduate physics cohort the use of ranking may not be necessary as there is a broader distribution of grades. Due to the highly selective nature of the summer school, the variation between students is limited within the assessment. The use of ranking creates a greater delineation between the students.

5.3.3 Results

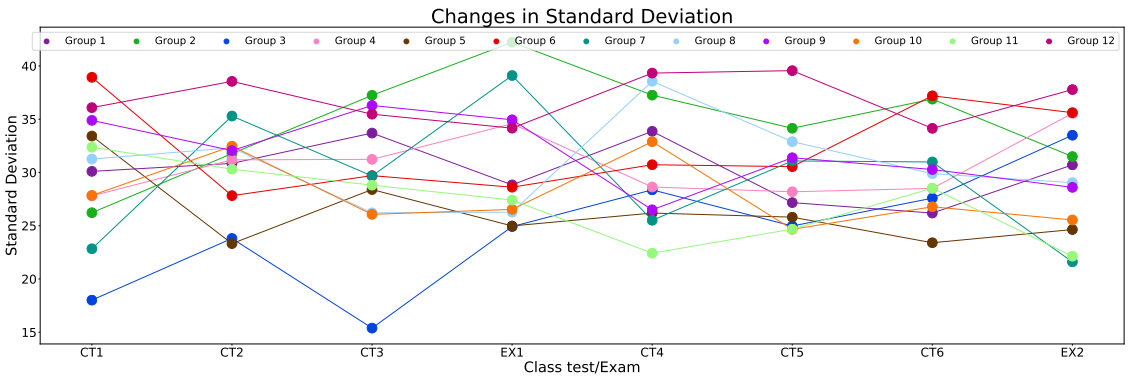


Figure 5.2: Standard Deviation of all groups throughout the summer school

There are no overall trends in the Standard Deviations within or between groups. The range of Standard Deviations are consistent from week to week.

The correlation between standard deviation and observations, while not strong, is positive and does support the hypothesis that a quantitative measure of group dynamics is possible, but that using the group correlations was too restrictive, seen in Table 5.4.

	Observations
Average Correlation (Ranking)	-70.6%
Average Standard Deviation (Ranking)	52.1%

Table 5.3: Summary of Correlations of All Measurements

	Average Performance
Average Correlation (Ranking)	19.2%
Average Standard Deviation (Ranking)	17.6%
Observations	19.8%

Table 5.4: Summary of Correlations with Average Performance

The measures of group cohesion in Table 5.3 have two very different correlations. The correlation of the standard deviation with the observation is the only positive and significant correlation found. Table 5.4 shows that the measures of group dynamics, be they qualitative or quantitative do not appear to have a strong relationship with the group's overall performance.

Figure 5.2, unlike Figure 5.1, does not have wildly different trends for each group, again suggesting that the use of full group correlation requires impossibly high similarities in student performance.

5.4 Discussion

It has been shown here that a quantitative approach to analysing group dynamics is possible. The use of measures such as standard deviation, that measure consistency within a group, can potentially be considered a numerical representation of cohesion within a group. How that data is interpreted, though, is not clear. There is still a need for further analysis to understand whether there is a particular aspect of the interaction that is captured within the standard deviation measure.

As the set up of these groups was specifically designed to create diverse groups it is perhaps unsurprising that the use of correlation analysis within the group was unsuccessful. The relative success of the standard deviation as a measure does suggest that if the groups were designed to be more similar initially then the correlation analysis may be able to identify significant differences.

While each of these approaches can potentially shed light on student dynamics, there is still no evidence linking positive group dynamics to overall performance. While this data cannot say categorically that group work has a positive impact on student performance in a general sense, it does show that quantitative data may have the power to answer this question. A combined approach is necessary to find a definite solution because, as has been shown here, the integration of the quantitative with the qualitative is often a difficult process.

Chapter 6

Does teaching students from multiple streams in the same groups improve outcomes?

6.1 Introduction

Teaching in higher education has an increasing focus on group work and encouraging collaboration within the student cohort. Group work has well documented benefits for students in terms of academic success and higher levels of motivation and engagement [Springer and Stanne, 1999, Kusurkar et al., 2013]. It is, therefore, important when implementing group work to consider how the groups are constructed and what impact this may have on the performance of the group as a whole or the individuals within the group. One of the most important aspects of the makeup of a group is the diversity within those groups, in particular cognitive diversity. Diversity of background and therefore diversity of knowledge leads to cognitive diversity [Curşeu and Pluut, 2013]. Students with different educational backgrounds will have not only different knowledge bases but also use different approaches to tackle problems. This is particularly evident in the cohorts of non-honours or introductory level courses [Murata, 2013].

Currently there is no strong consensus on the impact of diversity on group performance [Warner et al., 2012]. Partially due to variability in the definition of diversity and cohesion, [Carron and Brawley, 2012] there is also an issue with the definition of a group versus a team, or the differences between levels of education or groups within industry [Haughton, 2009]. While there is some support from qualitative data to suggest that diversity within group is positive, there is no strong quantitative measure of the impact. High levels of diversity provide a variety of points of view, allowing students to approach the problem from various angles and find the one most suited to their way of thinking. However, in a group setting this can create a lack of unity and impede cohesion, particularly if not all students understand all points of view. Much of the research in this area relies on individual self reporting [Roth et al., 2010], which can often

result in survey fatigue during a standard course.

It is also necessary that students are interacting in a measurable way. While it is easy to monitor the interactions of students in a primary or secondary education setting, this is much harder in tertiary education. Student have much more freedom and so to analyse student interactions the learning environment must ensure that students are interacting in such a way as their understanding can be measured. It is therefore important that the group work is structured, with a focus on peer instruction or mentoring, if a quantitative measurement is to be successful. The structure of classroom learning, especially in a primary education setting is complex, but consistent. The same students are in the same classroom environment all day and all year. An observer can identify small, incremental changes more clearly as there is more opportunity to observe. Additionally, with the majority of the learning occurring in the classroom, the control over how the students engage with the learning is much higher. This control is then lost as students progress through the stages of education, where finally in higher education, most of the learning is done in a self-directed way, outside of contact with staff.

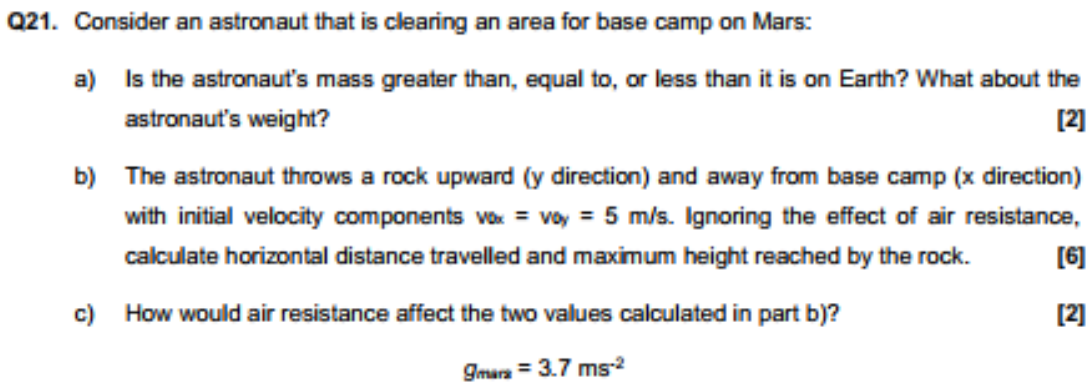
Here is presented a fully quantitative measure of the impact of cognitive and knowledge diversity on the performance of groups.

The formal assessment methods for the summer school are class tests and exams consisting of a mix of multiple choice questions (MCQs) and short answer questions (SAQs). There is also a small continual assessment component of the course which will not be discussed here. MCQs are well suited for testing conceptual understanding but were excluded from the analysis as the time spent focused on MCQ style questions was during the full class sessions. These sessions were not completed as part of the tutorial group, with students interacting with in various groups throughout all of the sessions. In contrast, the questions used during the small group tutorials were designed to emulate the style and difficultly of the SAQs in the formal assessment. As the tutorial questions are always completed only during the group sessions, differences seen in the performance of the SAQs can be ascribed directly to the small group tutorials.

The small groups, through which most of the learning is done, are either purely life science stream students or are a mix of life science and engineering stream students. Therefore a comparison can be made between the performance of these two sets of groups, characterised as low diversity (life science only) and high diversity (life science and engineering mix). Variations were limited between the groups as they were normalised based on gender balance, home university campus and GPA. Some consideration was also given to the major and year of study.

6.2 Method

To assess the impact of high diversity on the performance of a group a full statistical analysis of student performance in SAQs was conducted. Over the course of the summer school the students completed 6 class tests and 2 exams, each consisting of 20 MCQs and 3 SAQs of



Q21. Consider an astronaut that is clearing an area for base camp on Mars:

a) Is the astronaut's mass greater than, equal to, or less than it is on Earth? What about the astronaut's weight? **[2]**

b) The astronaut throws a rock upward (y direction) and away from base camp (x direction) with initial velocity components $v_{0x} = v_{0y} = 5$ m/s. Ignoring the effect of air resistance, calculate horizontal distance travelled and maximum height reached by the rock. **[6]**

c) How would air resistance affect the two values calculated in part b)? **[2]**

$g_{mars} = 3.7\ ms^{-2}$

Figure 6.1: Example of SAQ

This is an example question taken from Class Test 1. All questions are broken down into subsections, which do not necessarily all fall into the same category of simple or complex. For example, part a) would be considered simple as it is a basic application of the relationship between mass and gravitational pull. However, part c) would be considered complex as it requires more evaluation of the various components of Newtonian dynamics to understand where air resistance is relevant.

which students choose 2. An example of a standard SAQ is shown in figure 6.1. To provide a true comparison, only students on the life science stream were included in the analysis as the assessment questions for the engineering stream were not identical to those completed by the life science students. The comparison made was of the performance of life science students in a group either with engineering stream students or without engineering stream students.

Each question was analysed not as a whole but broken down into individual subsections. This allowed for a more in depth analysis of where exactly any differences in performance were coming from as a question could include many different styles of question in each subsection. For each part the average performance was taken for students in engineering and non-engineering groups. The averages for each subsection were then normalised based on the overall average for each paper, to eliminate variation from week to week based on the difficultly of the material.

Each subsection was also divided into either "simple" or "complex" questions. "Simple" was defined, as in Blooms' taxonomy, as remembering, understanding or simple apply questions. "Complex" was similarly defined as deeper apply, analyse and evaluate questions. The final type of question from blooms' taxonomy is "creating" which was disregarded as the design of the summer school puts the focus on conceptual understanding rather than complex application or problem solving. This subdividing was done to understand where the largest impact was seen as it was expected that the influence of diversity would not affect all questions equally. If a positive impact due to high levels of diversity was to be seen then this would be more evident in questions with a higher order of complexity as students will benefit from the improved discussion as they can approach the concept from multiple points of view. If a negative impact was to be seen this would be more evident in the simple questions as students would be struggling with basic concepts as the group dynamic would be interrupting the learning process.

The process of learning requires an element of failure. A student needs to be able to identify what they do not know to be able to develop their understanding. If students do not feel able to be open to asking questions and showing ignorance of a topic, then they will not be able to learn most effectively. A group with high levels of diversity will potentially have more active discussion as students approach the topic from different perspectives. This allows for an environment of asking questions, and an environment that shows that asking questions is not an inherent sign of a lack of understanding, but just of different perspective. Based on observations in tutorial settings, a group with lower levels of diversity may encourage a student to not open up in discussion for fear of seeming different to others in the group, especially if they believe that, as they have the same background as the students around them, they *should* understand a concept already. Shame and embarrassment can have negative impacts on student outcomes [Bynum et al., 2021].

All analysis was conducted using the Mann-Whitney-U test. This is a non-parametric analytical method often used in educational research settings, exemplified in Wieman's laboratory focused research [Wieman, 2015]. The test is designed to identify variations in distributions using a ranked sum method. This is well suited to analysis on test scores as the ranking also functions as a normalisation process. This test also does not require data sets of similar sizes like many analytical tests, which is particularly suited to this project as group sizes can vary. For this project the standard significance values were used, with a >95% p value considered significant.

6.3 Results

The analysis shows that there is an overall significant difference in the performance of the engineering and non-engineering groups with the engineering groups outperforming the non-engineering groups. This gives an average for the engineering group of 2.91% and an average of -3.32% for the non-engineering groups after normalisation. This is a large difference, but it is important to understand where this difference comes from.

Looking at the data broken down by question difficultly, as in figure 6.2, initially it is clear that the students performed better on the simple questions. This is to be expected, but confirms the categorisation of the questions using blooms taxonomy was fair and accurate. Looking at the differences between the groups, it is clear that overall the engineering students are outperforming but there is a particularly large difference in the complex questions. The difference is statistically significant for the complex questions but not for the simple questions. The difference for the complex questions represents a 7.05% increase for the engineering group students with the complex questions representing 33.7% of the total questions answered.

Initial analysis was also performed to determine if the ability of the student influenced the difference in performance seen between the engineer and non-engineer groups. While there was

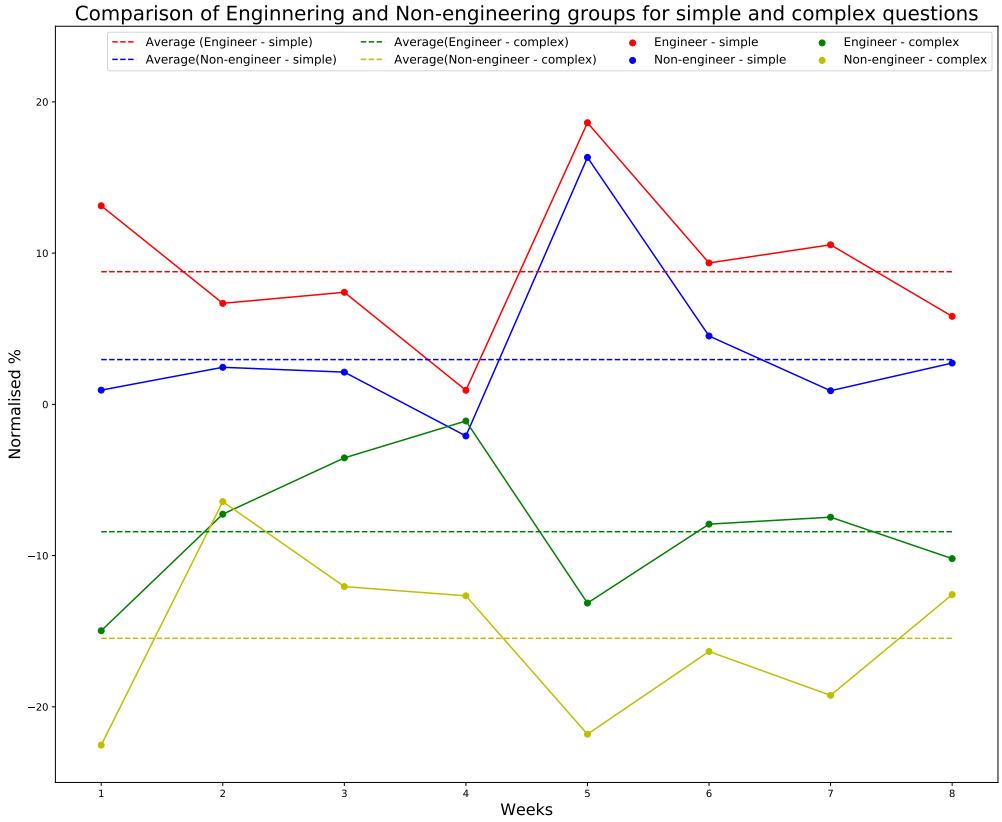


Figure 6.2: Comparison of high and low diversity groups for simple and complex questions

Results of formal assessment, week by week, averaged over all students. The red data points represents the simple questions for the engineer group students, and the blue data points represent the simple questions for the non-engineer students, the green points represent the complex questions for the engineer group students and the yellow points represent the complex questions for the non-engineer students. The dashed line is the average for each category and has been added to show the significant differences more clearly.

not a significant difference between the groups based on ability, for both the engineer and non-engineer groups the weaker students improved over the course of the summer school, showing the benefits of group work on student performance overall. It is possible that the effect is being obscured as the engineer group students are starting from a better position and so therefore the improvement may be smaller but more significant. Further analysis is required.

6.3.1 Additional Results

The results demonstrated in this chapter are potentially very significant and ideally would have been reinforced by data from 2019. While analysis was undertaken using the same methodology, only 2 groups contained no engineers and so group comparisons were inconclusive. This is also

true for comparisons of performance between groups with different numbers of engineering students. While analysis in aggregate for all years may highlight some reoccurring differences, normalising for changes in both overall cohort size and individual group size would be necessary.

6.4 Discussion

This study has demonstrated that cognitive and knowledge diversity within groups have a quantifiable and significant impact on the performance of students. This impact is large and consistent throughout the course and goes above and beyond the impact of group work alone. Given the style of assessment was focused on conceptual learning rather than problem solving and complex applications then we can link this improvement in performance directly to deeper understanding of the conceptual underpinning of the material. Given the improvement was particularly significant for the more complex questions, this further shows that the students are developing a deeper understanding of the material.

This is additionally supported by the drop in performance in week 5 for complex questions. Week 5 is the start of a new module, and with it, a move from classical dynamics to electromagnetism. Many of the engineering students who have a greater background in physics, those contributing to the knowledge diversity in the group will not have as much experience with electromagnetism, thus there is a drop in performance for those who were benefiting from that diversity. The improvement for the rest of the module, though never returning to the heights of module 1, is due to more accelerated improvement in the performance of the engineers. With a stronger basis in physics, they can get to grips with the new ideas faster, and thus return to a position of knowledge provider in the group.

This is also a fully quantitative approach to understanding group cohesion. Using the same style of questions used within the small group sessions to the assessment allowed for more simple analysis and a direct link of the improved performance to the group environment. High levels of homogeneity within the cohort also allowed for fluctuations in performance to be more readily identified by eliminating any noise created by variations between groups. This method could easily be implemented in any course containing group work as it is based solely on the final assessment for the course and does not require any further data.

Further projects will be designed to fully understand the nature of the interactions within the groups to be able to pinpoint the benefit provided by the engineering students themselves. It is likely that the engineering students provide an alternative view point that provides another access route to the material. This will also encourage more engaged and useful discussion within the group. Those projects will also seek to understand any impact on the engineering students performance as this project was purely focused on the experience of the life science students.

Chapter 7

Student-Led Laboratory Teaching

The efficacy of laboratory teaching has been considered from both a qualitative and a quantitative perspective, with a variety of quantitative analyses finding that there is no conclusive impact of laboratory teaching on student understanding and outcomes [Holmes and Wieman, 2018, Prades and Espinar, 2010, Wieman, 2015, Sobhanzadeh et al., 2017]. The expected outcomes of laboratory teaching can be varied but often focus on developing a mixture of technical and analytical skills. Measuring the efficacy of teaching approaches is often limited by the context and the tools used. As has been seen throughout this manuscript, a student cohort is a high dimensional system, controlled for and measuring the correct variable can be difficult. Individual variations between students can be high making identifying trends difficult. Careful choice of measurement and analytical tools is required.

In this project, the analytical tools used are standard. This kind of analysis has been done for laboratory teaching before, but the context of the Summer School provides an ideal set up and data to find conclusive results. Educational data is often so noisy that, once considering the context of implementation, identifying a conclusive result can be difficult. This project is a case study in how to use quantitative techniques to demonstrate, strong, conclusive impacts for teaching approaches.

7.1 Introduction

Instructional laboratories are considered an important part of undergraduate physics teaching [AAPT, 1997, AAPT, 2014, Hofstein and Lunetta, 2004]. Practical sessions are a key aspect of learning, giving students the chance to explore physics and see the concepts they are learning in action. Take for example, an experiment on projectile motion. The impact of gravity, velocity and angle can be demonstrated in a very simple experiment. Students can adjust the parameters easily and play with the concepts. There is also the opportunity to take data and analyse the results, moving through the theory, to the practical session, to the mathematical description. As such, students are expected to achieve a variety of intended learning outcomes from laboratory

	AAPT 1997	***AAPT 2014***
(1)	The art of experimentation	Designing experiments
(2)	Experimental and analytical skills	Developing technical and practical laboratory skills
(3)	**Conceptual learning**	**Analysing and visualising data**
(4)	Understanding the basis of knowledge in physics	Constructing knowledge
(5)	Developing collaborative learning skills	Communicating physics
(6)		Modelling

Table 7.1: Summary of Introductory Physics Laboratory Goals

This is the summary of goals for physics laboratories as outlined by the AAPT in 1997 and 2014. While many of the goals have remained the same, "conceptual learning" has been forgone in favour of a larger focus on "visualisation and modelling of data and systems".

sessions, as outlined in 1997 by the American Association of Physics Teachers [AAPT, 1997] (AAPT). These are primarily focused on the practical, analytical and collaborative skills needed for experimentation, with a very clear mention of developing conceptual understanding. Even without specific focus, laboratory teaching should have the flexibility to deliver all of these learning outcomes as each step of the learning process can be supported with practical teaching.

However, in 2014 the AAPT redesigned these goals [AAPT, 2014] contradicting the original overarching goals by removing the aim of developing conceptual knowledge as highlighted in table 7.1. This is reflected in recent papers which suggest that laboratory work has little to no impact on students' conceptual understanding [Holmes and Wieman, 2018, Prades and Espinar, 2010, Wieman, 2015, Sobhanzadeh et al., 2017]. This may be due, in part, to how broad the intended learning outcomes and overall goals are for laboratory sessions, with practical skills often coming at the expense of conceptual understanding [Wilcox and Lewandowski, 2017]. The practicality of laboratory sessions means that the focus often shifts, moving to technical and analytical experimental skills. With that shift of focus, come changes in the way the sessions are facilitated, delivered, and designed. Small changes in the structure of a laboratory course, change the way students engage with the learning. The students' expectations are also influenced by the materials they are provided; laboratory scripts which explain, step by step, how to complete an experiment. They become "cookbook-like" and prevent students from fully engaging with the laboratory session and the underlying concepts [Holmes et al., 2017, Wilcox and Lewandowski, 2016].

Alternative approaches for laboratory teaching have been designed, shifting the focus of the sessions by using problem or project based learning methods [Holmes and Wieman, 2018, Bouquet et al., 2017, Aslan-Tutak and Adams, 2006, Szott, 2014]. Such approaches seek to tackle many aspects of the traditional laboratory such as the "cookbook" style of instruction and lack of student engagement and ownership. These methods are often trying to emulate a more

realistic version of scientific research. Students are involved in the process from hypothesis to experimental design to results and analysis. These methods create an environment for students that allows for a much deeper engagement with the art of experimentation.

With this work as a basis, a method of laboratory teaching was designed, structured around conceptual understanding as the key learning outcome. While this work is building on the framework proposed by Wieman and Holmes, and Bouquet [Holmes and Wieman, 2018, Bouquet et al., 2017], this approach utilises a variety of methods to put student understanding at the core of the sessions.

7.2 New Approach

Prior to the introduction of the new approach the Summer School used the same laboratory sessions as those of the University of Glasgow Physics 1 and 2 cohorts. These labs, while successful at the University of Glasgow, do not serve the necessary function for the Summer School. The technical knowledge and experimental skill are key for physics majors but a focus on conceptual knowledge, especially in the intensive environment of the Summer School is necessary. A new approach was needed to support the overall aims of the Summer School.

Building on the work of Wieman, Holmes and Bouquet including the Investigative Science Learning Environments (ISLE) [Etkina and Heuvelen, 2007, Etkina et al., 2010], a new laboratory teaching approach was designed, putting conceptual understanding at the fore.

The ISLE labs utilise a long process of observation, hypothesis and testing to create a realistic research laboratory environment. The experiments took the form of projects and students worked on the same experiment for several weeks. This approach does tackle the issue of "cookbook" style of instruction as students are not following a lab script at all, they are only provided with an initial aim. However, it does therefore split the focus of the sessions to encompass both conceptual understanding and practical skills. The ISLE labs provide a good starting point, and show that much of what would be considered standard for a laboratory session is not necessary and can be stripped away.

For this to be viable for the Summer School a similar approach to the ISLE labs was used but with a shortened timescale. The experiments were kept simple; calculate g, measure the speed of sound, calculate the specific heat capacity of water. These simple experiments can be approached in different ways but the design process is kept simple. Focused on generally one central concept, the student needed to demonstrate understanding of that concept to succeed.

The scripts outlined the aims, what data should be collected and what results should be presented without providing a step-by-step guide. An example of the instruction provided for a projectile motion experiment is, *"1. Record the motion of a projectile with your smartphone in such a way that you can extract its position in the x and y axis. 2. Create a table on a spreadsheet with the position [of the projectile] in the x and y axis as a function of time"*. The scripts also

included questions to prompt the students to consider related ideas; how would the experiment change if these variables were different?

In addition, there were changes to what equipment was used in the laboratory, and how it was used. In traditional experiments, students use unfamiliar equipment, creating a barrier to fully understanding how the experiments work. The technical understanding of the equipment becomes a central part of the learning experience as the experiments cannot succeed without it. For the Summer School, this technical understanding is also unnecessary as these students are not physics majors. Taking advantage of the opportunity provided, in these experiments smartphones were used, which are extremely versatile and familiar to students. As well as capturing high quality video and still images, smartphones can be used as a variety of equipment such as magnetometers, tone generators and accelerometers. By using smartphones, students are no longer provided with equipment set up prior to the session, ready to complete an experiment, and so can take an alternative approach.

As with most lab work the students work in small groups. The use of design style labs and smartphones encourage discussion as the barrier to entry has been lowered for all students. There should be no experience or knowledge from outwith the course that is necessary to succeed in these labs.

By removing the barriers to student understanding in labs, as outlined by Wieman, Holmes, Bouquet and Etkina, and has been stated above, a new laboratory environment that encourages students to build from the conceptual underpinning can be designed. Within the summer school the focus was shifted through 2 mechanisms, the group mentors and the assessment. The group mentors created a consistent message for the student cohort - linking the lab work to the group work the students had already completed and allowing the students to make the connections between the different theoretical concepts and succeed in designing an appropriate experiment.

The traditional assessment method within practical laboratory teaching is to have students complete a report which discusses the methodology, results and analysis of the data collected during the laboratory session. This assessment method was used for the first year of the intervention (2017) to keep as much consistency with the pre-intervention assessment and allow a more robust comparison of student performance.

Additionally, changes in 2019 created a much more structured assessment framework which more clearly highlighted the broader theoretical context of the experiment. This shifted the focus even more significantly away from a purely practical mindset in the laboratory sessions. This was done by moving from a report-based assessment within the laboratory sessions to a discreet question and answer format. Rather than including some discussion of the conceptual basis of the experiment in the analysis section of the report, students were asked to answer specific conceptual questions. This encouraged the students to consider the conceptual basis of the experiment, not only while they were conducting the experiment (as they had done in previous years), but also when completing the assessment and demonstrating that understanding. The

more structured assessment also highlights to the students that this is a key learning outcome of the session.

7.3 Method

To assess the efficacy of the newly introduced practical laboratory sessions, student performance in formal assessments (8 end of week class tests and exams) was analysed. A comparison was made between student performance of two consecutive years of the International Physics Summer School, before and after introducing the new style of laboratory teaching. Student performance in lab related (LR) and non-lab related (NR) questions was compared within each year. Each question was coded as either an LR or NR question. Questions were classed as LR if the question directly referenced an experiment or asked about a concept as demonstrated within the experiment. Questions which referenced the experiment directly could take the form of Figure 7.2 and asked about scenarios that the student explicitly would have encountered during the experiment. Alternatively, the questions could ask about the concept without reference to the experimental set up, as in Figure 7.3. Non-lab related questions could follow a similar format, asking about experimental set ups but if those set ups were not directly related to an experiment completed during the Summer School then it was not categorised as lab related, see Figure 7.1.

For all assessments, each student completed a total of 160 MCQs. For all MCQs the percentage of students who correctly answered the question was determined. To compare LR and NR questions, the Mann-Whitney-U rank sum test was used, as the data has a non-normal distribution. This test is also useful within this context as it allows for comparisons between groups of varying sizes. The number of lab related questions is smaller than that of the non related questions. The values for each data set are shown in Table 7.2. Additionally,this statistical test has been used for analysis of student performance in assessment by Wieman [Wieman and Holmes, 2015]. No short answer questions (SAQs) were used in the analysis as there are very few questions that would fall into the category of LR. Any analysis that was undertaken for these SAQs would be unrepresentative of the overall trends.

In addition to analysing performance student feedback was also obtained through the use of surveys. Two surveys were distributed, one at the middle and one at the end of the course. A mix of Likert-type scale and semantic differential scale questions were used, as well as free response questions. These surveys were used to give a broad view of student engagement with the lab sessions and identify any potential disconnect between the intended approach of the labs and how they were perceived by the student cohort.

The quantitative analysis was carried out again for the 2 subsequent years (2018 & 2019) as the laboratory sessions continued to be facilitated in the same style.

Q14. A negatively-charged particle travels parallel to magnetic field lines within a region of space. Which one of the following statements concerning the force exerted on the particle is true?

A. The force is perpendicular to the magnetic field.
B. The force is perpendicular to the direction in which the particle is moving.
C. The force slows the particle.
D. The force accelerates the particle.
E. The force has a magnitude of zero newtons.

Q15. What must the initial state of motion of a charged particle be if it follows a helical path in a magnetic field?

A. It must be moving at an angle, neither parallel nor perpendicular, to the magnetic field.
B. It must be moving in the direction opposite to the magnetic field.
C. It must be initially at rest when it is placed in the magnetic field.
D. It must be moving perpendicular to the magnetic field.
E. It must be moving parallel to the magnetic field.

Figure 7.1: Example of NR Questions

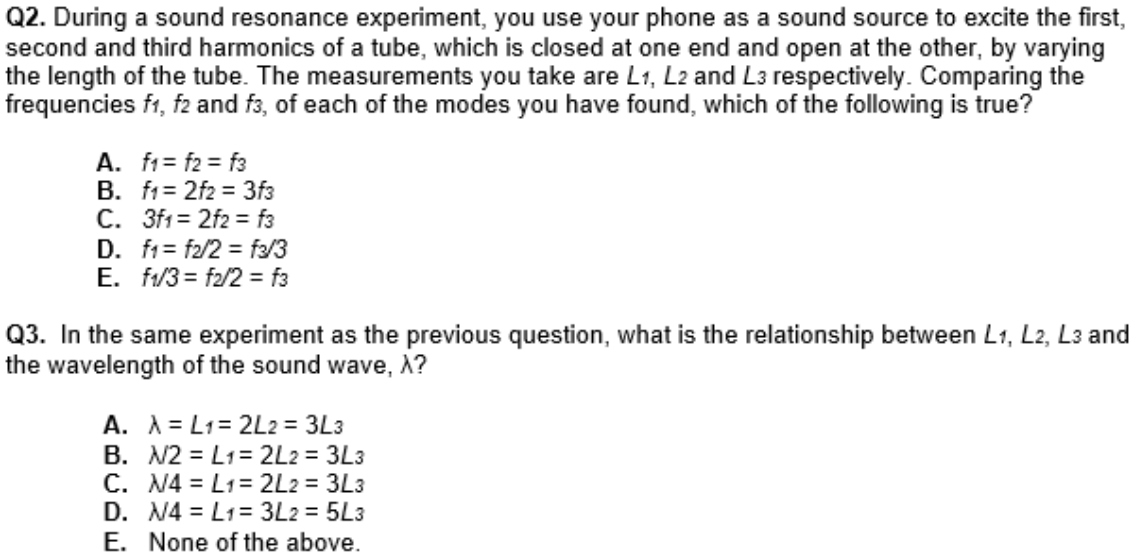
These questions are examples of Non-lab Related questions. These questions have no relation to any experiment the students completed during the Summer School. While students did complete experiments relating to magnetism, there was no experiment related to particle motion.

7.4 Results

7.4.1 Evidence from Initial Intervention

Before any analysis was conducted, the data was normalised for any variability in the complexity of the material between class tests and exams. The average performance was used for the normalisation calculation. The data for 2016, 2017, 2018 and 2019 is presented in Figures 7.4a, 7.4b, 7.5a, 7.5b. For all years there is no data for weeks 1 and 5, these weeks did not have laboratory sessions and so have been excluded from the data. The data for 2016 is pre-intervention and uses the traditional demonstrating and facilitating techniques used in physics laboratory teaching. The data for 2017, 2018 and 2019 is post-intervention, using the new laboratory teaching techniques. In all years there is a higher performance in lab related questions. As the students are exposed to concepts in a different environment, one that allows students to see the connections between different ideas, this follows.

However, in 2016 the difference is not statistically significant. The average performance for the non-lab related questions is -1.53% and the performance for lab related questions is 2.61%, as can be seen in Figure 7.4a. The data in Table 7.2 shows that there is no significant



Q2. During a sound resonance experiment, you use your phone as a sound source to excite the first, second and third harmonics of a tube, which is closed at one end and open at the other, by varying the length of the tube. The measurements you take are L_1, L_2 and L_3 respectively. Comparing the frequencies f_1, f_2 and f_3, of each of the modes you have found, which of the following is true?

A. $f_1 = f_2 = f_3$
B. $f_1 = 2f_2 = 3f_3$
C. $3f_1 = 2f_2 = f_3$
D. $f_1 = f_2/2 = f_3/3$
E. $f_1/3 = f_2/2 = f_3$

Q3. In the same experiment as the previous question, what is the relationship between L_1, L_2, L_3 and the wavelength of the sound wave, λ?

A. $\lambda = L_1 = 2L_2 = 3L_3$
B. $\lambda/2 = L_1 = 2L_2 = 3L_3$
C. $\lambda/4 = L_1 = 2L_2 = 3L_3$
D. $\lambda/4 = L_1 = 3L_2 = 5L_3$
E. None of the above.

Figure 7.2: Example of LR Questions

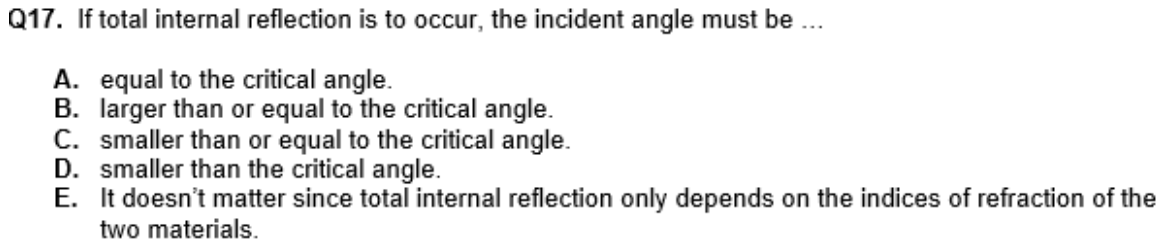
These questions are examples of Lab Related questions. These questions directly reference an experiment students have completed. Both the questions asked and the set up should be familiar to students.

difference. That there is some difference for all years indicates that there is some impact on student understanding based on the laboratory teaching.

The results from 2017 again show a higher performance in lab related questions than non-lab related questions, with an average performance of -1.33% for non-lab related questions and 7.27% for lab related questions, as can be seen in Figure 7.4b. This is a statistically significant difference, as shown in Table 7.2. Though there is no data for week 4 for 2017, the increased performance is more consistent throughout the 2017 Summer School.

Between 2016 and 2017 there were no changes to the structure or material of the Summer School. While the cohorts had different students the demographics were very similar, more so than with a traditional undergraduate cohort. The experiments were covering broadly the same material, with only the structure of the laboratory sessions changed.

To quantify the impact of the intervention, the effect size was calculated using the week by week difference in student performance in LR and NR questions for pre and post-intervention, shown in Figure 7.4a and Figure 7.4b respectively. This effect size is specifically comparing the performance in LR questions of the pre and post-intervention cohorts. As the data sets are of a similar size and standard deviation, the effect size Cohen's d was used and is defined as:



Q17. If total internal reflection is to occur, the incident angle must be ...

A. equal to the critical angle.
B. larger than or equal to the critical angle.
C. smaller than or equal to the critical angle.
D. smaller than the critical angle.
E. It doesn't matter since total internal reflection only depends on the indices of refraction of the two materials.

Figure 7.3: Example of LR Questions

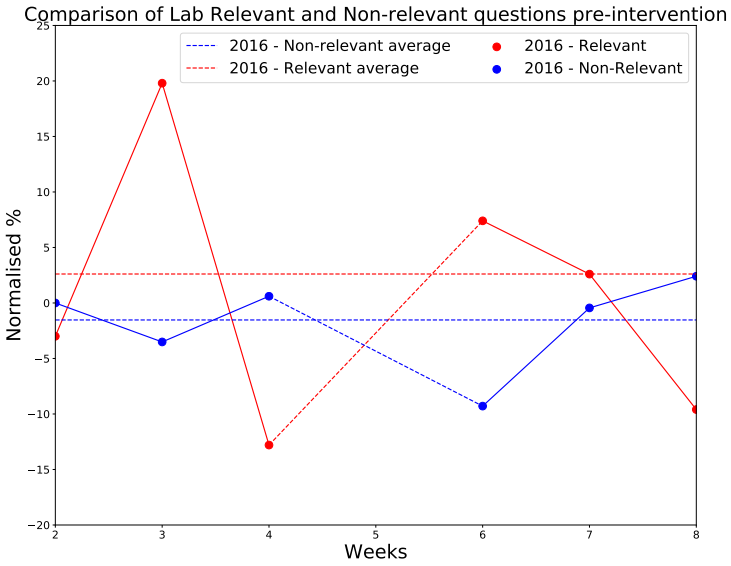
This question is an example of a Lab Related question. Students were asked to complete an experiment which included finding the critical angle and how this related to total internal reflection, which is reflected in this question. Unlike in Figure 7.2 this question does not directly reference the experiment, but it clearly references the same scenario.

$$d = \frac{\bar{x}_{post} - \bar{x}_{pre}}{SD_{pooled}} \tag{7.1}$$

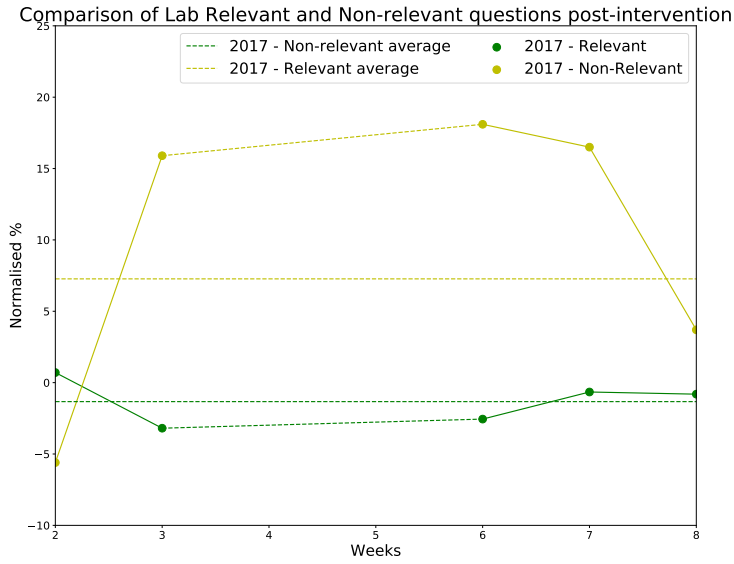
where $\bar{x}_{post}$ is the average of the post intervention cohort, $\bar{x}_{pre}$ is the average of the pre intervention cohort and SD is the standard deviation of the sample population. However, as there was no sample population to take the standard deviation from, a pooled standard deviation was calculated using:

$$SD_{pooled} = \sqrt{\frac{(N_{post}-1)SD^2_{post} - (N_{pre}-1)SD^2_{pre}}{(N_{post}-1)(N_{pre}-1)}} \tag{7.2}$$

where N_{post} & N_{pre} are the sample size and SD_{post} & SD_{pre} the standard deviation of the 2017 and 2016 cohorts respectively. This is outlined by Olejnik and Algina [Olejnik and Algina, 2000], with 2016 as the control group as is defined there. This gave an effect size of **d = 0.69**, or an improvement of 2/3 of a standard deviation. This improvement to the success rate on LR questions is equivalent to an increase of 9.6%, almost a full grade improvement. While LR questions only make up 14.2% of all MCQs students answered, the improvement is conclu-



(a) Summary of Pre-Intervention Results



(b) Summary of Post-Intervention Results

Figure 7.4: Pre and post intervention assessment results

Results of formal assessment, week by week, averaged over all students. The dashed line is the overall average for all assessment and has been added to show the significance of the improvement of the intervention. For weeks 1 and 5 for the pre-intervention (a) and weeks 1, 4, and 5 for the post-intervention (b), there are no LR questions and so a comparison cannot be made.

sive. Figure 7.4b shows the improvement in how well students are performing in LR questions post-intervention is significant and consistent. Given the similarity in the averages for the NR questions between the cohorts this improvement can be attributed purely to the introduction of

	Pre-intervention (2016)		**Post-intervention (2017)**	
	Related	*Non-Related*	*Related*	*Non-Related*
Count	18	102	19	114
U Value	816	1020	812	1354
U Critical	693.7		826.7	
Mean	918		1083	
Standard Deviation	136.1		155.5	
P-Value	0.280		0.04	
Significant	**No**		**Yes**	

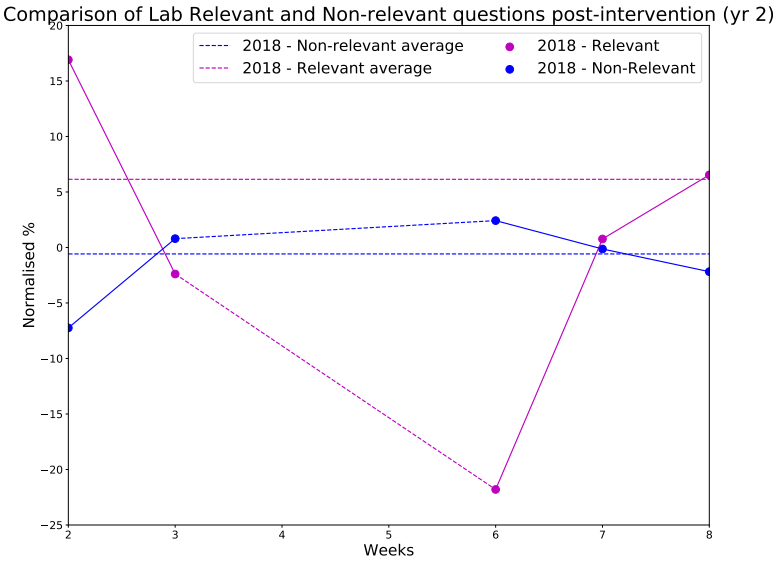
Table 7.2: Mann-Whitney-U Test Results

Summary of the Mann-Whitney-U test results for both pre and post-intervention. For both the pre and post-intervention data sets, 2 data subsets are being compared, the lab related and non-related questions. Significance was set at 0.05 so it is clear that only the post-intervention data has a significant difference.

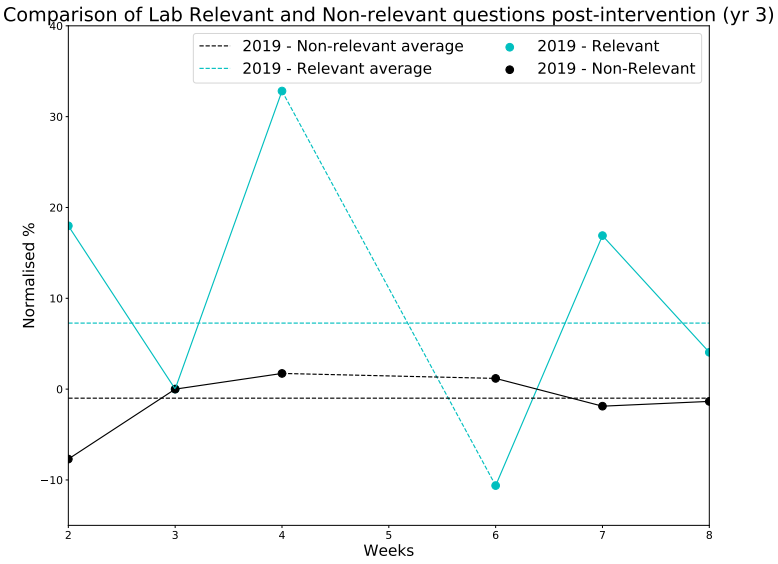
the new style of laboratory sessions.

Given the success of the intervention in 2017, the new laboratory structure was used again in 2018 and 2019. The results of the Mann-Whitney-U analysis are presented in Table 7.3 The results for 2019 show a similar performance to 2017, as can be seen in Figure 7.5b, with a statistically significant difference between non-lab related questions and lab related questions. The average performance for non-lab related questions was -1.00% and for lab-related questions was 7.27%. This is the same average performance in lab-related questions as was seen in 2017. Some changes had been introduced to the to the assessment of the laboratory sessions themselves, though this was focused on the production of the laboratory reports and not on the learning aims or the outcomes of the experiments. The implementation in 2018 was less successful.

Overall the data shows that there was no significant difference between LR and NR questions in 2018. Performance throughout the Summer School in 2018 was poor in comparison to other years and may account for a more limited impact of the intervention. This style of laboratory teaching requires high investment from both staff (including mentors) and students and must be clearly introduced and explained to staff and students before the course begins. The mentors have not experienced this style of laboratory teaching from the student perspective and can struggle to "buy into" the concept. This is exacerbated by the fact that the mentors are themselves students and still maintain the student perspective; looking at the additional work and engagement required for these laboratory sessions as a waste of time rather than valuable as it does not seem directed related to the main assessment. The average performance for non-lab related questions was -0.58% and for lab-related questions was 6.14%. The data is in-line with the results from all years, a higher performance was seen in lab-related questions for the pre-intervention year as well. Even with potentially poor implementation, the student performance is, at least, analogous to a traditional laboratory set-up, and still has the potential for much improved results.



(a) Summary of Post-Intervention Results (2018)



(b) Summary of Post-Intervention Results (2019)

Figure 7.5: Post intervention assessment results (2018 & 2019)

Results of formal assessment, week by week, averaged over all students for 2018 and 2019, the second and third year of the laboratory intervention. The dashed line is the overall average for all assessment and has been added to show the significance of the improvement of the intervention. Again weeks 1, 4, and 5 for 2018 and week 1 and 5 for 2019 included no LR questions and so a comparison cannot be made.

Every implementation of an intervention will vary in success. The individuals involved will invariably intersect with the intervention in different ways. With this, there is an inherent risk. Here it has been demonstrated that the risk of employing this intervention is low but the potential

	Post-intervention (2018)		**Post-intervention (2019)**	
	Related	*Non-Related*	*Related*	*Non-Related*
Count	20	139	21	139
U Value	1135	1645	1062	1858
U Critical	1073		1133.5	
Mean	1390		1460	
Standard Deviation	193		197.9	
P-Value	0.09		0.02	
Significant	**No**		**Yes**	

Table 7.3: Mann-Whitney-U Test Results

Summary of the Mann-Whitney-U test results for second (2018) and third (2019) intervention implementations. Significance was set at 0.05 so it is clear that only the 2019 data has a significant difference. The data shown in Figure 7.5a demonstrates where this variation in significant arises.

gain is large. The broader context of the education environment that interventions are employed within necessitates the use of several year studies to fully prove the worth of any intervention. A non-conclusive performance in one year should not represent an inherent failure, though of course the converse is also true. If an intervention has no impact then what has changed through that intervention is perhaps not a key component of the course.

7.4.2 Qualitative Evidence

The perception of the new laboratory sessions was positive, with the quantitative response to the student surveys shown in Table 7.4. These results are the aggregate response at the mid-point and end of the Summer School. The most positive responses were for how helpful the laboratory sessions were. Feedback also indicated that students felt there was a strong link between the theory and the experiment. If students feel that they are benefiting from the labs, they will engage more, which further enhances the benefit. There can be an expectation that students will not see the benefit of new or alternative approaches in learning and teaching, especially if there is not a direct link to assessment. While there were clear links to the laboratory sessions within the summative assessment, there was also a deliberate choice made to be clear about the benefits of the new laboratory structure. This clear communication and engagement with the students supported a more enthusiastic and positive response from the student cohort and is reflected in this positive feedback.

In the free response questions students highlighted that discussion in the labs was particularly helpful, as well as the support from the lab demonstrators. This feedback highlights that the student cohort was engaging with the intervention in the way it was designed. High levels of discussion are necessary for student-led laboratory teaching to succeed [Williams et al., 2017]. That the student cohort saw the link between discussion and success in the laboratory sessions demonstrates a very successful implementation of the intervention.

	Positive	***Negative***
Overall	76.0%	23.96%
Interesting	80.8%	19.2%
Enjoyable	69.9%	30.1%
Helpful	92.3%	7.7%
Satisfying	72.2%	27.8%
Understandable	72.4%	27.6%
Well Organised	68.6%	31.4%

Table 7.4: Summary of Student Feedback Survey

Results of the student surveys. The response from the students was overall positive, with no aspect of the laboratory sessions garnering a more negative response than positive. Overwhelmingly, students found the laboratory sessions extremely helpful.

A comparison was drawn between male and female students to identify if there was any variation in student preference. The distribution of responses was similar for both male and female students. As the experience for female student in laboratory settings is often less positive, especially in relation to confidence and discussion, this is a very positive result. The nature of the intervention facilitates discussion and normalises asking questions to both students and demonstrators.

A comparison was also made between the life sciences and engineering cohorts. The engineering students were more likely to have previous experience in physics laboratories or similar environments and therefore could make a direct comparison of their experiences. The response from engineering students was again incredibly similar to that of the life sciences students. This positive feedback is encouraging as the engineering students may have conducted these experiments before in their traditional setting and so the very positive responses suggest a directly comparable, and better experience in the new laboratory setting.

A similar survey distributed in 2018 showed a similar enthusiasm for discussion during the laboratory sessions, especially with the demonstrators. Students felt that designing the experiment was useful in enhancing their understanding of the course material, but highlighted that the use of smartphones was not necessarily a key part of that process. While almost all current students are very familiar with smartphones, the context in which the phones are used in the experiments is often unfamiliar. Many students are unaware of the existence of for example, a magnetometer within their phones. The phone can then become just another piece of laboratory equipment. However, students always retain a sense of ownership when using their own phones as part of the experiment, something which is not lost whatever the context of the experiment.

7.5 Discussion

This study demonstrates significant gains in conceptual learning and understanding from the approach to introductory physics teaching laboratory described above, as measured in the pre-existing course assessment tests ($p < 0.5$; $d = 0.69$ (a medium effect size)). This in turn shows that conceptual learning gains are an achievable learning objective for laboratory teaching, in contrast to the abandonment of this objective by the AAPT.

Key features of the approach reported here include, first, the use of simple physical equipment together with students' own mobile phones. Simplifying the equipment used greatly reduces the effort needed by students to conduct the experiment. The lower barrier to entry allows for less confident students to engage in the experimental process, and through that process, engage with the underlying conceptual knowledge.

Second, students are given instructions that provide light scaffolding, rather than comprehensive instructions to follow, leading to the laboratory work becoming a student-led, problem solving activity, providing a sense of self-determination and ownership. [Chan et al., 2014] This is enhanced by the use of students' own devices. The experiment then becomes something that the student has a personal record of, that can be shared and referred back to. Students are given the opportunity and agency to develop and implement their own ideas. Even when a student is replicating a "standard" experiment, they are reaching the conclusion via their own path and without pre-conceived notions of what the "correct" solution is. Removing the idea of the striving for a correct answer, but rather, striving to creating a method that works and analysing the gathered data to find an answer, not necessarily the correct answer but the logical answer, is a more authentic experience and connects students to the real world context of conceptual physical theories.

Finally, as with all laboratories, students work in small groups designed to elicit constructive peer interaction and discussion [Bennett et al., 2010]. And given the unique structure of the summer school, with a focus on peer learning throughout the course, these students can engage in this peer interaction in the most effective and constructive manner. Working in small groups allows the students to analyse the problem, propose solutions, communicate their thinking and critically evaluate and compare others' ideas, hence achieving the majority of goals laid out by the AAPT for laboratory teaching.

Future work will seek to clarify which aspects of the intervention are crucial and contribute the most to this learning gain. As the use of smartphones proliferates in teaching laboratories across STEM, smartphones are being introduced into laboratories as ways of visualising molecules in chemistry [Williams and Pence, 2011] or for collecting and analysing behavioural data in psychology [Miller, 2012], there is a need for careful consideration of the efficacy of the approach described in this paper. While the continued success of this intervention has been shown in subsequent years, it is necessary to use a similar approach in a variety of fields and contexts before assuming that this intervention will be universally successful.

Specific factors affecting our implementation should be noted. The International Physics Summer School is aimed at non-physicists with the focus of the course on the conceptual understanding rather than on technical aspects or skills, aligning with the global approach of the course. Hence, this approach to laboratory teaching would be ideal for similar courses that prioritise the conceptual rather than the technical. Also, the reported approach has yet to be tested on a standard introductory physics class. While such factors should be noted, the results illustrated above remain important to consider in the design of all laboratory teaching.

One question to which this study contributes is what the learning objectives of laboratory classes could and should be, and what types of laboratory class are feasible. Traditional laboratory work focuses on getting the experiment and equipment to produce a desired outcome. This study proposes that laboratory work need not be an exercise of trying to get equipment to work as advertised but can instead have a more meaningful relationship to scientific thinking (AAPT 1997/2014 objective 1) and can simultaneously raise scores on pre-existing assessment tests designed to measure conceptual understanding (AAPT 1997 objective 3).

Additionally, this study demonstrates the necessity of long term non-intrusive quantitative techniques. All courses are structured with multiple component sections. A standard science degree at university will include lectures, laboratories, seminars or tutorials and independent study. The reach and impact of any individual component of the course is hard to identify but this is also true of the internal aspects of each component. The use of quantitative methods can allow for interventions to be improved through iterative processes and for the specific impact of those changes to be identified. All education research needs to be considered over long term studies and only quantitative analyses can allow for large scale, long term, broadly applicable studies to be successful. However, there is a need for qualitative studies to support the interpretation and contextualisation of these long term studies.

Chapter 8

Discussion and Conclusion

8.1 Introduction

For any individual physics course there currently exists the tools to measure variables, to make predictions and measure outcomes of any intervention using a physicist's approach. A model can be created and tested through quantitative measurements and observation. While the results presented in this manuscript are specific to the cohorts and environment measured, the broader conclusion that positive, quantitative results from a physicist's approach are broadly possible in education research, can and should be drawn.

The methods used in this manuscript can be applied in any higher education physics environment. Given the intersectionality of education research, the democratisation of the research tools is the best way to deepen the understanding of the field. While the International Physics Summer School is an ideal environment, these methods can and must be applied in any environment that has the necessary data. Education research is always limited by scope and scale. By building quantitative tools that can be implemented by non-specialists, the scale of research can increase massively. A greater scale provides a much better chance of untangling the intersections and interweavings of the variables, skills and knowledge that underpin student performance.

This manuscript outlines how to understand the structure of educational experiment. Starting with the equipment (Chapters 2 & 3), what is being observed (Chapter 4), and the range of analyses that are possible with the same data sets (Chapters 5, 6 & 7).

8.2 Assessment

Assessment, like any analytical tool, needs to be calibrated. There are many external sources of calibration such as Blooms Taxonomy or the Force Concept Inventory (FCI). These calibrations show the true metric of the assessment used. This is especially important when multiple forms of assessment are being used concurrently, such as multiple choice questions (MCQs) and short answer questions (SAQs), or when there are multiple skills or areas of knowledge being assessed

at the same time, such as mathematical skill and conceptual physics knowledge.

In this manuscript I show that when considering how to interpret the results of assessment there is a wealth of data tied into the format of the assessment. How a student interprets a question is incredibly important and how a question is formatted has a significant impact on that. The results in this manuscript have shown that MCQs, no matter where they fall within Bloom's taxonomy, will create the same ranking within a cohort. This suggests that there is an impediment to student performance within the MCQ structure. A student should not be able to perform just as well in a question that requires problem solving as one that only requires recall, unless the barrier to success is within the structure of the question. This does not mean, however, that MCQs cannot be useful assessment tools. The correlations for both SAQ and MCQ with both the FCI and MST are consistent and indicate that both assessment types can be used interchangeably when considering physics knowledge. The key learning from this MCQ analysis is that assessment tools will always require some skill on the part of the student to engage with appropriately, and that teaching these skills is a necessary part of any course. How to interpret the language of an MCQ may seem like it would be the same in any subject, but how language, especially scientific language that is also in common usage, is framed in a question may not be familiar to all students.

Considering the structure of assessment is also inherently linked to how it is marked. If a student has the opportunity to demonstrate some understanding without requiring a complete knowledge then those questions may provide a more nuanced view of student understanding. This was highlighted in the results in Chapter 2 by the significantly higher levels of discrimination for calculation SAQs versus explain SAQs. These question types are taxonomically similar, have very similar average performances and assess the same kinds of conceptual understanding. The reason for the significantly different discrimination values can only be caused by another skill being necessary to complete the questions. While this is likely to be based on the mathematical skills necessary to complete these questions it may also be related to how a student demonstrates their conceptual understanding or the ability to apply that conceptual understanding and mathematical skill in a new context.

While the results provided in Chapters 2 & 3 have highlighted various skills required to successfully answer a question, these are not the only skills that a student can use. The skill of answering an MCQ is not one single process, it can incorporate problem solving, textual analysis, logic and many others. What these results demonstrate is that understanding how a student performs in an assessment goes far beyond the understanding of the course material.

There are many other variables that can impact student performance but this project has demonstrated that the simple use of correlations and discrimination analysis can identify which variables are the most significant. It is important to understand as many of these skills as possible to fully understand not only what is impacting student performance but also what can be influenced. This kind of analysis can become a standard part of the analysis conducted at the

end of any physics course in higher education. These are tools that will be familiar to most physicists, just applied in a different way.

It is important to consider that all of the results presented in Chapters 2 & 3, such as the disparate discrimination values between calculation and explain questions, may be only an artifact of a non-physicist cohort.

Many of the questions raised in the assessment projects remain unanswered. While it is clear that there is an impediment to student performance in MCQs, the mechanism of that impediment is not known and requires further study, likely taking a interdisciplinary approach. As MCQs are not always used as part of summative assessment in physics within the UK, the use of MCQs for assessing mathematical concepts can be debated [Main and of Physics (Great Britain), 2022], it may be difficult to dive deeper into these mechanisms within physics education specifically.

However, further work considering the impact of unfamiliar language on student performance may be more relevant while still potentially related to the impact of MCQs. The work presented in these projects is limited as the students are non-physicists and any further research on the impact of mathematical or scientific language should consider the difference between the novice and the expert physicist. Student interpretation will always have an impact on performance, but that interpretation will be built on experience and therefore will be impacted by demographics. It is important to conduct this research on the relevant cohorts as this kind of data cannot be extrapolated.

The results presented in these chapters may also contribute to the continued review of the Force Concept Inventory as a standardised tool. The FCI never had a very strong correlation with performance and, unlike other parts of this project, the FCI is designed to be used with physicists and non-physicists alike. As the Summer School is a majority female cohort, this data adds to the growing body of evidence that the FCI produces gendered results [Normandeau et al., 2017, Lorenzo et al., 2006]. This limitation of the FCI must be considered as part of the context of these results.

8.3 The Student

When considering the intrinsic variables for an individual student, all significant differences identified in this project can be linked to two categories, grade point average (GPA) and prior physics experience. While GPA is not a categorical predictor of student performance within the Summer School it does have a strong correlation. The vast majority of students on this course have not completed any physics based courses at university, however, GPA can still identify students with a strong ability to learn, something that is transferable to any course. The ability to build knowledge as a cohesive unit rather than as discrete knowledge centres is why measures such as GPA still have such a strong correlation with performance.

This project was limited, however, by the lack of data regarding socio-economic factors. This

limitation comes not only from a lack of data, but also a lack of context. The Summer School cohorts contain predominantly American students and the impact of socio-economic factors are not only unknown but set within a different culture context. These factors likely intersect with those variables identified, and may have another underlying effect as was seen with GPA and prior experience.

Due to these limitations, this kind of research needs to be conducted on as wide a scale as possible to fully understand the impact of metacognition and the ability to integrate that understanding of learning with prior knowledge and how that lowers the barrier to entry. This kind of data can be seen through the lens of threshold concepts. There is a question around the impact of a shared vocabulary that can arise from prior experience and whether that impacts how a student may pass through the liminal spaces that arise around these threshold concepts. This experience is likely contributing to the highly individual process of understanding and passing through these thresholds [Nicola-Richmond et al., 2018]. Application of knowledge is much easier when the only impediment is the situation in which the knowledge is to be applied. Adding the additional barrier of unfamiliar language can be especially taxing as it amplifies all other barriers to performance. This may be trapping students in these liminal spaces. As physics is often taught in a linear way, if these thresholds are not passed then the next set of material will only compound the lack of understanding. Considering the results of Chapters 2 & 3, these threshold concepts may also come from related academic areas (i.e. mathematics) or more broadly, in a metacognitive sense, such as knowing how to correctly apply knowledge. This can only be understood by considering larger data sets and comparing the various differences in prior knowledge and metacognitive skill. Further research in this area would also benefit from longitudinal studies, understanding the impact of these variables, not only on conceptual understanding during the course but also long term retention. Students who are successfully passing through these thresholds should also see better long term retention of their understanding. If these studies are also supported by qualitative approaches such as interviews and focus groups, a better understanding of the core mechanism, be it an aspect of metacognitive skill or purely greater experience with the subject, impacting student understanding can be found.

An understanding of the underlying intrinsic variables within a student population that can influence performance can allow educators to adapt their assessment to minimise these variations. If the cohort has limited experience of physics then it is likely that the language and philosophical approach to physics will need to be overcome before real learning can begin. This could be a simple change in the way a concept is introduced, but it could have a significant impact on how a student approaches that threshold.

GPA and prior knowledge may not be the key indicators for every cohort, but understanding these intrinsic variables is a very powerful tool. Even limited predicting power may allow an educator to make a small change at the start of a course that has an impact throughout.

8.4 Group Dynamics

A quantitative approach to analysing group dynamics is possible. The use of measures such as standard deviation, that measure consistency within a group, can potentially be considered a numerical representation of cohesion within a group. With analysis of the impact of groups on learning this could be a tool that provides a new approach to considering group dynamics. However, there is still a need for further analysis to understand whether there is a particular aspect of the interaction that is captured within the standard deviation measure.

While this method was designed to show that a quantitative approach to analysing group dynamics is possible it has demonstrated that a variety of approaches to education research are necessary. The understanding of group dynamics has potentially been limited by the inherent siloing of qualitative and quantitative researchers.

It is perhaps unsurprising that the use of correlation analysis within the group was unsuccessful, however, there may be learnings that can be taken, not only for quantitative education researchers, but qualitative as well. While the relative success of the standard deviation as a measure does suggest that if the groups were designed to be more similar initially then the correlation analysis may be able to identify significant differences, but it may also indicate that there is a different kind of cohesion forming in the group. Qualitative observations of the group may provide some light on how cohesion manifests in a group. There are still questions to be answered around how knowledge spreads within a group and whether the variables discussed in Chapter 4, for example metacognitive ability, may impact those vectors.

Additionally, while each of these approaches can potentially shed light on student dynamics in groups, this approach does not provide evidence linking positive group dynamics to overall performance. This data cannot say categorically that group work has a positive impact on student performance in a general sense.

In contrast, the results of Chapter 6 do show a strong positive impact due to group dynamics. This study demonstrated that cognitive and knowledge diversity within groups has a quantifiable and significant impact on the performance of students. This impact is large and consistent throughout the course and goes above and beyond any potential impact of group work alone. Given the style of assessment was focused on conceptual learning rather than problem solving and complex applications then we can link this improvement in performance directly to deeper understanding of the conceptual underpinning of the material.

This could be considered a fully quantitative approach to understanding group cohesion. Using the same style of questions used within the small group sessions to the assessment allowed for more simple analysis and a direct link of the improved performance to the group environment. This method could easily be implemented in any course containing group work as it is based solely on the final assessment for the course and does not require any further data.

There is a limit to extrapolation of this method. Diversity of physics knowledge can be easily measured quantitatively, and in the case of the Summer School existed within a binary, either

existent or non-existent prior knowledge. The results from Chapter 5 illustrate that this is not necessarily a method that would work outwith a binary variable.

Additionally, the Summer School used the flipped classroom method, with a specific focus on group work. This method is labour intensive and requires buy-in from both staff and students. In a course with a smaller component of group work it is possible that the striking impact found in chapter 6 would not be replicated.

Considering further avenues for this research, a qualitative approach, observing groups for a more specific variable than group cohesion, could potentially achieve similar results to Chapter 6. This avenue could be supported by the current research that focuses on student experiences within groups [Gapp and Fisher, 2012, Roth et al., 2010] The scope of the qualitative measures of Chapter 5 may have been too broad, where the specificity of the binary comparison in Chapter 6 allowed for a much clearer picture. An interdisciplinary approach would best support research into group dynamics going forward, especially considering the nuance of how the makeup of a group impacts performance, and not just quantifying the impact.

8.5 Student-led Laboratory Teaching

This study demonstrates significant gains in conceptual learning and understanding from the approach to introductory physics teaching laboratory described in Chapter 7, as measured in the pre-existing course assessment tests ($p < 0.5$; $d = 0.69$ (a medium effect size)). This in turn shows that conceptual learning gains are an achievable learning objective for laboratory teaching, in contrast to the abandonment of this objective by the AAPT. These results are built on the conclusions throughout the rest of this manuscript. The assessment is a well defined tool that can provide a clear and definite result, the student cohort is understood well enough to allow for a fair comparison between years and the statistical approach incorporates the methods tests in the group analysis. Additionally, a qualitative measure was used to supplement and better understand the statistical results.

The key features of the approach used include, first, the use of simple physical equipment together with students' own mobile phones. Second, students are given instructions that provide light scaffolding, rather than comprehensive instructions to follow, leading to the laboratory work becoming a student-led, problem solving activity, providing a sense of self-determination and ownership. Finally, as with all laboratories, students work in small groups designed to elicit constructive peer interaction and discussion. And given the unique structure of the summer school, with a focus on peer learning throughout the course, these students can engage in this peer interaction in the most effective and constructive manner.

All of these aspects have been implemented in laboratory teaching before, but to the best of our knowledge this is the first time that they have been implemented together in this way. These approaches have also not often been implemented in a physics environment. While there are

many techniques and approaches to education that are subject specific, this kind of pedagogic shift is applicable to any subject with a practical element, be that chemistry or fine art.

Traditional laboratory work focuses on getting the experiment and equipment to produce a desired outcome. This study proposes that laboratory work need not be an exercise of trying to get equipment to work as advertised but can instead have a more meaningful relationship to scientific thinking (AAPT 1997/2014 objective 1) and can simultaneously raise scores on pre-existing assessment tests designed to measure conceptual understanding (AAPT 1997 objective 3). This could drive change throughout STEM, as well as practical, non-STEM subjects.

Additionally, this study demonstrates the necessity of long term non-intrusive quantitative techniques. All courses are structured with multiple component sections. A standard science degree at university will include lectures, laboratories, seminars or tutorials and independent study. The reach and impact of any individual component of a course is hard to identify but this is also true of the internal aspects of each component. The use of quantitative methods can allow for interventions to be improved through iterative processes and for the specific impact of those changes to be identified. Demonstrating the impact of any intervention is at its most persuasive when considered over long term studies. Quantitative analyses can better facilitate large scale, long term, broadly applicable studies. This would then allow qualitative approaches to be used in more efficient ways, helping understand the underpinning mechanism of the results found quantitatively.

Future work will seek to clarify which aspects of the intervention are crucial and contribute the most to this learning gain. The use of smartphones is proliferating in teaching laboratories across STEM; smartphones are being introduced into laboratories as ways of visualising molecules in chemistry [Williams and Pence, 2011] or for collecting and analysing behavioural data in psychology [Miller, 2012]. With the broader potential applications, there is a need for careful consideration of the efficacy of the approach that was described in Chapter 7. The efficacy must be considered outside of the confines of the Summer School and of physics. While the physics approach to education provides a systematic and clear structure for implementation and evaluation, and helps communicate educational interventions to non education researchers, there must be a flexibility when sharing this research outside of the subject confines. All subject education researchers bring a bias towards their research styles and standards, and only through collaboration, will there be widespread change and development.

For this laboratory teaching approach there appears to be several factors that affect the implementation. The Summer School is aimed at non-physicists with the focus of the course on the conceptual understanding rather than on technical aspects or skills. Hence, this approach to laboratory teaching would be ideal for similar courses that prioritise the conceptual rather than the technical. However, it is also important to consider that development of technical skills is a key outcome for any physics degree. This laboratory approach has not been evaluated in that area and so likely cannot fully replace a standard physics laboratory session. If this approach was

used in conjunction with a more traditional laboratory set up the impact will likely be affected. This will limit the number of courses where this set up would be ideal.

Also, the reported approach has yet to be tested on a standard introductory physics class. And while the impact on a standard physics cohort may be different, the implementation in a life-sciences course may show similar effects as the Summer School cohort is primarily life-science majors. Were this approach to be implemented in a standard chemistry or biology course, a much better understanding of the impact of cohort on the efficacy of the intervention could be understood.

8.6 Conclusion

The results presented in this manuscript, for example, the impact of laboratory work on student results, will not be found in every physics course. What has been proven is that these results can exist, and that those results can be measured quantitatively. The same goes for the impact of group dynamics on individual success and how we assess students. The tools exist to demonstrate that significant, measurable change is possible and can be incredibly powerful. This research was conducted in a highly controlled environment, and as such does of course provide a best case scenario for education research. The results presented here are a kind of "idealised" version of an intervention. There are so few opportunities to try new techniques and be able to compare, statistically, the impact; having a cohort that is as consistent year on year as the summer school is not a common occurrence. Conversely, this cohort is not a standard physics cohort, and any individual results must always contain that caveat. While the methodologies and interventions can be implemented in any physics course, the results will be impacted by the demographics of the cohort. This is, however, true of any physics cohort as well. As was shown in Chapter 4, prior knowledge has significant impact on performance, and the range of prior physics knowledge, especially in an introductory class, may be vast. What this highlights is that qualitative approaches which can categorise and contextualise the variations in cohorts and learning environments are key even for quantitatively focused education research.

Trying a new pedagogical approach is also not a risk-free endeavour. It can only be ethically introduced if ways to identify if it is going wrong are baked into the approach. The Summer School is in an enviable position of having the freedom to explore, with students who are open to a new approach, and has the data and the opportunity to identify issues as they arise. In many cases educational reform is slowed by a lack of supporting data or data that is not conclusive. These methods and the context that they have been used in show that when the noise is removed these kinds of positive results can be found. Education can and should be a field that is data driven, and these results show that it is entirely possible.

Additionally, this approach provides an idea of where quantitative approaches, even in idealised situations, are limited. When considering situations where there are too many variables,

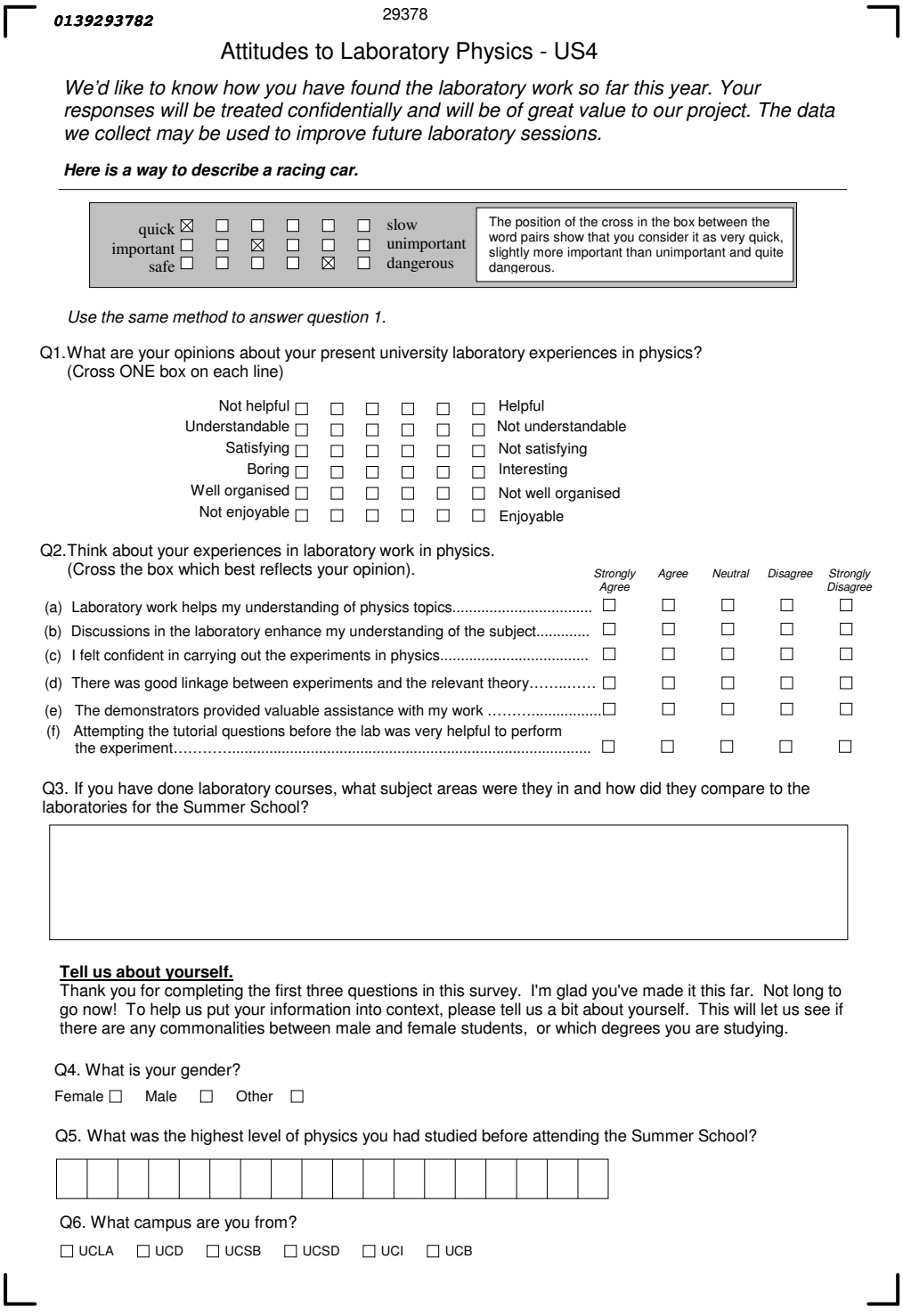
or variables that cannot be usefully codified or categorised, then quantitative approaches are not appropriate. What has been shown throughout this manuscript, however, is that while the attempt should always be made, qualitative approaches must always be used to support and contextualise quantitative methods. Educational data will always be noisy, there will always be more variables, more ways to categorise students or assessment. Qualitative analyses can shine a light on what components of the learning environment are key, directing the focus of quantitative studies. Educational interventions can and should be held to a high statistical standard but the aim should be an interdisciplinary approach which encompasses the best of both qualitative and quantitative approaches.

There is a question to consider with regards to what the goal of learning is. When analysing the metrics that are used within education they can vary quite significantly, many have been shown throughout this manuscript. All of these metrics are important aspects of any learning experience. As was stated in the first chapter of this manuscript, a standard must be outlined before any analysis can begin. An agreement must be made that the tools of measurement are fair and consistent. There are two kinds of analytical results. There are those that objectively measure the situation from outside perspective, observations and analyses of interactions and data. Then there are those that measure the subjective experience of those within the context. Even within this manuscript both of these categories of measurement have been used. What is most important, is that all physics educators have the tools that allow them to understand the goals of their learning environment and introduce the necessary interventions to allow that environment to reach its full potential.

Appendix A

Feedback Surveys

Provided below are the surveys used to receive the feedback analysed in Chapter 7. These surveys were adapted from feedback surveys that had been used for several years with the undergraduate physics cohorts at the University of Glasgow. The design was kept simple so as to avoid overwhelming the students thus garnering a more fair reflection of student attitudes.



0139293782 29378

Attitudes to Laboratory Physics - US4

We'd like to know how you have found the laboratory work so far this year. Your responses will be treated confidentially and will be of great value to our project. The data we collect may be used to improve future laboratory sessions.

Here is a way to describe a racing car.

quick	☒	☐	☐	☐	☐	☐	slow	The position of the cross in the box between the word pairs show that you consider it as very quick, slightly more important than unimportant and quite dangerous.
important	☐	☐	☒	☐	☐	☐	unimportant	
safe	☐	☐	☐	☐	☒	☐	dangerous	

Use the same method to answer question 1.

Q1. What are your opinions about your present university laboratory experiences in physics?
(Cross ONE box on each line)

Not helpful	☐	☐	☐	☐	☐	☐	Helpful
Understandable	☐	☐	☐	☐	☐	☐	Not understandable
Satisfying	☐	☐	☐	☐	☐	☐	Not satisfying
Boring	☐	☐	☐	☐	☐	☐	Interesting
Well organised	☐	☐	☐	☐	☐	☐	Not well organised
Not enjoyable	☐	☐	☐	☐	☐	☐	Enjoyable

Q2. Think about your experiences in laboratory work in physics.
(Cross the box which best reflects your opinion).

	Strongly Agree	*Agree*	*Neutral*	*Disagree*	*Strongly Disagree*
(a) Laboratory work helps my understanding of physics topics.....	☐	☐	☐	☐	☐
(b) Discussions in the laboratory enhance my understanding of the subject.....	☐	☐	☐	☐	☐
(c) I felt confident in carrying out the experiments in physics.....	☐	☐	☐	☐	☐
(d) There was good linkage between experiments and the relevant theory……..….	☐	☐	☐	☐	☐
(e) The demonstrators provided valuable assistance with my work …….....☐	☐	☐	☐	☐	☐
(f) Attempting the tutorial questions before the lab was very helpful to perform the experiment…….....	☐	☐	☐	☐	☐

Q3. If you have done laboratory courses, what subject areas were they in and how did they compare to the laboratories for the Summer School?

Tell us about yourself.
Thank you for completing the first three questions in this survey. I'm glad you've made it this far. Not long to go now! To help us put your information into context, please tell us a bit about yourself. This will let us see if there are any commonalities between male and female students, or which degrees you are studying.

Q4. What is your gender?

Female ☐ Male ☐ Other ☐

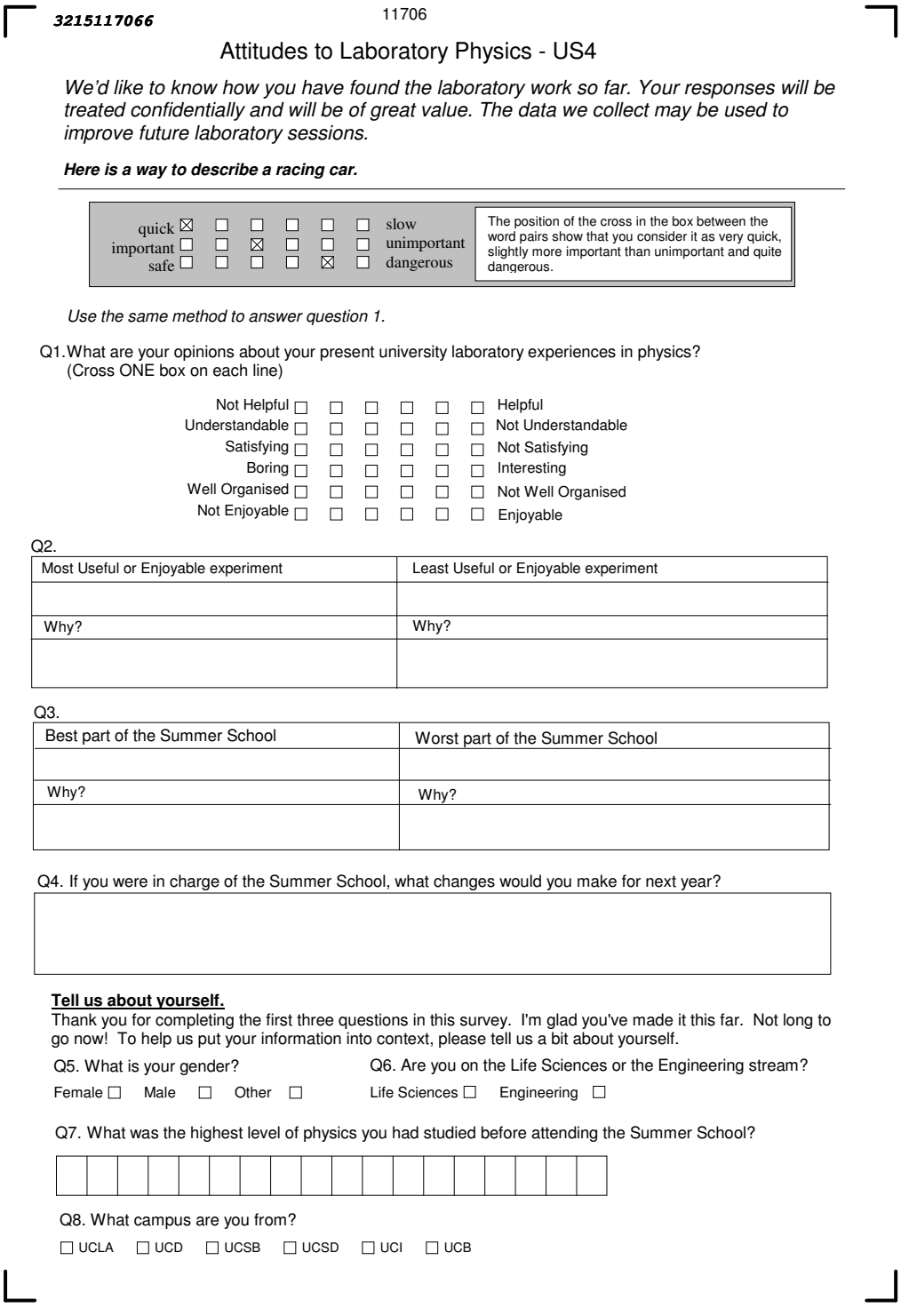
Q5. What was the highest level of physics you had studied before attending the Summer School?

Q6. What campus are you from?

☐ UCLA ☐ UCD ☐ UCSB ☐ UCSD ☐ UCI ☐ UCB

Figure A.1: Laboratory Experience Survey 2017 - Module 1

This survey was based on similar feedback surveys used within the University of Glasgow. Both Likert and semantic differential style questions were used.



3215117066 11706

Attitudes to Laboratory Physics - US4

We'd like to know how you have found the laboratory work so far. Your responses will be treated confidentially and will be of great value. The data we collect may be used to improve future laboratory sessions.

Here is a way to describe a racing car.

quick	☒	☐	☐	☐	☐	☐	slow
important	☐	☐	☒	☐	☐	☐	unimportant
safe	☐	☐	☐	☐	☒	☐	dangerous

The position of the cross in the box between the word pairs show that you consider it as very quick, slightly more important than unimportant and quite dangerous.

Use the same method to answer question 1.

Q1. What are your opinions about your present university laboratory experiences in physics? (Cross ONE box on each line)

Not Helpful	☐	☐	☐	☐	☐	☐	Helpful
Understandable	☐	☐	☐	☐	☐	☐	Not Understandable
Satisfying	☐	☐	☐	☐	☐	☐	Not Satisfying
Boring	☐	☐	☐	☐	☐	☐	Interesting
Well Organised	☐	☐	☐	☐	☐	☐	Not Well Organised
Not Enjoyable	☐	☐	☐	☐	☐	☐	Enjoyable

Q2.

Most Useful or Enjoyable experiment	Least Useful or Enjoyable experiment
Why?	Why?

Q3.

Best part of the Summer School	Worst part of the Summer School
Why?	Why?

Q4. If you were in charge of the Summer School, what changes would you make for next year?

Tell us about yourself.

Thank you for completing the first three questions in this survey. I'm glad you've made it this far. Not long to go now! To help us put your information into context, please tell us a bit about yourself.

Q5. What is your gender?

Female ☐ Male ☐ Other ☐

Q6. Are you on the Life Sciences or the Engineering stream?

Life Sciences ☐ Engineering ☐

Q7. What was the highest level of physics you had studied before attending the Summer School?

Q8. What campus are you from?

☐ UCLA ☐ UCD ☐ UCSB ☐ UCSD ☐ UCI ☐ UCB

Figure A.2: Laboratory Experience Survey 2017 - Module 2

This survey was employed at the end of the Summer School and included questions about the Summer School as a whole.

Bibliography

[AAPT, 1997] AAPT (1997). Goals of the introductory physics laboratory. *The Physics Teacher*, 35(9):546.

[AAPT, 2014] AAPT (2014). AAPT Recommendations for the Undergraduate Physics Laboratory Curriculum. *Report*, page 29.

[Aslan-Tutak and Adams, 2006] Aslan-Tutak, F. and Adams, T. L. (2006). Project-Based Learning in Post-Secondary Education: Theory, Practice and Rubber Sling Shots. *Higher Education*, 51(2):287–314.

[Barker, 1980] Barker, K. N. (1980). Data collection techniques: observation. *American journal of hospital pharmacy*, 37(9):1235.

[Bennett et al., 2010] Bennett, J., Hogarth, S., Lubben, F., Campbell, B., and Robinson, A. (2010). Talking science: The research evidence on the use of small group discussions in science teaching. *International Journal of Science Education*, 32(1):69–95.

[Bhaw and Kriek, 2020] Bhaw, N. and Kriek, J. (2020). A review of the final and supplementary Grade 12 physics examinations from 2014 to 2018 based on a modified Bloom's taxonomy. In *Journal of Physics: Conference Series*, volume 1512. Institute of Physics Publishing.

[Blair et al., 2016] Blair, E., Maharaj, C., and Primus, S. (2016). Performance and perception in the flipped classroom. *Education and Information Technologies*, 21(6):1465–1482.

[Bloom et al., 1956] Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., and Krathwohl, D. R. (1956). The Classification of Educational Goals, Handbook 1 Cognitive Domain. *Taxonomy of educational objectives*, page 207.

[Bodin and Winberg, 2012] Bodin, M. and Winberg, M. (2012). Role of beliefs and emotions in numerical problem solving in university physics education. *Physical Review Special Topics - Physics Education Research*, 8(1):1–14.

[Bouquet et al., 2017] Bouquet, F., Bobroff, J., Fuchs-Gallezot, M., and Maurines, L. (2017). Project-based physics labs using low-cost open-source hardware. *American Journal of Physics*, 85(3):216–222.

[Bynum et al., 2021] Bynum, W. E., Varpio, L., Lagoo, J., and Teunissen, P. W. (2021). 'I'm unworthy of being in this space': The origins of shame in medical students. *Medical Education*, 55(2):185–197.

[Carron and Brawley, 2012] Carron, A. V. and Brawley, L. R. (2012). Cohesion: Conceptual and Measurement Issues Conceptual and Measurement Issues. *Small Group Research*, 43(6):726–743.

[Chan et al., 2014] Chan, P. E., Graham-Day, K. J., Ressa, V. A., Peters, M. T., and Konrad, M. (2014). Beyond Involvement: Promoting Student Ownership of Learning in Classrooms. *Intervention in School and Clinic*, 50(2):105–113.

[Chatman et al., 2008] Chatman, J. A., Boisnier, A. D., Spataro, S. E., Anderson, C., and Berdahl, J. L. (2008). Being distinctive versus being conspicuous: The effects of numeric status and sex-stereotyped tasks on individual performance in groups. *Organizational Behavior and Human Decision Processes*, 107(2):141–160.

[Cotton et al., 2010] Cotton, D. R., Stokes, A., and Cotton, P. A. (2010). Using observational methods to research the student experience. *Journal of Geography in Higher Education*, 34(3):463–473.

[Crouch and Mazur, 2001] Crouch, C. H. and Mazur, E. (2001). Peer Instruction: Ten years of experience and results. *American Journal of Physics*, 69(9):970–977.

[Curşeu and Pluut, 2013] Curşeu, P. L. and Pluut, H. (2013). Student groups as learning entities: The effect of group diversity and teamwork quality on groups' cognitive complexity. *Studies in Higher Education*, 38(1):87–103.

[Day and Bonn, 2011] Day, J. and Bonn, D. (2011). Development of the concise data processing assessment. *Physical Review Special Topics - Physics Education Research*, 7(1):1–14.

[Dewi et al., 2022] Dewi, W. S., Murtiani, Sari, S. Y., and Mairizwan (2022). The validity of physics learning evaluation teaching material based on project-based learning and portfolio assessment. *Journal of Physics: Conference Series*, 2309(1).

[Downing, 2003] Downing, S. M. (2003). Validity: On the meaningful interpretation of assessment data. *Medical Education*, 37(9):830–837.

[Downing, 2004] Downing, S. M. (2004). Reliability: on the reproducibility of assessment data. *Medical Education*, pages 1–8.

[Doğan et al., 2021] Doğan, Y., Batdı, V., and Yaşar, M. D. (2021). Effectiveness of flipped classroom practices in teaching of science: a mixed research synthesis. *Research in Science and Technological Education*, 00(00):1–29.

[Eddy and Brownell, 2016] Eddy, S. L. and Brownell, S. E. (2016). Beneath the numbers: A review of gender disparities in undergraduate education across science, technology, engineering, and math disciplines. *Physical Review Physics Education Research*, 12(2):1–20.

[Etkina and Heuvelen, 2007] Etkina, E. and Heuvelen, A. V. (2007). Investigative Science Learning Environment – A Science Process Approach to Learning Physics. In *Research-based reform of university physics*, pages 1–48.

[Etkina et al., 2010] Etkina, E., Karelina, A., Ruibal-Villasenor, M., Rosengrant, D., Jordan, R., and Hmelo-Silver, C. E. (2010). Design and reflection help students develop scientific abilities: Learning in introductory physics laboratories. *Journal of the Learning Sciences*, 19(1):54–98.

[Foreman and Gubbins, 2015] Foreman, J. L. and Gubbins, E. J. (2015). Teachers See What Ability Scores Cannot: Predicting Student Performance With Challenging Mathematics. *Journal of Advanced Academics*, 26(1):5–23.

[Fraser et al., 2014] Fraser, J. M., Timan, A. L., Miller, K., Dowd, J. E., Tucker, L., and Mazur, E. (2014). Teaching and physics education research: Bridging the gap. *Reports on Progress in Physics*, 77(3).

[Gapp and Fisher, 2012] Gapp, R. and Fisher, R. (2012). Undergraduate management students' perceptions of what makes a successful virtual group. *Education and Training*, 54(2-3):167–179.

[Gijbels et al., 2005] Gijbels, D., Van De Watering, G., Dochy, F., and Van Den Bossche, P. (2005). The relationship between students' approaches to learning and the assessment of learning outcomes. *European Journal of Psychology of Education*, 20(4):327–341.

[Gilboy et al., 2015] Gilboy, M. B., Heinerichs, S., and Pazzaglia, G. (2015). Enhancing student engagement using the flipped classroom. *Journal of Nutrition Education and Behavior*, 47(1):109–114.

[Graham et al., 2017] Graham, M., McLean, J., Read, A., Suchet-Pearson, S., and Viner, V. (2017). Flipping and still learning: experiences of a flipped classroom approach for a third-year undergraduate human geography course. *Journal of Geography in Higher Education*, 41(3):403–417.

[Hake, 2002] Hake, R. (2002). Lessons from the Physics Education Reform Effort Linked references are available on JSTOR for this article : You may need to log in to JSTOR to access the linked references . Lessons from the Physics Education Reform Effort. *Conservation Ecology*, 5(2).

[Halloun and Hestenes, 1985] Halloun, I. A. and Hestenes, D. (1985). Common sense concepts about motion. *American Journal of Physics*, 53(11):1056–1065.

[Haughton, 2009] Haughton, B. (2009). Identifying Diversity and Cohesion in Small Group Interaction. *International Journal of Diversity in Organisations, Communities and Nations*, 8(6):57–66.

[Hestenes et al., 1992] Hestenes, D., Wells, M., and Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher*, 30(3):141–158.

[Hofstein and Lunetta, 2004] Hofstein, A. and Lunetta, V. N. (2004). The Laboratory in Science Education: Foundations for the Twenty-First Century. *Science Education*, 88(1):28–54.

[Holmes et al., 2017] Holmes, N. G., Olsen, J., Thomas, J. L., and Wieman, C. E. (2017). Value added or misattributed? A multi-institution study on the educational benefit of labs for reinforcing physics content. *Physical Review Physics Education Research*, 13(1):1–12.

[Holmes and Wieman, 2018] Holmes, N. G. and Wieman, C. E. (2018). Introductory physics labs: We can do better. *Physics Today*, 71(1):38–45.

[James, Alisa R;Griffin, Linda L;France, 2005] James, Alisa R;Griffin, Linda L;France, T. (2005). PERCEPTIONS OF ASSESSMENT IN ELEMENTARY PHYSICAL EDUCATION: A CASE STUDY. *Physical Educator*, 62(2):85.

[Jandaghi, 2010] Jandaghi, G. (2010). Assessment of validity, reliability and difficulty indices for teacher-built physics exam questions in first year high school. *Educational Research and Reviews*, 5(11):651–654.

[Klein et al., 2017] Klein, P., Müller, A., and Kuhn, J. (2017). Assessment of representational competence in kinematics. *Physical Review Physics Education Research*, 13(1):1–18.

[Kusurkar et al., 2013] Kusurkar, R. A., Ten Cate, T. J., Vos, C. M., Westers, P., and Croiset, G. (2013). How motivation affects academic performance: A structural equation modelling analysis. *Advances in Health Sciences Education*, 18(1):57–69.

[Lamm and Trommsdorff, 1973] Lamm, H. and Trommsdorff, G. (1973). Group versus individual performance on tasks requiring ideational proficiency (brainstorming): A review. *European Journal of Social Psychology*, 3(4):361–388.

[Lin and Singh, 2013] Lin, S.-Y. and Singh, C. (2013). Can free-response questions be approximated by multiple-choice equivalents? *American Journal of Physics*, 81(8):624–629.

[Lindblom-Ylänne et al., 2003] Lindblom-Ylänne, S., Pihlajamäki, H., and Kotkas, T. (2003). What makes a student group successful? Student-student and student-teacher interaction in a problem-based learning environment. *Learning Environments Research*, 6(1):59–76.

[Lizzio et al., 2002] Lizzio, A., Wilson, K., and Simons, R. (2002). University students' perceptions of the learning environment and academic outcomes: Implications for theory and practice. *Studies in Higher Education*, 27(1):27–52.

[López-Zambrano et al., 2020] López-Zambrano, J., Lara, J. A., and Romero, C. (2020). Towards portability of models for predicting students' final performance in university courses starting from moodle logs. *Applied Sciences (Switzerland)*, 10(1).

[Lorenzo et al., 2006] Lorenzo, M., Crouch, C. H., and Mazur, E. (2006). Reducing the gender gap in the physics classroom. *American Journal of Physics*, 74(2):118–122.

[Lorge et al., 1958] Lorge, I., Fox, D., Davitz, J., and Brenner, M. (1958). A survey of studies contrasting the quality of group performance and individual performance, 1920-1957. *Psychological Bulletin*, 55(6):337–372.

[Main and of Physics (Great Britain), 2022] Main, P. C. and of Physics (Great Britain), I. (2022). *Assessment in university physics education*. IOP Publishing, Bristol [England] (Temple Circus, Temple Way, Bristol BS1 6HG, UK).

[Mazur, 1997] Mazur, E. (1997). *Peer instruction: a user's manual*. Prentice Hall, Upper Saddle River, N.J.

[Miller, 2012] Miller, G. (2012). The Smartphone Psychology Manifesto. *Perspectives on Psychological Science*, 7(3):221–237.

[Murata, 2013] Murata, A. (2013). Diversity and High Academic Expectations Without Tracking: Inclusively Responsive Instruction. *Journal of the Learning Sciences*, 22(2):312–335.

[Nicola-Richmond et al., 2018] Nicola-Richmond, K., Pépin, G., Larkin, H., and Taylor, C. (2018). Threshold concepts in higher education: a synthesis of the literature relating to measurement of threshold crossing. *Higher Education Research and Development*, 37(1):101–114.

[Normandeau et al., 2017] Normandeau, M., Iyengar, S., and Newling, B. (2017). The Presence of Gender Disparity on the Force Concept Inventory in a Sample of Canadian Undergraduate Students. *The Canadian Journal for the Scholarship of Teaching and Learning*, 8(1).

[Nuttall et al., 1987] Nuttall, D. L., European, S., June, N., and Nuttall, D. L. (1987). The Validity of Assessments Linked references are available on JSTOR for this article : The Validity of Assessments. *European Journal of Psychology of Education*, 2(2):109–118.

[Olejnik and Algina, 2000] Olejnik, S. and Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology*, 25(3):241–286.

[Prades and Espinar, 2010] Prades, A. and Espinar, S. R. (2010). Laboratory assessment in chemistry: An analysis of the adequacy of the assessment process. *Assessment and Evaluation in Higher Education*, 35(4):449–461.

[Qazdar et al., 2019] Qazdar, A., Er-Raha, B., Cherkaoui, C., and Mammass, D. (2019). A machine learning algorithm framework for predicting students performance: A case study of baccalaureate students in Morocco. *Education and Information Technologies*, 24(6):3577–3589.

[Qu et al., 2019] Qu, S., Li, K., Wu, B., Zhang, S., and Wang, Y. (2019). Predicting student achievement based on temporal learning behavior in MOOCs. *Applied Sciences (Switzerland)*, 9(24).

[Riegle-Crumb and Moore, 2014] Riegle-Crumb, C. and Moore, C. (2014). The gender gap in high school physics: Considering the context of local communities. *Social Science Quarterly*, 95(1):253–268.

[Roth et al., 2010] Roth, W.-M., Lee, Y.-J., and Hsu, P.-L. (2010). The Group Cohesion Scale-Revised: Reliability and Validity. *Compendium Newtown Pa*, 17(3):147–56.

[Rusk and Rønning, 2020] Rusk, F. and Rønning, W. (2020). Group work as an arena for learning in stem education: negotiations of epistemic relationships. *Education Inquiry*, 11(1):36–53.

[Scott et al., 2006] Scott, M., Stelzer, T., and Gladding, G. (2006). Evaluating multiple-choice exams in large introductory physics courses. *Physical Review Special Topics - Physics Education Research*, 2(2):1–14.

[Seyranian et al., 2018] Seyranian, V., Madva, A., Duong, N., Abramzon, N., Tibbetts, Y., and Harackiewicz, J. M. (2018). The longitudinal effects of stem identity and gender on flourishing and achievement in college physics. *International Journal of STEM Education*, 5(1):40–14.

[Simmons and Heckler, 2020] Simmons, A. B. and Heckler, A. F. (2020). Grades, grade component weighting, and demographic disparities in introductory physics. *Physical Review Physics Education Research*, 16(2):20125.

[Sivarajah et al., 2018] Sivarajah, S., Smith, S. M., and Thomas, S. C. (2018). Tree cover and species composition effects on academic performance of primary school students. *PLoS ONE*, 13(2):1–11.

[Sobhanzadeh et al., 2017] Sobhanzadeh, M., Kalman, C. S., and Thompson, R. I. (2017). Labatorials in introductory physics courses Labatorials in introductory physics courses. *European Journal of Physics*, 38(6):65–70.

[Springer and Stanne, 1999] Springer, L. and Stanne, M. E. (1999). Effects of Small-Group Learning on Undergraduates in Science , Mathematics , Engineering , and Technology : A Meta-Analysis Author (s): Leonard Springer , Mary Elizabeth Stanne and Samuel S . Donovan Published by : American Educational Research Associat. *American Educational Research Assocaition*, 69(1):21–51.

[Steinmayr et al., 2019] Steinmayr, R., Weidinger, A. F., Schwinger, M., and Spinath, B. (2019). The importance of students' motivation for their academic achievement-replicating and extending previous findings. *Frontiers in Psychology*, 10(JULY).

[Stöhr et al., 2020] Stöhr, C., Demazière, C., and Adawi, T. (2020). The polarizing effect of the online flipped classroom. *Computers and Education*, 147(December 2019).

[Street et al., 2015] Street, S. E., Gilliland, K. O., McNeil, C., and Royal, K. (2015). The Flipped Classroom Improved Medical Student Performance and Satisfaction in a Pre-clinical Physiology Course. *Medical Science Educator*, 25(1):35–43.

[Szott, 2014] Szott, A. (2014). Open-ended Laboratory Investigations in a High School Physics Course: The difficulties and rewards of implementing inquiry-based learning in a physics lab. *The Physics Teacher*, 52(1):17–21.

[Warner et al., 2012] Warner, S., Bowers, M. T., and Dixon, M. A. (2012). Team Dynamics: A Social Network Perspective. *Journal of Sport Management*, 26(1):53–66.

[Watty et al., 2010] Watty, K., Jackson, M., and Yu, X. (2010). Students' Approaches to assessment in accounting education: The unique student perspective. *Accounting Education*, 19(3):219–234.

[Whitcomb and Singh, 2020] Whitcomb, K. M. and Singh, C. (2020). For physics majors, gender differences in introductory physics do not inform future physics performance. *European journal of physics*, 41(6):65701.

[Wieman, 2015] Wieman, C. (2015). Comparative Cognitive Task Analyses of Experimental Science and Instructional Laboratory Courses. *The Physics Teacher*, 53(6):349–351.

[Wieman and Holmes, 2015] Wieman, C. and Holmes, N. G. (2015). Measuring the impact of an instructional laboratory on the learning of introductory physics. *American Journal of Physics*, 83(11):972–978.

[Wilcox and Lewandowski, 2016] Wilcox, B. R. and Lewandowski, H. J. (2016). Open-ended versus guided laboratory activities: Impact on students' beliefs about experimental physics. *Physical Review Physics Education Research*, 12(2):1–8.

[Wilcox and Lewandowski, 2017] Wilcox, B. R. and Lewandowski, H. J. (2017). Developing skills versus reinforcing concepts in physics labs: Insight from a survey of students' beliefs about experimental physics. *Physical Review Physics Education Research*, 13(1):1–9.

[Williams and Pence, 2011] Williams, A. J. and Pence, H. E. (2011). Smart phones, a powerful tool in the chemistry classroom. *Journal of Chemical Education*, 88(6):683–686.

[Williams et al., 2017] Williams, J. L., Miller, M. E., Avitabile, B. C., Burrow, D. L., Schmittou, A. N., Mann, M. K., and Hiatt, L. A. (2017). Teaching Students To Be Instrumental in Analysis: Peer-Led Team Learning in the Instrumental Laboratory. *Journal of Chemical Education*, 94(12):1889–1895.

[Zaidi et al., 2018] Zaidi, N. L., Grob, K. L., Monrad, S. M., Kurtz, J. B., Tai, A., Ahmed, A. Z., Gruppen, L. D., and Santen, S. A. (2018). Pushing Critical Thinking Skills with Multiple-Choice Questions: Does Bloom's Taxonomy Work? *Academic Medicine*, 93(6):856–859.