The Assessment of Executive Function and Fluid Intelligence in Traumatic Brain Injury

Greg Halliday, MSc, B.A. Hons

Submitted in partial fulfilment of the requirements for the degree of
Doctorate in Clinical Psychology

School of Health and Wellbeing
College of Medical, Veterinary and Life Sciences
University of Glasgow

July 2023

Table of contents

List of Tables

List of Figures

Acknowledgements

Few significant milestones are the work of a sole individual, and the completion of a doctorate is no exception. There are countless people to whom I owe an acknowledgement of their contribution towards this achievement, so to everyone who has provided me with support and guidance throughout my life, I offer you my sincere thanks. I would like to start by thanking my academic supervisor Dr Jessica Fish, who provided me with patience, wisdom and reassurance, as well as her vast knowledge of neuropsychology. Thank you also, to Dr Breda Cullen, my university advisor for your compassion and kindness through the hardest parts of this endeavour. Thank you to Dr Jala Risaq for your stats knowledge, and to Tracey McKee for your Systematic Review wisdom.Thank you to my family, particularly my Mum and Dad (Kenny and Seonaid), who have given me everything you had in life (and more). Through your sacrifices, and teaching, you have provided me with the opportunity to achieve, and have been examples of the importance of hard work. I owe everything to you. To my partner, Jenna, who has given endless support, cheerleading, and opportunities to complain, thank you for always being there, for proof-reading and for not letting me give up. Thank you also to the Wine and Cheese crew, for your solidarity throughout this difficult career journey. Special shout outs to my friends, the Steve Clarke Fan Club and the Philm Club for allowing me to have a life outside of work, and always giving encouragement. Thank you also to my friends and colleagues on the DClinPsy for your solidarity and companionship. Finally, to my Auntie Dianne, the most selfless person I have ever known, who exemplified putting others first, thank you for everything.

Chapter 1

The psychometric properties of the Zoo Map Test in traumatic brain injury: a systematic review

Word Count: 8606

Prepared in accordance with the author requirements for PlosOne: https://journals.plos.org/plosone/s/submission-guidelines

Abstract

Background: Traumatic Brain Injury (TBI) has a range of short, and long-term consequences which may include significant cognitive deficits. The Zoo Map Test (ZM), a subtest of the Behavioural Assessment of the Dysexecutive Syndrome (BADS), is one of many tools used to assess executive function, a cognitive domain of particular functional relevance. This systematic review aims to examine the psychometric properties of the ZM, to provide recommendations regarding its use for clinical and research purposes within the TBI population.

Methods: Six electronic databases were searched in May 2023. This review followed COSMIN guidance and assessed risk of bias using the COSMIN risk of bias checklist. Results from measurement properties were collated and synthesised to provide a summary of evidence, and provided a categorisation for quality of evidence using a modified GRADE approach.

Results: Eight studies with 497 participants with TBI were identified as meeting inclusion criteria (studies published in English, assessing psychometric properties of ZM in a sample of adults with moderate-severe TBI, where injury severity is reported). None of these studies reported on content, structural, cross-cultural or criterion validity, internal consistency, reliability or measurement error. The evidence for convergent validity could not be determined due to inconsistency. The evidence for ecological validity was considered insufficient, but high quality. Evidence for responsiveness to change was considered insufficient and of low quality. Evidence for known-group validity was found to be sufficient and quality of evidence rated as moderate.

Conclusions: There is some support for using the ZM as a measure in TBI. Concerns relating to ecological and construct validity are discussed. Further research is needed to specifically examine the psychometric properties of the ZM to provide further evidence regarding its soundness for clinical and research purposes when used as an isolated tool, distinct from the broader BADS battery.

Keywords: neuropsychological assessment, behavioural assessment of the dysexecutive syndrome, executive function, psychometric properties, systematic review

Introduction

Background

Acquired brain injury (ABI) is a term that describes a variety of injuries causing damage to the brain and which impair, in some way, the brain's ability to perform its necessary functions to its usual standards. These injuries may result from a variety of mechanisms, including: through infections in the brain or through interruption of oxygen supply to the brain or as a result of a "mechanical force injury", more frequently referred to as a Traumatic Brain Injury (TBI; Elbaum & Benson, 2007). TBIs are often categorised according to their severity, with mild injuries being distinguished from those of moderate or greater severity (Malec et al., 2007). Though mild TBI can cause lasting symptoms, most people make good recoveries without specific treatment (Turner-Stokes et al., 2015). By contrast, moderate-severe TBI is consistently associated with long term disability and people with more severe injuries are likely to require specialist rehabilitation (Turner-Stokes et al., 2015).

Frequent symptoms of TBI include physical and emotional difficulties, alongside cognitive sequelae. Some of the most frequent cognitive impacts are in the domains of attention, memory, executive function and language (Stuss, 2011). Patients also often present with difficulties related to interpersonal functioning (Temkin et al. 2009). Males are twice as likely as females to experience a TBI and incidence varies significantly by age, with the highest incidence in children aged below 4, and adolescents aged 15-19 years (Langlois et al., 2006). An estimated 1.3 million people in the UK live with symptoms of TBI (All-Party Parliamentary Group on Acquired Brain Injury, 2018). As incidence of TBI continues to increase, and as we continue to develop our understanding of how to support and treat those who have experienced head injury in the acute stages, the mortality rate for severe TBI is decreasing. As a result, there is a growing population living with long-term symptoms of TBI, which may cause significant disability (Galgano et al., 2017). TBI is therefore an emergent health concern, and as such, the British government is developing a health and social care strategy to support those impacted by ABI; including those who have experienced TBI.

In order to assess the impact of TBI on cognition, and to obtain information about the person's likely level of long-term disability, healthcare services routinely incorporate neuropsychological assessment (Andrews, 2005). Neuropsychological assessment allows for a systematic examination of brain functioning for a variety of medical and neurodevelopmental difficulties. It frequently involves the use of standardised assessment batteries aimed at testing various areas of cognitive functioning (Catanese, 2007). These batteries usually consist of assessment tools, which allow clinicians to identify areas where functioning is impaired or reduced from the expected level, and can allow for the development of a profile of a patient's cognitive strengths and vulnerabilities (Stebbins, 2007).

Moderate-severe TBI is associated with difficulties with emotion regulation and mood, inhibitory control, planning, social cognition, and insight into acquired difficulties (McDonald & Genova, 2021). Previous research has also highlighted that those who have experienced more severe head injuries are more likely to experience a greater decline in life satisfaction than those with mild injuries (Caplan et al., 2016). In order to provide increased support for

this particularly vulnerable population, it is important that clinicians are able to accurately identify the impact of TBI on brain functioning, and how this may impact on daily life.

Assessment of Executive Function

An important aspect of neuropsychological assessment post-TBI is the examination of executive function. Executive functions are the high-level cognitive processes we use to facilitate our abilities to problem solve through developing new behaviour, particularly in response to novel stimuli. These processes are thought to be, in part, supported by structures within the frontal lobes (Gilbert & Burgess, 2008). The assessment of executive function (EF) in TBI is inherently complex due to the heterogeneous nature of TBI presentations, as impairments may vary depending on the nature, location, and severity of the injury. Cognitive assessment may also be complicated by additional physical/sensory disability (Ponsford, 1995). The nature of executive function complicates these matters further, with significant debate surrounding whether the areas of the brain associated with EF function autonomously as a single area with unified purpose, or as a fractionated system, responsible for heterogeneous functions (Stuss, 2011).

Previous research has highlighted that some commonly used tools associated with executive function, such as the Stroop test (Stroop, 1935) and the Wisconsin Card-Sorting test (Grant & Berg, 1948), have questionable predictive and ecological validity (Burgess et al., 2006). It has been argued that performance on these tests is not representative of one's ability to perform daily tasks, which may be impaired by compromised executive function. Similarly, it has also been suggested that these tests may fail to accurately discriminate between clinical and non-clinical samples (George & Gilbert, 2018; Newstead et al., 2018). One potential explanation may be that neuropsychological assessments are usually conducted in a standardised, distraction-free environment to maximise performance and support patients to remain engaged with the task at hand. It has been suggested that the structured testing environment that many traditional neuropsychological tests of executive function require may influence the expression of executive dysfunction in patients (i.e. mask their deficits), and performance is therefore not representative of the true effects of executive function deficits on daily living (Gioia & Isquith, 2004).

The Zoo Map Test

In response to the assessment issues outlined above, Wilson (1993) considered the extent to which traditional neuropsychological assessments were able to predict daily functioning and called for new procedures which applied to real life functioning. This led to the development of measures of executive function which were believed to be more ecologically valid and which more accurately captured symptoms associated with executive deficits, such as the Multiple Errands Test (Shallice & Burgess, 1996). Previous research has supported the premise that the predictive and ecological validity of these types of tests makes them more sensitive than traditional neuropsychological measures to elicit and detect executive failures, as well as in predicting behavioural problems in daily life (Alderman et al., 2006).

The Behavioural Assessment of the Dysexecutive Syndrome (BADS) [Wilson et al., 1996] was developed to provide an ecologically valid assessment of executive functioning, with an assessment battery which requires participants to plan, initiate, monitor, and adjust

behaviour in response to the explicit and implicit demands of the tasks involved. The BADS is designed to assess for characteristics of the "dysexecutive syndrome"; a constellation of varied cognitive and behavioural symptoms that can be disabling in everyday life. As part of the assessment process, patients' ability to plan is assessed using a subtest called the Zoo Map Test.

The Zoo Map Test asks patients to describe how they would plan a visit to a series of locations within a zoo, ensuring that certain rules are obeyed. The Zoo Map examines complex planning and strategy, by requiring participants to draw a route they would follow to visit all animals presented on the list provided. In the first condition, patients are asked to visit each of the animals instructed. The second condition requires planning of a route to visit the animals in a prescribed order. As an assessment tool regularly used within clinical practice and within research, the psychometric properties of the BADS have been examined in several studies. Norris & Tate (2000) examined the psychometric properties of the BADS in a sample of 36 ABI patients, and highlighted that the BADS and most of its subtests correlate significantly with other standard tests of executive function, suggesting that the BADS as an assessment battery has acceptable concurrent validity. They also highlighted that the Zoo Map correlated significantly with other executive measures which assessed planning behaviour and problem solving, displaying the concurrent validity of this subtest. They argued that the Zoo Map was a particularly useful tool in discriminating between healthy controls and those with ABI, indicating that the construct validity of the Zoo Map is comparable to other standard executive tests. Similarly, Emmanouel et al. (2014) investigated the validity of BADS subtests in order to highlight whether they successfully discriminated between healthy controls and participants with brain injury, and found significant group differences between ABI patients and healthy controls on almost every BADS executive variable. They noted that the Zoo Map Test adequately discriminated between ABI patients and healthy controls, but could not be used to localise the brain injury. The use of the Zoo Map subtest as an ecologically valid test of executive function, separated from the BADS, has become more common in research studies. There has, however, yet to be a systematic review specifically examining the utility of the Zoo Map Test as a measure of executive function in relation to specific patient groups.

Aims

The aim of the current study is to systematically review, critically appraise, compare, and summarise the quality of evidence regarding the psychometric properties of the Zoo Map Test for use in a moderate-severe TBI population.

Methods

This review adhered to the Consensus-based Standards for the Selection of Health Measurement Instruments (COSMIN) guidelines for evaluating studies on measurement properties of patient-reported outcome measures (PROMS) [Mokkink et al., 2018]. COSMIN guidelines were originally developed to evaluate which measure is most appropriate from a range of possible tools, however, the guidelines may also be used to evaluate the psychometric properties of a single clinician reported or performance-based outcome measure (such as the Zoo Map Test). The study protocol outlining the aims and methods for the review was registered with the Prospective Register of Systematic Reviews (PROSPERO) in May 2023 (reference number: 413806).

Literature Search

Initial searches were completed to determine whether any existing similar reviews had already been registered, using PROSPERO, and the Cochrane Database of Systematic Reviews (CDSR). Relevant articles to be included in the review were then identified using a systematic search of the following databases: Medline, EmBase, Psycinfo, CINAHL, Web of Science core collection and The British Library EThOS database. The search considered only papers published from 1996, the year of publication for Wilson et al.'s (1996) initial article presenting the BADS battery (including the Zoo Map Test).

The initial search strategy included a combination of the terms "Zoo Map", "brain injury" and "neuropsychological assessment". Consultation with a librarian trained in systematic review methodology highlighted that the use of three search strings limited the sensitivity of the search, producing too few results and potentially excluding relevant studies. It was agreed that to maximise sensitivity, the final search strategy would include only terms to identify studies which used the Zoo Map Test as a measure. This would increase the total number of papers being screened, but would reduce the likelihood of potentially relevant studies being missed. Search strategies were then adapted for each database to reflect their functionality (specifically, the search of the British Library e-thesis online service EThOS database excluded the "BADS" term due to the library search terms including the root of included words, which resulted in an excessive number of irrelevant results). A detailed description of different iterations of the search strategy can be found in Appendix 1. Reference lists of included studies were checked (i.e. hand-searched) to identify any other potentially relevant literature.

The following criteria were applied:

Eligibility Criteria

- Original studies published in peer-reviewed journals and unpublished PhD theses.
- Studies that examine the Zoo Map Test in an adult population with moderate-severe TBI.
- Articles that provide psychometric data on the Zoo Map Test, relevant to at least one of the following: reliability, content validity, convergent validity, construct validity, discriminant validity, known-group validity and responsiveness to change.
- Intervention studies using the Zoo Map Test as an outcome measure were used as evidence of responsiveness to change.
- Articles published in English.
- Participants' age >18 years.
- Participants with moderate-severe TBI.

Exclusion criteria

- Conference abstracts and non-peer-reviewed published evidence.
- Participants included those with ABI of various aetiologies and did not report TBI data separately.
- Studies which did not report on severity of TBI, or that included mild TBI (mTBI), but did not report these data separately.
- Review articles (e.g., systematic review, narrative review).

- Articles written in any language other than English.

The results of the systematic search were exported to EndNote (Clarivate, 2013) reference management software, and duplicates were removed. Two rounds of screening took place, with a sample (20%) from each round reviewed by a second reviewer. The first stage of screening examined the title and abstracts of articles to exclude those which clearly did not include the relevant psychometric measure, study design, or participant group. The second round of screening included examination of the full text. Inter-rater agreement on study selection was k= 0.740, with any disagreements resolved through discussion. No third reviewer was required to resolve consensus disputes.

Data extraction

Data were extracted by the first author. A data extraction form was developed that included information about the studies: author, year of publication, country, study design, sample size, and descriptive information for the patient and control groups where applicable (number of participants, severity of TBI, age, and gender). Measurement properties assessed by each study and comparator measure used (where applicable) were also recorded.

Risk of Bias

Following data extraction, all studies were assessed by two reviewers for risk of bias using the COSMIN risk of bias checklist (Mokkink et al., 2018; Terwee et al., 2018). The COSMIN checklist was developed for studies on health-related patient-oriented outcomes, to allow for analysis of study quality when investigating psychometric qualities of measures. The COSMIN risk of bias checklist comprises 10 items, assessing the methodology for different measurement properties (details displayed in Table 1.1). Items are rated as “very good”, “adequate”, “doubtful”, or “inadequate” based on the COSMIN criteria, utilising a “worst score” approach to identify the overall quality of the methods on each measurement criteria (a methodological quality score per measurement property is obtained by taking the lowest rating of any of the items in a box). The COSMIN approach allows for a modular assessment of study quality; therefore, all relevant boxes are used independently for studies assessing more than one measurement property. Convergent and ecological validity (Box 9) is measured separately for each comparator measure analysed within a study. Evidence for validity was then considered separately if different measures within the same study were assessed as having different quality. For example, if a study is assessed as having “very good” statistical analysis and two comparator measures were assessed as “very good” and “doubtful”, the study was considered as two separate studies and by applying the COSMIN recommended “worst score counts” approach, a score of “very good” and “doubtful” overall quality respectively would be applied.

Table 1.1: Categories and corresponding Boxes within COSMIN Risk of Bias tool

Category of measurement	***Associated Boxes from COSMIN Risk of Bias tool***
PROM development/content validity	1 & 2
Structural validity	3
Internal consistency	4
Cross-cultural validity	5
Reliability	6
Measurement error	7
Criterion validity	8
*Construct validity**	9
Responsiveness	10

*Construct validity measurement within Cosmin Risk of Bias tool includes convergent, and known-group validity

Evaluation of Measurement Properties

The results of each study on a measurement property were then rated by the first author as sufficient (+), insufficient (–), or indeterminate (?) according to COSMIN's criteria for good measurement properties (Mokkink et al., 2018; Terwee et al., 2018). In line with this guideline's recommendations, evidence-based hypotheses regarding the magnitude and direction of expected results were applied in circumstances where these had not been predefined by the study authors.

Synthesis

In line with the guidance provided in the user manual for the COSMIN methodology for systematic reviews of patient-reported outcome measures (Mokkink et al., 2018), the nature of the existing evidence on measurement properties of the Zoo Map Test was summarised by the primary reviewer by providing a rating of sufficient (+), insufficient (-), or indeterminate (?) if at least 75% of the total studies results were rated as such for the individual measurement property. In studies where convergent validity was measured, only correlations with measures of cognition were considered as relevant to the purpose of this review and included within data synthesis. To enable consistent analysis of the synthesised data, correlations for relevant outcomes were reported, and effect sizes were calculated or converted to Cohen's d for studies examining known-group validity and responsiveness to change. Ecological validity was considered an additional psychometric property, and was assessed using the same criteria as convergent validity, to provide information relevant to the applicability of Zoo Map results when predicting real-world functional ability.

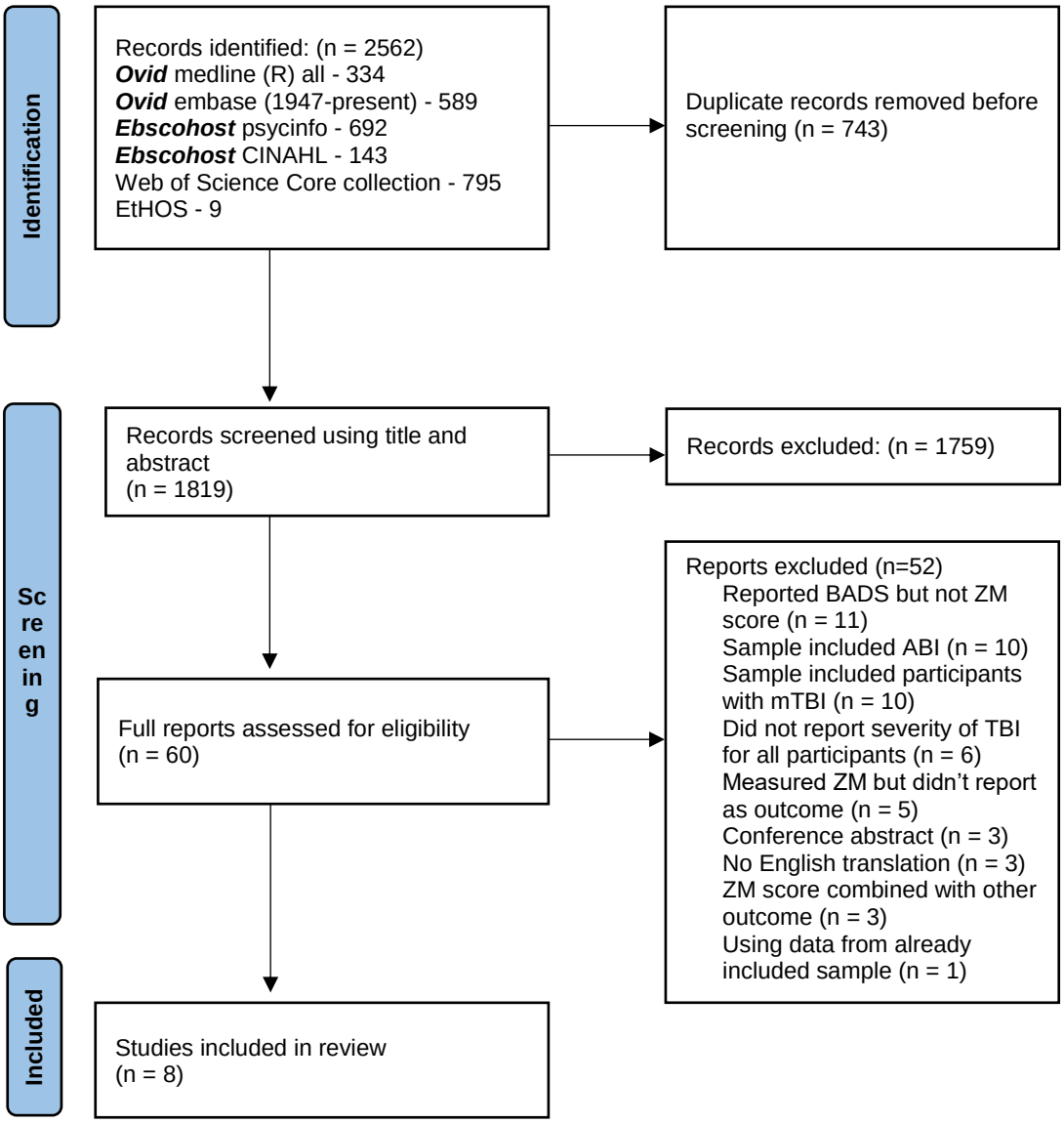
The primary reviewer graded the quality of summarised evidence using the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach recommended within the COSMIN guidelines. This approach categorises the overall quality of evidence as it pertains to a measurement property as: "high", "moderate", "low", or "very low" based on the overall risk of bias, (in)consistency of results, precision (in relation to sample size), and directness.

Results

Study Selection

The search identified a total of 2562 records. After duplicates were removed, 1819 studies were screened for eligibility, with 60 accessed for full-text screening. Of the 52 decisions to exclude studies at the full-text stage, 6 were of particular note. One study (Weddell & Wood, 2018) was excluded as the sample duplicated that of another study already included within the review (Weddell & Wood, 2016). The 2016 study was more relevant to the aims of the current review than the 2018 study, thus the 2018 study was excluded. Three studies used the Zoo Map as a measure, but combined scores with another measure to provide a composite, without reporting Zoo Map data individually, and were therefore excluded (Krpan et al., 2013; Lamberts, Evans & Spikman, 2011; Ponsford et al., 2017). One study (Fitzgerald et al., 2017) met all criteria, but was excluded as it did not provide data related to injury severity for one of their eleven participants, and did not report data for the group, excluding this participant separately. Wilson et al.'s (1998) original study investigating the validity of the BADS was excluded as the sample included participants with both ABI and TBI, and TBI data was not reported separately. A final total of eight studies were included in this review. The study selection process is illustrated in Figure 1.

Figure 1.1: Prisma diagram illustrating search process and outcomes of each screening stage



Study Characteristics

Details of the eight included studies, including study design, participant details, and relevant measurement properties are highlighted in Table 2. Only two of the included studies (Wood & Liossi, 2006; Wood & Liossi, 2007) aimed to specifically evaluate the psychometric properties of the Zoo Map Test. Wood & Liossi (2006) examined the ecological validity of the Zoo Map and 3 other tests of EF compared to the Dysexecutive Questionnaire from the BADS, whilst their 2007 study examined the relationship between tests of executive function (including the Zoo Map), and the concept of *g*, also known as general intelligence (Spearman, 1904).

Seven of the included studies reported the mean Zoo Map profile scores for their sample, and three of these compared this with a control group. Ownsworth et al. (2017) reported

mean scores for version 1 of the Zoo Map Test. All 8 included studies used Glasgow Coma Scale (GCS) scores to determine injury severity within their inclusion criteria, but only 7/8 reported these scores. All studies reported post traumatic amnesia data. All but one study considered the time since injury at which participants were assessed for their study (Wood & Liossi, 2006). Rakers et al. (2018) did not report individual or mean time since injury for the sample, though highlighted that all participants had been injured more than three months prior to data collection. The evidence per measurement property is detailed below and a summary can be found in Table 1.2. Further information regarding the details of the methodological appraisal, study results, and criteria ratings can be found in Appendices 2 and 3.

None of the studies reported data on content validity, and though Wilson et al.'s (1993, 1996, 1998) BADS studies do assess content validity, they do not report TBI data separately. Content validity of the Zoo Map Test was not therefore assessed as part of this review. Similarly, none of the included studies reported on structural validity, internal consistency, cross-cultural validity, reliability, measurement error, or criterion validity. Results are therefore presented for convergent validity, ecological validity, known-group validity, and responsiveness to change (details of these measurement properties are reported in Appendix 4). A summary of GRADE evidence scores is provided in Appendix 5.

Convergent Validity

Four studies reported data on convergent validity (Spikman et al., 2012; Weddell & Wood, 2016; Wood and Liossi, 2006; Wood & Liossi, 2007). Spikman et al. (2012) examined correlations between the Zoo Map and other tests of executive function with tests associated with social cognition and emotional empathy. In line with COSMIN guidance, the comparison instruments were assessed as individual PROMs, however, as the authors did not provide information related to the psychometric measurement properties of any comparison proms, a combined risk of bias was provided based on the overall study design. Weddell and Wood (2016) examined the association between the Zoo Map profile score and self-reported personality change, and six other psychometric assessment tools. Four of these tools were deemed relevant to the purposes of this review, though only two were relevant to convergent validity, with two tests (Dysexecutive Questionnaire [Burgess et al., 1998] and a custom measure referred to as the Frontal and Social Behaviour Scale) considered to be more relevant to ecological validity. The psychometric properties of the remaining two scales (Barratt Impulsiveness Scale [Patton et al., 1998] and STAXI-II Anger Expression Index [Spielberger et al., 1983]) are not established within a TBI population, and no information was provided by the study authors relating to their psychometric properties. A combined risk of bias was therefore provided based on the overall study design.

Wood and Liossi (2006) examined correlations between the Zoo Map and 10 other measures of executive function. Eight of these measures were considered relevant to convergent validity, with two (self-report and other report versions of the Dysexecutive Questionnaire) considered to be more relevant to ecological validity. All measures included in these analyses have previously established good psychometric properties within a TBI population (as noted by study author or identified by the review team). This study was therefore provided with a combined risk of bias rating. Similarly, Wood and Liossi (2007) examined correlations between a group of executive tests (10 subscales), as well as directly

comparing the Zoo Map to traditional measures of executive function. All 10 subscales included in this analysis were considered to provide evidence for convergent validity of the Zoo Map. This study also examined the association between these tests after controlling for fluid intelligence. As in their previous study, the measures all have good psychometric properties in brain injury, and have therefore been provided with a combined risk of bias rating based on the overall study design as recommended by COSMIN guidance (see Appendix 2). All relevant Zoo Map correlations for the above 4 studies are reported in Appendix 3 (convergent validity), which details the individual rating per psychometric property.

Of the four studies that provided evidence relating to convergent validity, two were rated as "very good" (Wood & Liossi, 2006; Wood & Liossi, 2007), and two of "doubtful" quality (Spikman et al., 2012; Weddell & Wood, 2016). Of the hypotheses generated either by the original study authors, or as part of this review, seventeen of thirty-three (52%) were confirmed. When separating the studies of "very good" quality from those of "doubtful" quality, fifteen of the seventeen confirmed hypotheses are attributed to the studies of a higher quality, raising the percentage of confirmed hypotheses to 65%. Even when accounting for study design, the evidence for convergent validity did not meet the pre-determined threshold of 75% confirmed hypotheses, and was therefore rated as inconsistent. One explanation for the apparent inconsistency in results could be related to the issues raised by Wood and Liossi (2007), indicating that general intellectual ability could be accounting for significant variance in executive function scores, though it is highlighted that this does not account for the full variance of performance on these tests. As this hypothesis has not been comprehensively resolved within the literature, the overall quality of evidence must be rated as "could not be determined" according to COSMIN guidance.

Ecological Validity

Three studies reported data on ecological validity (Hendry et al., 2016; Weddell & Wood, 2016; Wood & Liossi, 2006). Hendry et al. (2016) examined correlations between the Zoo Map and a home-based cooking task (HBCT; Chevignard et al., 2009), comparing each HBCT subscale measure to Zoo Map profile scores, and Zoo Map version 1 scores. This analysis was considered to represent ecological validity due to the functional and ecological nature of the comparison task. The psychometric properties of the HBCT were briefly reported by Hendry et al., (2016) noting that previous research highlighted evidence of discriminative, convergent, and concurrent validity, and reported inter-rater reliability for four of the HBCT subtest scores within the HBCT. Associations between HBCT performance overall and the subscores representing component tasks/contributions of more specific cognitive functions were examined to assess the associations between the HBCT and other cognitive test performance. In line with COSMIN guidance, individual HBCT subscores, as well as total error score were assessed as individual PROMs, but provided a combined risk of bias rating based on the overall study design.

Weddell and Wood (2016) examined the relationship between the Zoo Map and other cognitive tests considered to have ecological validity due to the nature of their assessment methodology. The psychometric properties of the Dysexecutive Questionnaire (DEX) are well-known and have already been established within a TBI population, whilst the authors provide no information relating to the psychometric properties of the Frontal and Social

Behaviour Scale or its relevance to this population group. In line with COSMIN guidance, these analyses are considered to be separate PROMs, and were therefore treated as individual studies and provided individual risk of bias ratings.

Wood and Liossi (2016) also examined the relationship between the Zoo Map Test and the DEX providing analyses related to two versions of the DEX (self and other). These analyses were considered as separate PROMs but provided with a combined risk of bias rating. All relevant Zoo Map correlations for the above three studies are reported in Appendix 3 (ecological validity), which details the risk of bias rating per psychometric property.

Of the four sources of evidence which provided evidence relating to the ecological validity of the Zoo Map Test, two were "very good" quality (Weddell & Wood, 2016; Wood & Liossi, 2006), one was rated as "adequate" (Hendry et al., 2016) and one of "doubtful" quality (Weddell & Wood, 2016). Of the thirteen hypotheses generated by these studies, relating to ecological validity, only five (39%) were confirmed. The evidence for the Zoo Map Test's ecological validity quality is considered insufficient. The quality of this evidence was rated as high due to the high quality of included studies, with minimal inconsistency (only one hypothesis generated from the study was considered to be of "doubtful" quality, and there was no evidence of imprecision or indirectness).

Known-Group Validity

Three studies reported on known-group validity (Rakers et al., 2018; Spikman et al., 2012; Westerhof-Evers et al., 2019) of which, one was rated to have "very good" methodological quality (Spikman et al., 2012), whilst one (Rakers et al., 2018) had "adequate" methodological quality, and the other "doubtful" quality (Westerhof-Evers et al., 2019). Two of three (66%) hypotheses related to known-group validity were accepted. However, when removing the "doubtful" quality study, 100% of hypotheses were accepted. The evidence for known-group validity for the Zoo Map was therefore rated as sufficient, and the quality of evidence graded as moderate (one point downgraded for inconsistency).

Responsiveness to Change

One study reported on responsiveness to change (Ownsworth et al., 2017). This study examined the effects of two theory-based interventions (error-based learning and errorless learning) on skill improvement, self-awareness, behavioural competency, and psychosocial functioning in a sample of participants with TBI. The study quality was rated as "adequate" and the hypothesis tested relating to responsiveness to change was rejected. The evidence for the Zoo Map's responsiveness to change was therefore rated as insufficient, and the quality of this evidence rated as low (one point downgrade for risk of bias, one point downgrade for imprecision).

Table 1.2 Characteristics of studies included in SR

Author (Year)	*Country*	*Study design*	*Measurement properties*	*Zoo Map measurement*	*TBI Severity Assessment*	*Clinical sample*	*Comparison group*
				How was ZM reported?	**GCS mean/cut-off (SD), Mpta (SD), Mtsi (SD)**	**n;, Age, Mean (SD); Gender (n, %)**	**Group details, n;, Age, Mean, SD; Gender (n, %)**
Hendry et al. (2016)[1]	Australia	Prospective Case Control	Val: Eco	Mean profile and errors mean	GCS: Mean=5.51 (3.04), Mtsi=31.6 (36.92)	n=45, 37.9 (13.43), M=36 (80%), F=9 (20%)	N/A
Ownsworth et al. (2017)	Australia	RCT	Responsiveness to change	Mean Version 1 (errors) score	Sample 1: GCS Mean=6.19 (3.9), Mpta=76.16 (60.5),Mtsi=36.44 (45.8) Sample 2: GCS mean=5.12 (2.9), Mpta=81.5 (42.4), Mtsi=40.81 (49.3)	Sample 1: n=27, 37.37 (13.6) M=20 (74%), F=7 (26%). Sample 2: n=27, 37.86 (13.3), M=23 (85%), F=5 (15%)	N/A

Rakers et al. (2018)	Netherlands	Multi-Cohort	Val: KG	Mean profile score compared to control	GCS Mean=9.2 (3.1), Mpta=31 (32), Mtsi ≥3	n= 59, 42.9 (13.1), M=48 (81%), F=11 (19%)	Healthy Control: n=51, 41.9 (14.2), M=35 (69%), F=16 (31%) Mild TBI: n= 47, 37.5 (14.5), M=31 (66%), F=16 (34%)
Spikman et al. (2012)	Netherlands	RCT	Val: Conv Val: KG	Mean profile score compared to control	GCS Mean=9.5 (3.6), Mpta=41 (42), Mtsi=35	n=28, 30.1 (12.9), M=20 (71%), F=8 (29%)	Healthy Control: n=55, 30 (12.5) (13.2) M=30 (55%), F=25 (45%)
Weddell & Wood (2016)[3]	UK	Cross-sectional	Val: Conv Val: Eco	Mean profile score for each sample	Sample 1: GCS Mean=9, Mpta=26.8, Mtsi=64.5 Sample 2: GCS Mean=8.7, Mpta=28.5, Mtsi=49.3	Sample 1: n=40, 39.8, M=30 (75%), F=10 (25%) Sample 2: n=31 (33.8), M=22 (71%), F=9 (29%)	N/A
Wood & Liossi (2006)	UK	Single Cohort	Val: Conv Val: Eco	Mean profile score and mean profile score separated by brain injury area	GCS Mean=7.37 (3.6), Mpta=21.4 (34.09)	n=59, 33.86 (12.72) M=41 (70%), F=18 (30%)	N/A
Wood & Liossi (2007)	UK	Cross-sectional	Val: Conv	Mean profile score	GCS Mean=8.4 (3.97),	n=118, 35.47 (13.54), M=84	N/A

					Mpta=17.3 (3.65), Mtsi=34.8	(71%), F=34 (29%)	
Westerhof-Evers et al. (2019)[2]	Netherlands	RCT	Val: KG	Mean profile score compared to control	Mpta=32.4 (41.1), Mtsi=105 (103)	n=63, 42 (13), M=51 (81%), F=12(19%)	Healthy Control Group 1: n=72, 45 (15.4) M=50, (68%), F=22 (32%) Healthy Control Group 2: n= 45, 47.1 (11.8), M= 30 (67%), F=15 (33%)

Abbreviations: GCS = Glasgow Coma Scale, Mpta = Mean post traumatic amnesia duration (measured in days), Mtsi = Mean time since injury (measured in months), Val: Conv = Convergent Validity, Val: Eco = Ecological Validity, Val: KG = Known-Group Validity

[1]Did not report Mpta

[2]Did not report individual participant GCS or mean score

[3]Did not report Standard Deviations

Discussion

Although frequently used in the assessment of TBI, this is the first systematic review and synthesis of evidence relating to the psychometric properties of the Zoo Map Test within a TBI population. The review identified 8 studies which reported data relevant to the COSMIN taxonomy for reviewing the psychometric properties of patient-reported outcome measures. The COSMIN allows for a modular approach to reviewing the literature, permitting examination of current available evidence as well as highlighting gaps in the evidence base. No studies were found which reported evidence on content validity, structural validity, internal consistency, cross-cultural validity, reliability, measurement error, or criterion validity.

Overall, we found the Zoo Map to have sufficient evidence of known-group validity, highlighting its ability to differentiate between healthy controls and individuals with TBI, as well as between varying severity of TBI (mild vs moderate-severe). This evidence was found to be of moderate quality, and supports the use of the Zoo Map as an assessment tool to measure the impact of TBI on cognition. These results also underline that, absent from the rest of the BADS battery, the Zoo Map could still have relevance as a specific measure which can be utilised in research and clinical settings.

Our review highlighted insufficient evidence of responsiveness to change, though it should be noted that the quality of evidence for this was designated as "low" due to risk of bias (only one available study of "adequate" quality) and imprecision (lack of participants in the overall sample). This result is not entirely unexpected, as the Zoo Map was not designed as a tool to be sensitive to changes over time, but as a tool to identify potential executive deficits and how they may impact planning behaviour in everyday scenarios. This is particularly relevant when considering that moderate-severe TBI is associated with debilitating impairments in memory, attention and executive function which endure over time (Adnan et al., 2012). With this in mind, these results do not undermine the utility of the Zoo Map as an assessment tool, but are representative of its significant limitations as an outcome measure when used in isolation to identify potential changes in presentation over time, especially when considering the potential for test-retest effects, where practice may improve individuals' outcomes. It has even been suggested by Wilson (2003), one of the creators of the Zoo Map, that cognitive measures should never be used as measures of treatment outcome, that rehabilitation should aim to evoke meaningful change in quality of life, functional ability, and the achievement of patient-specific goals.

The studies included in this review examined the relationship between the Zoo Map and many different psychometric tools, used for a variety of purposes, including (but not limited to) other measures of executive function, measures of intellectual functioning, social cognition measurement tools, and measures of impulsivity and anger. The convergent validity of the Zoo Map could not, however, be determined due to variability in study quality and inconsistency in the outcomes of included studies' hypothesis testing. The significant heterogeneity of the comparison measures could explain some of this inconsistency, though, does not account for the variability in

outcomes when comparing the Zoo Map to other measures of executive function, where 66% of hypotheses were confirmed.

The Zoo Map Test was introduced as a measure thought to be more 'ecologically valid' than traditional executive tests, which would suggest that the evidence for ecological validity should be clear. The results of this review, however, highlight that there is insufficient evidence of ecological validity of the Zoo Map Test alone, within a TBI population. This evidence was found to be of a high quality, suggesting that this represents the "true" ecological validity of the Zoo Map. This would, however, be an oversimplification of the circumstances surrounding this outcome. The significant heterogeneity of outcome measures (DEX, HBCT, F&SB) compared to the Zoo Map in the three studies examining ecological validity makes it difficult to consider them as a collective group in order to provide an accurate synthesis of findings. The degree to which these tests represent other ecologically valid measures of executive function can also be questioned. Although the DEX is considered a well-established, and ecologically valid test of executive functioning, the other two tests used as comparison measures (the HBCT and F&SB) have not been robustly assessed to determine their psychometric properties. The results of this review in relation to the ecological validity of the Zoo Map should not be ignored, but should serve to highlight the need to examine long-held assumptions about the ecological validity of many tests of executive function.

The results of this review highlight the significant gaps in the literature surrounding executive function. The limited research available from which to draw conclusions means that this review cannot present comprehensive conclusions regarding the utility of the Zoo Map Test as a measure of executive function in patients with TBI. This highlights/reflects the problems facing this area of research. Wood and Liossi (2007) suggest that the ecological validity of executive tests is likely to vary across different populations and severities of those experiencing neuropsychological difficulties, whilst emphasising the role of Spearman's *g* in accounting for variance in performance in executive tests. The complex nature of executive function, and the lack of consensus relating to its relationship with *g* continues to be relevant more than 15 years on from these studies, with no clear resolution.

Future Research

More studies designed to examine the psychometric properties of the Zoo Map Test are required, in order to provide consistent evidence of a high quality which can be used to draw conclusions to benefit both clinical practice and future research. Specifically, studies examining the content validity of the Zoo Map for use in TBI would provide justification for its regular use, independent of the rest of the BADS, in this population. Similarly, novel studies examining the reliability and measurement error of the Zoo Map Test would provide valuable support as to its credibility as a tool to measure the severity of TBI. More high-quality studies are also necessary to clarify whether there is evidence for convergent validity of the Zoo Map, as this could not be determined within this review. The question of the role of Spearman's *g* has been raised with regards to all measures of executive function, and is addressed by one

study in this review in relation to the Zoo Map (Wood & Liossi, 2007). Further research investigating this relationship is necessary to support the identification and development of novel psychometric tests which can be used to identify deficits in functioning and support tailored interventions for individuals following a TBI.

Review Strengths and Limitations

The Cochrane Handbook for systematic reviews of interventions (Mokkink et al., 2018) recommends that data extraction is undertaken by two reviewers, independently, to avoid missing relevant information, and to ensure reliability of the extracted data. Due to time and resource constraints, only the primary reviewer participated in data extraction. Similarly, only a proportion of identified studies (20%) were screened by the second reviewer, increasing the possibility that potentially relevant studies may have been excluded. The exclusion of non-English language articles is also a limitation of this study which may have led to the exclusion of potentially relevant research evidence. It may also have been informative to conduct a meta-analysis of convergent validity correlations and known-group validity effect sizes, but unfortunately this was not possible within the scope of the current review. This review did, however follow a systematic process, following the clear structure provided by the COSMIN methodology, which allows for a thorough examination of the available evidence. The study also benefitted from the comprehensive examination of several research databases, with the inclusion of a grey literature repository, reducing the likelihood that publication bias influenced the pattern of results. The inclusion of a second-rater in the screening and Risk of Bias rating stages of this review is also a strength. This study addresses a clinically relevant question, making use of the currently existing evidence, and represents a starting point for further studies examining the psychometric properties of the Zoo Map.

Conclusions

This systematic review has implications for clinical practice, highlighting some benefits of the Zoo Map as an independent assessment tool, as well as indicating the need for critical analysis of its properties, and what they mean for an individual's ability to function in their daily life. The findings of this review also highlight the limitations of the current literature, and the need for future research with improved methods of examining and reporting the utility of tools used to examine executive functioning. There is a distinct lack of studies critically appraising the psychometric properties of the tools used in clinical practice and research on executive function, indicating that this is an area which should be a priority for future research.

References

Adnan, A., Crawley, A., Mikulis, D., Moscovitch, M., Colella, B., & Green, R. (2013). Moderate–severe traumatic brain injury causes delayed loss of white matter integrity: Evidence of fornix deterioration in the chronic stage of injury. *Brain Injury*, *27*(12), 1415-1422.

Alderman, B. (2003). Alderman, N., Burgess, P. W., Knight C., Henman, C. *Ecological validity of a simplified version of the multiple errands shopping test, J. Int. Neuropsychol. Soc*, *9*, 31-44.

All-Party Parliamentary Group on Acquired Brain Injury. (2018). Acquired brain injury and neurorehabilitation: time for change. https://cdn.ymaws.com/ukabif.org.uk/resource/resmgr/campaigns/appg-abi_report_time-forcha.pdf

Andrews, K. (2005). Rehabilitation practice following profound brain damage. *Neuropsychological Rehabilitation*, *15*(3–4), 461–472.

Burgess, P. W., Alderman, N., Evans, J. O. N., Emslie, H., & Wilson, B. A. (1998). The ecological validity of tests of executive function. *Journal of the International Neuropsychological Society*, *4*(6), 547-558.

Burgess, P. W., Alderman, N., Forbes, C., Costello, A., Coates, L. A., Dawson, D. R., Anderson, N. D., Gilbert, S. J., Dumontheil, I., & Channon, S. (2006). The case for the development and use of 'ecologically valid' measures of executive function in experimental and clinical neuropsychology. *Journal of the International Neuropsychological Society*, 12(2), 194-209.

Caplan, B., Bogner, J., Brenner, L., Williamson, M. L., Elliott, T. R., Dreer, L. E., ... & Perrin, P. B. (2016). Trajectories of life satisfaction over the first 10 years after

traumatic brain injury: race, gender, and functional ability. *Journal of Head Trauma Rehabilitation*, *31*(3), 167-179.

Catanese, C. (2007). *Traumatic Brain Injury: Methods for Clinical and Forensic Neuropsychiatric Assessment, Second Edition* (2nd ed.). CRC Press.

Chevignard, M. P., Taillefer, C., Picq, C., Poncet, F., Noulhiane, M., & Pradat-Diehl, P. (2008). Ecological assessment of the dysexecutive syndrome using execution of a cooking task. *Neuropsychological Rehabilitation.* 18, 461–485.

Clarivate. (2013). Endnote 20.

Elbaum, J. (2007). Acquired Brain Injury and the Family. In J. Elbaum & D. M. Benson (Eds.), *Acquired Brain Injury: An Integrative Neuro-Rehabilitation Approach* (pp. 275–285). Springer.

Emmanouel, A., Kessels, R. P., Mouza, E., & Fasotti, L. (2014). Sensitivity, specificity and predictive value of the BADS to anterior executive dysfunction. *Neuropsychological rehabilitation*, *24*(1), 1-25.

FitzGerald, M. C., O'Keeffe, F., Carton, S., Coen, R. F., Kelly, S., & Dockree, P. (2019). Rehabilitation of emergent awareness of errors post traumatic brain injury: A pilot intervention. *Neuropsychological Rehabilitation*, *29*(6), 821-843.

Galgano, M., Toshkezi, G., Qiu, X., Russell, T., Chin, L., & Zhao, L. R. (2017). Traumatic Brain Injury. *Cell Transplantation*, *26*(7), 1118–1130.

George, M., & Gilbert, S. (2018). Mental Capacity Act (2005) assessments: Why everyone needs to know about the frontal lobe paradox. *The Neuropsychologist*, 5(1), 59-66.

Gioia, G. A., & Isquith, P. K. (2004). Ecological assessment of executive function in traumatic brain injury. *Developmental neuropsychology*, 25(1-2), 135-158.

Grant, D. A., & Berg, E. A. (1948). A behavioural analysis of degree of reinforcement and ease of shifting to new responses in a Weigl-type card sorting problem. *Journal of Experimental Psychology,* 38, 404–411.

Guyatt, G. H., Oxman, A. D., Vist, G. E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., Schünemann, H. J., & GRADE Working Group. (2008). GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. *BMJ (Clinical Research Ed.)*, *336*(7650), 924–926.

Hendry, K., Ownsworth, T., Beadle, E., Chevignard, M. P., Fleming, J., Griffin, J., & Shum, D. H. K. (2016). Cognitive Deficits Underlying Error Behavior on a Naturalistic Task after Severe Traumatic Brain Injury. *Frontiers in Behavioral Neuroscience*, *10*, 190.

Krpan, K. M., Stuss, D. T., & Anderson, N. D. (2011). Coping behaviour following traumatic brain injury: What makes a planner plan and an avoider avoid? *Brain Injury*, *25*(10), 989–996.

Lamberts, K. F., Evans, J. J., & Spikman, J. M. (2010). A real-life, ecologically valid test of executive functioning: The executive secretarial task. *Journal of Clinical and Experimental Neuropsychology*, *32*(1), 56-65. https://doi.org/10.1080/13803390902806550

Langlois, J. A., Rutland-Brown, W., & Wald, M. M. (2006). The Epidemiology and Impact of Traumatic Brain Injury: A Brief Overview. *The Journal of Head Trauma Rehabilitation*, *21*(5), 375.

Malec, J. F., Brown, A. W., Leibson, C. L., Flaada, J. T., Mandrekar, J. N., Diehl, N. N., & Perkins, P. K. (2007). The mayo classification system for traumatic brain injury severity. *Journal of Neurotrauma*, *24*(9), 1417-1424.

McDonald, S., & Genova, H. (2021). The effect of severe traumatic brain injury on social cognition, emotion regulation, and mood. *Handbook of Clinical Neurology*, *183*, 235-260.

Mokkink, L. B., De Vet, H. C., Prinsen, C. A., Patrick, D. L., Alonso, J., Bouter, L. M., & Terwee, C. B. (2018). COSMIN risk of bias checklist for systematic reviews of patient-reported outcome measures. *Quality of Life Research*, 27(5), 1171-1179.

Newstead, S., Lewis, J., Roderique-Davies, G., Heirene, R. M., & John, B. (2022). The Paradox of the Frontal Lobe Paradox. A Scoping Review. *Frontiers in Psychiatry*, *13*, 913230.

Norris, G., & Tate, R. L. (2000). The Behavioural Assessment of the Dysexecutive Syndrome (BADS): Ecological, concurrent and construct validity. *Neuropsychological Rehabilitation*, *10*(1), 33–45.

Ownsworth, T., Fleming, J., Tate, R., Beadle, E., Griffin, J., Kendall, M., Schmidt, J., Lane-Brown, A., Chevignard, M., & Shum, D. H. K. (2017). Do People With Severe Traumatic Brain Injury Benefit From Making Errors? A Randomized Controlled Trial of Error-Based and Errorless Learning. *Neurorehabilitation and Neural Repair*, *31*(12), 1072–1082.

Patton, J. H., Stanford, M. S., & Barratt, E. S. (1995). Factor structure of the Barratt impulsiveness scale. *Journal of Clinical Psychology*, *51*(6), 768-774.

Ponsford, J. (1995). Mechanisms, recovery and sequelae of traumatic brain injury: a foundation for the REAL approach. *Traumatic brain injury Rehabilitation for every day adaptive living*. 1-31

Ponsford, J., Rudzki, D., Bailey, K., & Ng, K. T. (2007). Impact of apolipoprotein gene on cognitive impairment and recovery after traumatic brain injury. *Neurology*, *68*(8), 619-620.

Rakers, S. E., Scheenen, M. E., Westerhof-Evers, H. J., de Koning, M. E., van der Horn, H. J., van der Naalt, J., & Spikman, J. M. (2018). Executive functioning in relation to coping in mild versus moderate-severe traumatic brain injury. *Neuropsychology*, *32*(2), 213–219.

Shallice, T., & Burgess, P. (1991). Deficits in strategy application following frontal lobe damage in man. *Brain 114*, 727–741

Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology: 15:* 201-293.

Spielberger, C. D., Jacobs, G. A., Russell, S. F., & Crane, R. S. (1983). Assessment of anger: The State-Trait Anger Scale. In Butcher J. N., Spielberger C. D. (Eds.), *Advances in personality assessment* (Vol. 2, pp. 161-189).

Spikman, J. M., Timmerman, M. E., Milders, M. V., Veenstra, W. S., & van der Naalt, J. (2012). Social cognition impairments in relation to general cognitive deficits, injury severity, and prefrontal lesions in traumatic brain injury patients. *Journal of Neurotrauma*, *29*(1), 101-111.

Stebbins, G. (2007). Neuropsychological Testing. *Textbook of Clinical Neurology: Third Edition*, 539–557.

Stuss, D. T., & Alexander, M. P. (2007). Is there a Dysexecutive Syndrome? *Philosophical Transactions of the Royal Society B: Biological sciences, 362(1481), 901-915.* h

Stuss, D. T. (2011). Traumatic brain injury: relation to executive dysfunction and the frontal lobes. *Current Opinion in Neurology*, *24*(6), 584-589.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. Journal of experimental psychology, 18(6), 643.

Temkin, N. R., Corrigan, J. D., Dikmen, S. S., & Machamer, J. (2009). Social functioning after traumatic brain injury. *The Journal of Head Trauma Rehabilitation*, *24*(6), 460-467.

Turner-Stokes, L., Pick, A., Nair, A., Disler, P. B., & Wade, D. T. (2015). Multi-disciplinary rehabilitation for acquired brain injury in adults of working age. *Cochrane Database of Systematic Reviews*, (12).

Weddell, R., & Wood, R. (2016). Exploration of correlates of self-reported personality change after moderate–severe traumatic brain injury. *Brain Injury*, *30*, 1–10.

Weddell, R. A., & Wood, R. L. (2018). Perceived personality change after traumatic brain injury II: comparing participant and informant perspectives. *Brain Injury*, *32*(4), 442-452.

Westerhof-Evers, H. J., Fasotti, L., van der Naalt, J., & Spikman, J. M. (2019). Participation after traumatic brain injury: The surplus value of social cognition tests beyond measures for executive functioning and dysexecutive behavior in a statistical prediction model. *Brain Injury*, *33*(1), 78–86.

Wilson, B. A. (1993). Ecological Validity of neuropsychological assessment: Do neuro-psychological indexes predict performance in everyday activities. *Applied and Preventive Psychology,* 2, 209-215.

Wilson, B. A. (2003). Goal planning rather than neuropsychological tests should be used to structure and evaluate cognitive rehabilitation. *Brain Impairment*, *4*(1), 25-30.

Wilson, B. A., Alderman. N., Burgess, P. W., Emslie, H., & Evans, J. J. (1996). *Behavioural assessment of the dysexecutive syndrome.* St Edmunds UK: Thames Valley Test Company.

Wilson, B. A., Alderman, N., Burgess, P. W., Emslie, H., Evans, J. J., & Chamberlain, E. (2003). Behavioural assessment of the dysexecutive syndrome (BADS). *Journal of Occupational Psychology, Employment and Disability*, *5*(2), 33–37.

Wilson, B. A., Evans, J. J., Emslie, H., Alderman, N., & Burgess, P. W. (1998). The development of an ecologically valid test for assessing patients with a dysexecutive syndrome. *Neuropsychological Rehabilitation*, *8*(3), 213–228.

Wood, R. L., & Liossi, C. (2006). The ecological validity of executive tests in a severely brain injured sample. *Archives of Clinical Neuropsychology*, *21*(5), 429–437.

Wood, R. L., & Liossi, C. (2007). The relationship between general intellectual ability and performance on ecologically valid executive tests in a severe brain injury sample. *Journal of the International Neuropsychological Society*, *13*(1), 90–98.

Chapter 2

Executive Function and Fluid Intelligence in Traumatic Brain Injury

Prepared in accordance with the author requirements for PLOS ONE: https://journals.plos.org/plosone/s/submission-guidelines

Word Count: 8275

Plain Language Summary

Executive Functioning is responsible for influencing performance in tasks of general mental ability, flexible thinking, self-control, and goal-oriented action. These skills help us to learn, work, and complete everyday tasks. Competing theories argue whether executive function relies on smaller areas of the frontal region of the brain that work to contribute towards different elements of executive function but communicate together; or whether it functions using a distinct area, which specifically controls all executive function.

Some researchers have suggested that tests which clinicians believe to be assessing executive functioning, may measure a different psychological concept known as “fluid intelligence”, thought to be representative of our ability to solve complex problems.

This study looked at routinely collected cognitive test data from a group of 150 people with traumatic brain injury seen in a regional specialist neurorehabilitation centre. We compared results from tests thought to be associated with executive functioning with those thought to be associated with fluid intelligence to see how much these concepts are linked, as well as examining whether the tests used by the service fit best into a model where executive function is separated into different elements, or a model where executive function is an independent construct.

Our analyses suggest that results in tests of executive function can account for some of the variation in fluid intelligence seen in the group. Tests which measure executive function fit best in a model where scores can be explained by two factors. The first factor seems consistent with the idea of general-purpose or fluid intelligence and the second with more specific cognitive functions, and in this study those functions were multi-tasking and taking another person’s perspective.

These results indicate: (i) we may need to reassess how we examine executive function in clinical practice to consider the contributions of both fluid intelligence and specific executive functions, (ii) that we should develop new tests that accurately measure both categories of ability. Such assessments may give more accurate indicators of the specific challenges faced after brain injury, and help clinicians to plan rehabilitation and management interventions to help offset the difficulties presented by these challenges.

Abstract

Background: There is considerable debate regarding the nature and role of executive function (EF) within the prefrontal cortex (PFC), and whether or not there are discrete categories of EF within different areas of the frontal lobe, or a "central executive". The role of fluid intelligence (Spearman's *g*) is significant, as it has been suggested that *g* accounts for much of the variation in EF test performance (Roca, 2010).

Aims: We examined whether scores in EF tests account for variability in fluid intelligence in a clinical sample of patients with Traumatic Brain Injury (TBI), where PFC damage is common.

Methods: We analysed routinely-collected neuropsychological assessment data from 150 patients with moderate-severe TBI who had undergone a comprehensive assessment of intellectual ability and executive function at an NHS rehabilitation service. The relationship between *g* and EF was examined through regression analyses, and an exploratory factor analysis (EFA).

Results: Multiple regression highlighted consistent statistically significant relationships between scores on EF tests and *g*, including both traditional and naturalistic EF tasks. The EFA highlighted that many tests loaded strongly onto a factor theoretically similar to the concept of *g*, explaining 49.8% of variance in *g* (β=.730, $p<.001$). A second factor, which did not significantly predict *g*, comprised of tests examining more specific cognitive functions.

Conclusions: Traditional and naturalistic EF tests were significant predictors of *g*. EFA highlighted a two-factor structure which is likely representative of *g*, and more discrete executive functions, consistent with the findings of Roca et al. (2010). Clinical assessment of EF should incorporate a good measure of *g*, alongside specific EF tests to measure more discrete functions including multi-tasking and perspective-taking. Future research should examine the relationship between fluid IQ loss and EF in clinical samples and develop a gold-standard approach to measuring both *g* and EF.

Keywords: Executive function, Traumatic Brain Injury, fluid intelligence, neuropsychological assessment

Introduction

Executive functions (EFs) are the high-level cognitive processes which facilitate our ability to problem solve, through developing new behaviour, particularly in response to novel stimuli. These processes are thought to be, in part, supported by structures within the frontal lobes of the brain, as well as the parietal cortex and the limbic system (Gilbert & Burgess, 2008). For several decades, there has been debate surrounding the frontal-parietal network that is considered responsible for EFs. Performance on EF tasks is variable, with many measures of EF showing low inter-correlations (Wood & Liossi, 2007). Previous research has, therefore, examined whether this network functions autonomously, as a "central executive" responsible for EF, or whether there are specific sections within the frontal-parietal network which are responsible for distinct aspects of EF (Duncan & Owen, 2000; Roberts, 1996). Monsell (1996) described our knowledge of this area as an "embarrassing zone of almost total ignorance", and although a considerable volume of research has since furthered our understanding of executive function, no definitive conclusions have been made, and debate continues. For example, Duncan et al., (2020) recently presented a model of fluid intelligence based upon its relationship with executive function, which incorporates a frontoparietal network responsible for completing tasks with multiple executive demands. Conversely, Shallice and Cipolotti (2018) propose that most executive tasks involve many components, and that the frontal-parietal network is more complex than neuropsychological group studies can account for.

It has been widely held that specific cognitive functions are somewhat localised to specific regions of the brain (Savoy, 2000; Shallice, 1998). Studies examining the impact of frontal lobe lesions on cognitive functioning have led some researchers to propose that there are discrete categories of functions within the frontal lobes, for example that task setting is associated with the left lateral, and monitoring associated with the right lateral areas of the frontal lobes (Stuss, 2011). Based on these observations, it has been suggested that the concept of a sole "central executive" is not supported, and that a fractionation model may better explain the mechanisms underlying Executive Function (Stuss & Alexander, 2007). However, evidence supporting the separability of executive functions is not consistent or conclusive (Roca et al., 2010).

Relevant to this issue is the role of general intelligence, also referred to as Spearman's *g* (Spearman, 1904). Spearman initially proposed that this process is responsible for overall performance in mental ability tests, and later claimed that it also influences performance on all cognitive tasks (Spearman, 1927). Tasks which are thought to be the best measure of this are known as "fluid intelligence" tests, and are thought to be more reliant on frontal lobe function than tests measuring "crystallized" intelligence (Ziegler et al., 2012). Crystallized intelligence relates to learned procedures and knowledge that develops over time, and is a construct which remains relatively stable across neuropsychological tests even in the context of neurological deficits. Horn (1967) describes crystallized intelligence as being based on prior experience, and the development of education and culture that comes with such experience, whilst fluid

intelligence represents the use of deliberate mental operations such as concept formation, hypotheses generating/testing, problem solving and the use of inductive and deductive reasoning.

The relationship between *g* and frontal lobe function has been hypothesised as being related to a phenomenon known as "goal neglect" (Duncan, et al., 1996), whereby people with prefrontal cortex (PFC) damage may exhibit behaviour that is contradictory to their stated goals, and which may in turn link to the everyday disorganisation seen in this patient group. By contrast, fluid intelligence is characterised by the ability to adhere to a plan through the activation of task-oriented goals. Several studies have attempted to investigate the relationship between *g* and frontal lobe function. For example, Duncan et al. (1997) examined common elements in different tests conventionally used to examine EF in a sample of 90 people with head injury. They found that the common element in EF tests related closely to *g*. Similarly, Roca et al.'s (2010) study of EF and fluid intelligence in patients with frontal lobe lesions highlighted that when patients and controls were matched on fluid intelligence, few further frontal deficits remained, regardless of precise lesion location. Of note, measures of multitasking, verbal abstraction and social cognition were found to be relatively distinct from *g*, and Roca et al. (2010) suggested that these tests may be distinguished by their particular association with the functions of the most anterior part of the PFC, Brodmann area 10. These findings suggest that with some exceptions, many tests purporting to measure *g* or EF may actually be measuring one, or overlapping, construct(s). A challenge for Roca et al.'s findings is the dissociations seen in clinical practice between tests of executive function – e.g. it is commonly the case when administering a range of EF tests that performance can be impaired on some and spared on others.

It is also common in clinical practice to see dissociations in performance across tests of EF. That is, performance in one test may not be indicative of performance on another test purportedly measuring the same construct (Burgess, 2004). This is also true when considering that performance on EF tests does not necessarily correspond with performance in complex, real-world situations (Burgess et al., 1998). If the relationship between *g* and EF was as clear as Duncan et al. (1997) propose, then we might expect to see more stable performance across all executive tests. However, an example of the complexity associated with EF assessment can be seen in what has been referred to as the "frontal lobe paradox" (Walsh, 1985). This concept refers to the phenomena whereby people with frontal lobe damage who show significant impairment in everyday functioning can sometimes still verbally describe a logical course of action relating to a specified task, but are unable to complete said task using the logic they previously described. This led to the development of a range of more naturalistic tasks, thought to be more representative of challenges associated with everyday functioning, such as the Multiple Errands Test (Shallice & Burgess, 1996), and the Behavioural Assessment of the Dysexecutive Syndrome (BADS; Wilson et al., 1996). These tests were introduced as 'ecologically valid' tests of EF, in that they were believed to replicate real-life scenarios likely to be encountered by those with executive deficits, though some have challenged this notion (Chan et al., 2008).The literature examining EF continues to debate several theoretical perspectives. Miyake

et al. (2000) studied EF in a neurologically healthy population to examine whether the factor structure of EF was best represented by a singular system or multiple systems co-ordinating in one area of the brain (unitary vs non-unitary). They examined the organisation and role of three proposed EFs: shifting, updating and inhibition. Their conclusions conflicted with Duncan's (1996) conceptualisation of EF as one construct, and argued that the three-factor model provided a significantly better fit to the data than a unitary model of EF, though highlighted that these three systems display some level of commonality. They further explained that EFs may be best characterised as distinct yet related constructs which share at least one commonality. Similarly, Hampshire et al. (2012) used factor analysis of brain imaging, behavioural and simulated data, and provided support to the non-unitary theory. They highlighted that individual components of intelligence correlated with questionnaire variables that they claimed had previously been associated with *g* in a unique and discrete fashion. The authors did not, however, describe in detail what these questionnaires were or how they had been previously associated with *g.* They concluded by suggesting that human intelligence comprises multiple complex systems which have unique functions and are capable of working independently.

Should the conclusions of Roca et al. (2010) be replicable and generalisable to other populations and to broader tests of executive function, these findings would have significant implications for the clinical assessment, interpretation and formulation of losses in EF. Specifically, it may impact the types of tests used to examine these neuropsychological deficits. Traditionally, EF is assessed using a comprehensive battery, aimed at testing various facets of this complex domain of functioning (Harvey, 2022). As highlighted by Roca et al. (2010), developing our understanding of deficits in EF may support the design of more appropriate assessment tools.

Given the potential clinical implications of Roca et al.'s (2010) study, it is important to identify whether this pattern holds within clinical samples. Roca et al.'s (2010) study used a research sample whose characteristics may be unlike those seen within usual clinical practice. A clinical sample is more likely to represent increased variety, both in terms of the participants' biological presentation (mechanism of injury, comorbidities), and wider individual and systemic factors such as educational background and cultural diversity. It is, therefore, necessary to examine whether the pattern identified within Roca et al.'s (2010) study, that fluid intelligence loss was highly correlated to losses in EF, is replicable when examining a sample of patients more representative of those seen within clinical practice.

A condition often associated with frontal lobe damage, and where the accurate measurement of EF is crucial, is Traumatic Brain Injury (TBI). Those who survive a TBI may experience debilitating symptoms which significantly impair their everyday functioning. As mortality rates for TBI improve, the number of people living with symptoms of TBI has increased, and is now considered a leading cause of disability worldwide (Maas et al., 2017). TBI is often described as heterogeneous, including in its severity, neuropathology, injury location, and in the unique presentation of cognitive difficulties presented by each patient (Covington & Duff, 2021). It is, however,

associated with a characteristic pattern of focal injury to areas such as the frontal and temporal lobes (Ng & Lee, 2019), alongside more diffuse axonal injury resulting from initial physical impact, combined with tearing and shearing from rotational forces as part of the primary injury, alongside further damage from possible secondary injuries (associated with for example, inability to manage raised intracranial pressure or other medical and surgical complications; Mckee & Daneshvar, 2015). Previous research has highlighted that TBI can result in altered patterns of functional connectivity in intrinsic connectivity networks such as the salience network and default mode network (Sharp et al., 2011; Sharp et al., 2015). These findings highlight a relationship between TBI and neuropathology that continues to develop in TBI patients following their injury, leading to cognitive and functional impairment. Due to the nature of these types of injury, patients often present with common impairments including: apathy, behavioural challenges, organisation difficulties and deficits in one's ability to problem solve. It is therefore common for patients presenting to NHS neurorehabilitation services to exhibit losses in EF. Stuss (2011) previously described the impairments commonly seen following TBI, and argued that research into focal lesions identified four categories of frontal lobe functions: energization, executive cognitive, emotional/behavioural and integrative/metacognitive. They suggested that the potential impairments associated with frontal injury map onto this fractionation model, and that tests such as the Stroop test and the Wisconsin card sorting test accurately measure discrete processes related to different frontal lobe regions.

This research aimed to examine whether the findings of previous research, which suggests that poor EF performance can be explained by fluid intelligence (*g*), and loss of fluid intelligence can be replicated within a clinical TBI sample assessed with an extensive clinical neuropsychological test battery. This issue was important to address, as the original findings imply that there is little value in conducting extensive assessments of EF in clinical practice, whereas in clinical practice comprehensive assessment of EF is common. These aims were operationalised with the following research questions:

1. Do EF tests account for significant variance in fluid IQ, across the following test types:
    a. 'traditional' executive tests including verbal fluency, trail making test-switching, and Stroop interference
    b. 'ecologically valid' executive tests including Six Elements and Zoo Map tests from the Behavioural Assessment of the Dysexecutive Syndrome

2. Do EF scores account for variance in fluid IQ beyond what is accounted for by 'non-cognitive' variables including demographic characteristics, mood, and anxiety?

3. Do EF tests used within the service fit best onto a unitary or fractionated factor structure?

Methods

Sample

This project analysed routinely collected assessment data from patients referred for neuropsychological rehabilitation at the Wolfson Outpatient Cognitive Rehabilitation Service (WOCRS) based at St George's University Hospital NHS Foundation Trust in London. WOCRS sees a broad age range of patients with all neurological conditions including TBI, and similarly accepts referrals from all genders (though it should be noted that the population of those experiencing TBI is likely to be younger adults, and males are more likely to experience TBI than females; Leitgeb et al., 2011). A power calculation for regression analysis suggested that a regression analysis with medium effect size ($f^2 = 0.15$) and power of 0.8 using 10 or fewer predictor variables would require a minimum sample of 117 participants to achieve a significant result at $p<.05$. A total of 741 patients were screened against the inclusion criteria for this study to assess their eligibility for participation. 556 participants were excluded in our initial screening as they did not meet inclusion criteria (principally as they had sustained brain injuries via a stroke or other non-trauma cause). A second screening then determined the amount of missing data across all cases for variables to be included in the regression models. Due to the use of a clinical sample, and a non-uniform method of data collection (multiple clinicians assessing different patients with complex and variable presentations), it was assumed that there was an element of non-random bias related to missing data. Previous research (Jacobsen, Gluud & Winkel, 2017) suggested that variables included in a regression model should only be excluded in cases where proportions of missing data are very large (40% or higher). For these reasons, only variables and cases meeting the predetermined threshold of 60% of relevant data were included in the analysis. Data for a final sample of 150 patients (100 male, 50 female) were included in the statistical analyses, surpassing the predetermined minimum sample size.

Procedure

Patients completed a detailed cognitive assessment which included a 90-minute clinical interview with a clinical neuropsychologist (usually with a relative also present), and full day of cognitive testing using a comprehensive neuropsychological assessment battery. This battery included a range of measures of intellectual ability, attention, and executive function, usually administered by a trained assistant psychologist supervised by a qualified clinical psychologist or neuropsychologist, or occasionally administered by a qualified clinical psychologist/neuropsychologist.

The criteria for participant inclusion in the study were:

- aged 18 or over
- referred to WOCRS from 2015 onwards (when the current test battery was introduced)
- experienced a TBI of at least moderate level of severity (i.e. those with mild TBI were excluded)
- presenting with cognitive or behavioural symptoms associated with the TBI

- IQ >70 on estimates of current intellectual functioning (Vocabulary and Matrix Reasoning subtests of WAIS-IV)
- No history of a learning disability or language disorder precluding administration of the standard battery
- No previously identified underlying organic neurological conditions such as other types of acquired brain injury or dementia

Time since injury for patients involved in this study varied, with the lower limit around 3 months, reflecting the time in the care pathway at which patients are typically referred for post-acute rehabilitation from acute services, and with no upper limit, as the service also accepts referrals from GPs and community services without any restriction according to factors such as age or time post-injury. Data relating to time since injury at time of assessment was not available for this study.

The measures used in this study were selected by mapping those used by Roca et al. (2010) onto equivalent assessment tools used within WOCRS, in an attempt to examine the same cognitive domains as the prior study, alongside additional executive measures thought to be of interest based on clinical experience (see Table 2.1).

Table 2.1: Measurement tools used by Roca et al. (2010) mapped against their equivalent tests from WOCRS neuropsychological testing battery plus additional measures not included by Roca et al. (2010)

Test Domain	*Measures used by Roca et al. (2010)*	*Measures in WOCRS battery*	*Brief description of WOCRS measure*
Fluid IQ/g	Cattell Culture Fair (IQ)	WAIS-IV PRI (comprising Matrix Reasoning, Block Design, Visual Puzzles)	Index score of 3 sub-measures from WAIS-IV used to highlight subject's ability to find relationships between non-verbal stimuli, and test reasoning skills
IQ loss	NART-Cattell IQ	TOPF estimated premorbid PRI minus current fluid IQ	Estimated premorbid perceptual reasoning index using test of premorbid functioning, minus current WAIS-IV PRI score
Executive function	WCST errors	DKEFS Sort Free Sorting (correct)	Scaled score based on number of correct categories/rules achieved
	Verbal Fluency (FAS, total)	DKEFS VF Phonemic	Scaled score based on number of correct responses generated across three trials FAS
	Interference (motor-reverse)	Stroop Trial B (repeat) percentile score	Percentile score for naming speed in colour-word interference condition
	Backwards digit span (Score 1 regardless of whether or not 1 or 2 per trial correct)	WAIS-IV Digit Span Backwards	Scaled score based on total number of digit strings correctly recalled in reverse order
	Spatial working memory (reverse block span) n correct spans 2 to 6	WMS-IV Symbol Span	Scaled score based on number of correctly identified and positioned symbols after brief exposure

Spatial working memory (reverse block span) n correct spans 2 to 6	WMS-IV Symbol Span	--
Spatial working memory (reverse block span) n correct spans 2 to 6	WMS-IV Symbol Span	--
Proverbs (3 items, 1 or 0.5 per item)	WMS-IV Similarities	Scaled score based on accuracy of participants' description of link between word pairs
Hotel sum total deviation from optimum	BADS Six Elements Test (overall)	Scaled score based on overall test performance, incorporating multi-tasking efficiency and errors
Hayling (short – 6 item, scoring for errors)	Stroop Trial B (repeat) percentile score	--
Months backward (errors 0,1,2)	--	--
Go-No Go	--	--
Motor programming (/3)	--	--
Iowa Gambling Test (total conservative minus risky choices)	--	--
Faux pas (1 for each Faux Pas correctly identified and 1 for each non-faux pas identified)	DKEFS Sort Recognition vs Free Sorting Description	Scaled score of contrast between individual ability to describe self-sorting vs examiner's sorting behaviour based on set of prescribed rules
Mind in the Eyes (total correct)	--	--

	--	VF Switching	Scaled score based on participant ability to switch between generating words fitting two semantic categories
	--	WAIS-IV PSI	Index score based on two subtests from WAIS-IV measuring visuo-motor processing speed
	--	Zoo Map 1 Errors/sequencing	Scaled score of number of errors made during complex route planning task
	--	Trail Making (Switching)	Scaled score based on speed at which participants complete visuo-motor switching task
Mood	--	GAD-7	7-item self-report measure of generalised anxiety based on diagnostic criteria
Anxiety	--	PHQ-9	9-item self-report measure of depression based on diagnostic criteria

List of abbreviations: BADS = Behavioural Assessment of Dysexecutive Syndrome, DKEFS = Delis-Kaplan Executive Function System, GAD-7 = General Anxiety Disorder-7, PHQ 9 = Patient Health Questionnaire 9, NART = National Adult Reading Test, TOPF = Test of Premorbid Functioning, WAIS = Wechsler Adult Intelligence Scale, WCST = Wisconsin Card Sorting Test, WMS = Wechsler Memory Scale, VF = Verbal Fluency, PRI = Perceptual Reasoning Index, PSI = Processing Speed Index

Table 2.2: Separation of executive tests ("traditional" vs "naturalistic")

"Traditional" Tests	***'Ecologically Valid tests***
Stroop B	Zoo map 1 (errors)
WAIS-IV digit span backwards	Six elements
DKEFS sort (correct)	
DKEFS trail making task	
Verbal fluency (phonemic)	

Research Governance and Data Protection

As this project used anonymised data routinely collected as part of clinical practice, it was exempt from NHS Research Ethics Committee review, as confirmed by an approvals manager from the Health Research Authority on 17/01/23. However, approval was sought and approved by the Health Research Authority and Health and Care Research Wales on 13/02/23 (HRA Reference: 23/HRA/0185). Confirmation of capacity and capability was received by a representative of St George's University Hospitals NHS Foundation Trust. Sponsorship was undertaken by the University of Glasgow (UoG).

A member of the clinical care team anonymised routinely collected clinical data, using a secure, remote desktop connection. The anonymised data was transferred via secure encrypted file transfer (see https://transfer.gla.ac.uk/ for specifications) to the researcher, and stored on the University of Glasgow OneDrive electronic storage system alongside all other project documents. The researcher used a template database developed by the service to develop a more comprehensive repository of data for relevant neuropsychological tests, alongside demographic information such as age, gender, and level of education. No personally identifiable information was accessed by the primary researcher at any point.

Statistical Analysis

All statistical analyses were conducted using SPSS version 28 (IBM, 2021). Descriptive statistics were used to analyse the characteristics of the sample, including demographics, average scores on cognitive domains, mood, and anxiety. Index/scaled scores were used for all variables to account for age as a confounding factor. The study's primary analysis addressed research question 1, examining the relationship between *g* and EF. We analysed correlations between fluid intelligence scores (*g*) and results on a broad range of executive tests to examine the variance in *g* accounted for by EF.

This study examined the relationship between tests considered to be 'traditional tests' (i.e. those previously identified as overlapping with the construct of *g*) and 'ecologically valid' tests (i.e. those previously identified as being distinct from *g*), with fluid intelligence in order to determine how much variance in *g* can be explained by EF test scores (see Table 2.2 for details on included measures). Two regression models were developed, one for 'traditional tests', and one for 'ecologically valid' tests, to examine the *g* variance accounted for by each type of EF test. To answer our second research question regarding the amount of unique variance EF accounts for, beyond that of 'non-cognitive' variables for each test category, the following potential confounding variables were included within the regression models: gender, education, mood, and anxiety.

Given the lack of consensus in the literature relating to the structure of EF, an exploratory factor analysis (EFA) of scores from all executive tests included in WOCRS battery was also conducted to provide evidence relating to the factor structure of EF. Finally, a third regression model was developed which included the factor scores based on the results of our factor analysis, to examine the relationship between these factors and *g*.

Results

Descriptive statistics on a small selection of demographic variables were computed, in order to characterise the sample. There were 100 male (66%) and 50 female (33%) participants, aged from 18-70 years (M=39.72, SD=13.78). Participants had on average 15.27 years of education (SD=3.15, range 9-22). The estimated premorbid intellectual ability of the sample was in line with the population mean (mean TOPF-estimated FSIQ=100.42, SD=11.24, range 70-127), as was current fluid intelligence as measured by the WAIS-IV perceptual reasoning index (PRI) (M=101.38, SD=18.46, range=58-142).

Relationship Between g and Executive Function

Hierarchical multiple regression was used to examine the relationship between fluid intelligence and the two 'ecologically valid' EF tests, after controlling for the influence of non-cognitive variables (see Table 3 for descriptive statistics). Preliminary analyses were conducted to ensure there were no violations of the assumptions of normality, linearity, multicollinearity, and homoscedasticity. Gender, years of education, mood, and anxiety were entered at step 1, explaining 8.7% of the variance in *g*. Following entry of Zoo Map 1 (errors) and the six elements test at step 2 of the model, 36% of variance was explained by the model as a whole, $F (6, 97) = 9.11$, $p<.001$. The two EF measures explained an additional 27.3% of variance in *g*, after controlling for gender, education, mood, and anxiety (R squared change = .273, F change $(2, 96) = 20.72$, $p<.001$). In the final model, only the two EF measures were statistically significant. The six elements test explained the most variance ($\beta=.414$ $p<.001$), then the Zoo Map 1 (errors) [$\beta=.338$, $p<.001$). Full correlations for this regression analysis are reported in Table 4.

A second hierarchical multiple regression was used to examine the relationship between five "traditional" EF tests, after controlling for the influence of non-cognitive variables (descriptive statistics in Table 5). Preliminary analyses were again conducted to ensure there were no violations of the assumptions of normality, linearity, multicollinearity and homoscedasticity. Gender, education (in years), mood, and anxiety were entered at step 1, again explaining 8.7% of the variance in *g*. Following entry of DKEFS trail making task, DKEFS sorting (correct), digit span (backwards), Stroop B, and verbal fluency (phonemic) at step 2 of the model, 54.3% of the variance was explained by the model as a whole $F (9, 99) = 15.28$, $p<.001$. The 5 EF measures explained an additional 49.4% of variance in *g,* after controlling for gender, education, mood and anxiety (R squared change = .494, F change $(5, 99) = 23.372$, $p<.001$). In the final model, three of the five EF measures were statistically significant, with the DKEFS trail making task having the highest beta value ($\beta=.331$, $p<.001$), followed by the DKEFS sorting task ($\beta=.250$, $p<.05$) and backwards digit span ($\beta=.199$, $p<.05$). Full correlations for this regression analysis are reported in Table 6.

Factor Structure of WOCRS Executive Tests

Participant data from the twelve EF tests from the WOCRS battery was examined by EFA using principal axis factoring (PAF). Previous research has indicated that PAF is the preferred EFA technique over maximum likelihood analysis when attempting to limit potential overextraction (De Winter & Dodou, 2011). Prior to performing this analysis, the suitability of data for factor analysis was assessed. Due to the nature of the clinical data used for this project, there were varying amounts of available data for each variable included in the

factor analysis (lowest n=136 for six elements test). Inspection of the correlation matrix revealed the presence of coefficients of .3 and above for the majority of included variables.

The communality of each factor was assessed to ensure acceptable levels of explanation (>.03), only one item (zoo map 1 errors) had a communalities value of <.03 (Factor 1=.245, Factor 2=.228). This item was ultimately included in the final factor model due to a moderate loading of .464 on Factor 1. No other variables were required to be excluded from the analysis, though it should be noted that the only other factor with similarly low values in communalities was the other 'ecologically valid' test, the six elements test (Factor 1=.324, Factor 2=.287). The Kaiser-Meyer-Olkin measure of sampling adequacy, representative of appropriateness for factor analysis, was .883, exceeding the recommended minimum value of .6, and achieving a "meritorious" rating (Kaiser, 1970, 1974). Bartlett's Test of Sphericity (Bartlett, 1954) which provides a measure of the statistical probability that the correlation matrix has significant correlations amongst its components reached statistical significance p<.001, supporting the factorability of the correlation matrix.

PAF revealed the presence of two factors with eigenvalues exceeding 1, explaining 43.54%, and 10.85% of the variance respectively. An inspection of the scree plot (see Appendix 9) using Catell's (1966) screen test revealed a clear break after the second factor, further supporting the retention of this two-factor solution. Given there was a weak, negative component correlation between the two factors (r=-.357), it is reasonable to assume that the two components are distinct factors which are weakly correlated, further supporting the two-factor solution. The sum of squared percentage of variance indicated that this two-factor solution explained a cumulative total of 45.14% of variance. Reproduced correlations indicated that less than 50% of reproduced residuals had absolute values greater than 0.05, indicating that the two-factor model had a good fit.

To aid in the interpretation of this solution, oblimin rotation was performed (based on initial unrotated component correlation of less than 0.3, r=.169). Factor loading significance cut-offs were based on Tabachnick and Fidell's (2007) suggested values of: 0.32=poor, 0.45=fair, 0.55=good, 0.63=very good and 0.71=excellent. The rotated solution revealed "fair" loadings for Zoo Map 1 (errors) and the six elements, whilst the remaining variables scored good-excellent factor loadings on Factor 1. Comparatively, Factor 2 has only two significant loadings, the six elements test and DKEFS sort (recognition vs free sorting description) loadings which had "poor" and "very good" loadings respectively (see table 7 for details).

There was a degree of communality between the factors, with both factors showing significant loadings for the six elements test. Factor 1 had significant loadings from all but one of the included executive tests (DKEFS sort recognition vs free sorting description). This fits with the theoretical proposition suggesting the underlying role of fluid intelligence in relation to executive function, and is in line with the conclusions made by Roca et al. (2010). Factor 2 showed significant loadings on trail-making (switching), symbol span, WAIS IV PSI, six elements test and DKEFS sort (self-other discrepancy), though it should be noted that all but the DKEFS sort (self-other) factor loadings were considered to be "poor".

Relationship Between Identified Factors and g

A final multiple regression was used, to examine the relationship between the factors identified in our EFA and our measure of fluid intelligence. Preliminary analyses were again conducted to ensure there were no violations of the assumptions of normality, linearity, multicollinearity, and homoscedasticity. This regression model accounted for 57.3% of the total variance in fluid intelligence scores $F(2, 117) = 78.468$, $p<.001$. The Factor 1 scores significantly predicted fluid intelligence ($\beta=.730$, $p<.001$) and explained 49.8% of this variance, whilst Factor 2 scores did not significantly predict fluid intelligence ($\beta=.088$, $p=1.64$). The correlations for this regression analysis are reported in Table 2.8.

Table 2.3: Descriptive Statistics for variables included in first hierarchical multiple regression

Measure	*Mean*	*Standard Deviation*	*N*
Sex	-	-	150
Education (years)	15.27	3.15	117
PHQ-9	7.46	5.86	144
GAD-7	11.78	5.18	144
Six elements	4.53	1.90	131
Zoo map 1 (errors)	5.05	1.11	148
WAIS-IV PRI	101.38	18.46	139

Table 2.4: Correlations for regression model including "naturalistic" executive tests

Measure	*WAIS-IV PRI*	*PHQ-9*	*GAD-7*	*Sex*	*Education (years)*	*Six elements*	*Zoo map 1 (errors)*
WAIS-IV PRI	-	-.092	-.072	-.07	.254	.419	.339
PHQ-9	-.092	-	.775	-.156	-.160	-.126	.046
GAD-7	-.072	.775	-	-.168	-.085	-.095	.061
Sex	-.07	-.156	-.168	-	.224	0.64	.004
Education	.254	-.160	-.085	.224	-	0.086	.077
Six elements	.419	.126	-.095	.064	.086	-	-.028
Zoo map 1 (errors)	.339	.046	.061	.004	-.028	-.028	-

Table 2.5: Descriptive statistics for variables included in second hierarchical multiple regression

Measure	*Mean*	*Standard Deviation*	*N*
Sex	-	-	150
Education (years)	15.27	3.15	117
PHQ-9	7.46	5.86	144
GAD-7	11.78	5.18	144
Digit span (backwards)	9.05	2.858	148
DKEFS sorting (correct)	9.63	3.331	147
Verbal fluency (phonemic)	8.54	4.106	145
Stroop B	44.68	39.789	147
DKEFS trail making task	7.61	4.008	148
WAIS-IV PRI	101.38	18.46	139

Table 2.6: Correlations for regression model including "traditional" executive tests

Measure	*WAIS-IV PRI*	*PHQ-9*	*GAD-7*	*Sex*	*Education (years)*	*Digit span (backwards)*	*DKEFS sorting (correct)*	*Verbal fluency (phonemic)*	*Stroop B*	*DKEFS trail making task*
WAIS-IV PRI	-	-.092	-.072	-.07	.254	.568	.586	.482	.487	.644
PHQ-9	-.092	-	.775	-.156	-.126	-.199	-.124	-.163	-.287	-.101
GAD-7	-.072	.775	-	-.168	-.095	-.119	-.073	-.024	-.164	-.072
Sex	-.07	-.156	-.168	-	0.64	-.022	.150	-.003	.095	-.013
Education (years)	.254	-.160	-.085	.224	-	.182	.395	.227	.221	.136
Digit span (backwards)	.568	-.199	-.119	-.022	.182	-	.499	.506	.409	.503
DKEFS sorting (correct)	.586	-.124	-.073	.150	.395	.499	-	.513	.403	.485
Verbal fluency (phonemic)	.482	-.163	-.024	-.003	.227	.506	.513	-	.425	.490
Stroop B	.487	-.287	-.164	.095	.221	.409	.403	.425	-	.483
DKEFS trail making task	.644	-.101	-.072	-.013	.136	.503	.485	.390	.483	-

Table 2.7: Rotated factor matrix from principal axis factoring

Measurement variable	***Factor 1***	***Factor 2***
DKEFS sorting (correct)	.754	.067[1]
Verbal fluency (switching)	.737	.106[1]
Verbal fluency (phonemic)	.702	.168[1]
DKEFS trail making	.695	.290[1]
Symbol span	.692	.310[1]
WAIS-IV PSI	.680	.317[1]
Similarities	.653	.079[1]
Digit span backwards	.646	.049[1]
Stroop B	.644	.146[1]
Six elements test	.432	.385
Zoo map 1 (errors)	.427	-.138[1]
DKEFS sort (recognition vs free sorting description)	-121[1]	.676

Notes: [1]factor loading <0.32, not significant

Table 2.8: Correlations for regression model including factor scores following EFA

	WAIS-IV PRI	*Factor 1 scores*	*Factor 2 scores*
WAIS-IV PRI	-	.752	.274
Factor 1 scores	.752	-	.256
Factor 2 scores	.274	.256	-

Discussion

This study examined the relationship between EF and *g* in a clinical sample of NHS TBI patients undergoing neuropsychological assessment. Although this relationship has been examined within the literature, there remains considerable debate as to the nature of EF, how it relates to *g,* and how we assess each in clinical practice. By using a clinical sample, we hoped to explore the real-world significance of this relationship and consider its implications for clinical practice.

We used two hierarchical multiple regression analyses to examine how well scores on EF tests, categorised as "traditional" or "naturalistic", could predict fluid intelligence after accounting for non-cognitive variables such as age, gender, education, mood, and anxiety. The first regression examined the relationship between two 'ecologically valid' executive tests, zoo map 1 (errors) and the six elements test, in comparison with *g*, and highlighted that these tests explained a significant amount of variance (27.3%, $p<.001$) in fluid intelligence, after accounting for the non-cognitive variables. This displays a clear link between *g* and 'ecologically valid' measures of EF. The second regression analysis, examining the relationship between five "traditional" tests and *g*, highlighted that these tests also explained a significant amount of variance in *g* (49.4%, $p<.001$), after accounting for non-cognitive variables, further displaying a clear relationship between *g* and EF. Five of seven measures of EF were found to be significant predictors of variance in fluid intelligence across these two regression analyses, with verbal fluency (phonemic) and Stroop B the only two measures to not show this relationship. This suggests that these two tasks may measure cognitive functions that are distinct from *g*. This is in line with previous research (Kahneman & Frederick, 2002; Shalice & Cipolotti, 2018) which has suggested a dual process model of cognitive functioning within the PFC, which accounts for automatic and non-automatic processes.

The factor structure of the executive tests used in the WOCRS neuropsychological assessment battery was also examined, using EFA, to provide further insight into the structure of EF. The results of this analysis suggested a clear two factor structure, where all but one EF test (DKEFS sort recognition vs free sorting description) showed significant loadings on one factor, while the second factor showed significant loadings for two EF tests (DKEFS sort recognition vs free sorting description, plus the six elements test). These findings are similar to those of Wood and Liossi (2007), who examined the relationship between 'ecologically valid' tests of EF and fluid intelligence in a sample of 118 TBI patients. Their factor analysis also showed that EF tests loaded onto two factors, and proposed that the most prominent factor was likely to be representative of *g*, whilst the less prominent factor was likely to represent specific executive components such as planning or multi-tasking.

To further investigate this relationship, we conducted a third regression analysis, extracting factor scores from the two factors shown by the PAF, and examining the extent to which they predicted variance in *g*. The results are particularly theoretically significant, as they showed that Factor 1 significantly predicted a large amount of variance (49.8%) in *g* ($p<.001$). In our factor analysis, the strong loadings on Factor 1 for the majority of executive tests is also likely to be representative of *g*. Conversely, Factor 2 did not predict significant variance in *g* ($\beta=.088$, $p=1.64$), indicating that this factor is distinct from the construct of *g*. It

is therefore likely that, similar to Wood and Liossi's (2007) results, Factor 2 is representative of specific cognitive components being tested by the EF tests which loaded onto Factor 2. In our model, this is represented by scores in tests which engage participants' ability to multi-task and take alternative perspectives.

The results of this study provide support for Duncan's (1995) hypothesis, that components of fluid intelligence are measured by tests which are considered to be representative of executive function. Given the comprehensive battery of executive tests included in our analyses, which are thought to be representative of various distinct EFs, and the results which indicate that EF tests do not account for all of the variance in *g*, it is likely that many current neuropsychological tests of EF instead provide a general indication of current fluid intelligence. Burgess et al. (2000) suggest that multi-tasking depends upon the most anterior portion of the PFC, BA 10, whereas other EFs recruit from other areas of the PFC. Although our findings cannot address the issue of localisation, the results of our factor analysis are consistent with the idea that multi-tasking is somewhat distinct from other EFs, supporting Burgess et al.'s (2000) findings. Similarly, it has been suggested that perspective-taking (associated with social cognition) is distinct from executive function, in that it has little relationship with set-shifting or executive control (Michael et al., 2018; Złotogórska-Suwińska & Putko, 2019). Our findings are of particular relevance to clinical assessment following TBI, as they suggest that it may be more prudent for clinicians to include a good measure of *g*, rather than a comprehensive battery of EF tests, alongside specific EF tests aimed at assessing discrete functions (e.g. planning, multi-tasking, perspective taking).

Study Strengths and Limitations

Unfortunately, due to limits on time and resources, two variables initially intended to be included as confounding variables in our regression models, time since injury, and socio-economic status, were not available for inclusion in our final analyses. These variables may both have meaningfully contributed to the development of an accurate regression model. Further, one of the intended analyses for this study was to examine the relationship between *g* loss (measured by TOPF-estimated premorbid PRI minus current WAIS-IV PRI), and the executive battery. The previous study by Roca et al. (2010) highlighted that performance in executive tests was attributable to losses in fluid intelligence, though found this within a non-clinical sample. However, our data showed that reading-based estimates of premorbid IQ were lower than observed post-TBI fluid IQ in a sizable proportion of the clinical sample. As it is unlikely that TBI served to improve fluid IQ, it seems reasonable to assume that the TOPF did not provide a reliable measure of estimated premorbid IQ within this clinical group (i.e. it had likely under-estimated premorbid ability, perhaps due to the cultural sensitivity of the measure, and the culturally and linguistically diverse population served by the service. Our analysis therefore excluded this variable. The under-estimation of premorbid ability raises questions about the validity of the TOPF and estimated-observed discrepancy analyses for future research.

The main limitation of this study is also its greatest strength. Utilising a clinical sample meant that the available data was somewhat inconsistently gathered. Patients in the sample received an assessment that was based on their individual presentation and needs. Not all patients received a standardised experience usually seen in clinical trials, as some tests were not relevant to clinical presentation. Similarly, it may have been difficult to meaningfully

interpret the results for some tests due to potential issues such as language or physical/motor functioning. This missing data reduced the sample size in our regression analyses, though both remained adequately powered. The clinical sampling does, however, represent the true diversity of patients presenting to neuropsychological rehabilitation studies, and provides novel insight into the realities of neuropsychological assessment within clinical settings. Our sample likely represents a more accurate cross-section of TBI patients in relation to education, employment, culture, and ethnicity, as we did not sample a homogeneous group, removing the sampling bias experienced by non-clinical research. This study has considerable generalisability due to its sampling techniques, and the variety of EF tests included in our analyses.

Future Research

Although all of our analyses were suitably powered, due to the nature of our clinical sample, there was some inconsistency in the data available to use for analyses. We therefore were unable to include important measures such as a measure of fluid intelligence loss. Future research examining this variable in the context of EF will provide further insight into the relationship between EF and *g*, and how this impacts neuropsychological assessment. Further studies using larger and different clinical samples, as well as a variety of EF tests would provide further insight into the nature of *g*, and how best to assess this in clinical practice. This study used the WAIS-PRI value as a measurement of fluid intelligence, as this is commonly used in clinical practice, however, the out-of-print Cattell culture fair has previously been used as a measure of *g* in research studies (Roca et al., 2010). Future research developing a gold standard measure of *g* for use in both clinical practice and academic research would help to provide future high-quality evidence, and provide clinicians with a valuable assessment tool.

Conclusions

The findings of this study have important implications for the clinical assessment of TBI. We found that commonly used EF tests are significantly correlated with fluid intelligence, and that when the factor structure of these tests is assessed, they load onto a two-factor model representative of fluid intelligence, and discrete cognitive functions (such as multi-tasking, and perspective-taking). Future research should focus on examining the relationship between fluid IQ loss and EF in a variety of clinical samples, developing a gold-standard measure of *g*, and improving the identification and measurement of executive functions that are distinct from *g*.

References

Burgess, P. W. (2004). Theory and methodology in executive function research. In *Methodology of Frontal and Executive Function* (pp. 87-121). Routledge.

Burgess, P. W., Alderman, N., Evans, J. O. N., Emslie, H., & Wilson, B. A. (1998). The ecological validity of tests of executive function. *Journal of the international Neuropsychological Society*, *4*(6), 547-558.

Burgess, P. W., Veitch, E., de Lacy Costello, A., & Shallice, T. (2000). The cognitive and neuroanatomical correlates of multitasking. *Neuropsychologia*, *38*(6), 848-863.

Catell, R. B. (1955). The scree test for number of factors. *Multivariate Behavioural Research, 1, 245-76*.

Chan, R. C. K., Shum, D., Toulopoulou, T., & Chen, E. Y. H.(2008). Assessment of executive functions: Review of instruments and identification of critical issues. *Archives of Clinical Neuropsychology*, *23*(2), 201–216.

Covington, N. V., & Duff, M. C. (2021). Heterogeneity is a hallmark of traumatic brain injury, not a limitation: A new perspective on study design in rehabilitation research. *American Journal of Speech-Language Pathology*, *30*(2S), 974-985.

De Winter, J. C., & Dodou, D. (2012). Factor recovery by principal axis factoring and maximum likelihood factor analysis as a function of factor pattern and sample size. *Journal of Applied Statistics*, *39*(4), 695-710.

Duncan, J., Assem, M., & Shashidhara, S. (2020). Integrated intelligence from distributed brain activity. *Trends in Cognitive Sciences*, *24*(10), 838-852.

Duncan, J., Emslie, H., Williams, P., Johnson, R., & Freer, C. (1996). Intelligence and the frontal lobe: The organisation of goal-directed behaviour. *Cognitive Psychology, 30*, 257–303.

Duncan, J., Johnson, R., Swales, M., & Freer, C. (1997). Frontal lobe deficits after head injury: Unity and diversity of function. *Cognitive Neuropsychology*, *14*(5), 713-741.

Duncan, J., & Owen, A. M. (2000). Common regions of the human frontal lobe recruited by diverse cognitive demands. *Trends in Neurosciences*, *23*(10), 475-483.

Gilbert, S. J., & Burgess, P. W. (2008). Executive function. *Current Biology*, *18*(3), R110-R114.

Hampshire, A., Highfield, R. R., Parkin, B. L., & Owen, A. M. (2012). Fractionating human intelligence. *Neuron*, *76*(6), 1225-1237.

Harvey, P. D. (2022). Clinical applications of neuropsychological assessment. *Dialogues in clinical neuroscience*.

Horn, J. L. (1967). Intelligence-why it grows, why it declines. *Transaction*, *5*(1), 23-31.

IBM Corp. (2021). IBM SPSS Statistics for Windows (Version 28.0). *Computer software*. IBM Corp.

Jakobsen, J. C., Gluud, C., Wetterslev, J., & Winkel, P. (2017). When and how should multiple imputation be used for handling missing data in randomised clinical trials–a practical guide with flowcharts. *BMC Medical Research Methodology*, *17*(1), 1-10.

Kaiser, H. (1970). A Second generation Little Jiffy. *Psychometrika, 35,* 401-415.

Kaiser, H. (1974). An index of factorial simplicity. *Psychometrika, 39,* 31-36.

Leitgeb, J., Mauritz, W., Brazinova, A., Janciak, I., Majdan, M., Wilbacher, I., & Rusnak, M. (2011). Effects of gender on outcomes after traumatic brain injury. *Journal of Trauma and Acute Care Surgery*, *71*(6), 1620-1626.

Maas, A. I., Menon, D. K., Adelson, P. D., Andelic, N., Bell, M. J., Belli, A., ... & Francony, G. (2017). Traumatic brain injury: integrated approaches to improve prevention, clinical care, and research. *The Lancet Neurology*, *16*(12), 987-1048.

Manly, T., Hawkins, K., Evans, J., Woldt, K., & Robertson, I. H. (2002). Rehabilitation of executive function: facilitation of effective goal management on complex tasks using periodic auditory alerts. *Neuropsychologika*, *40*(3), 271-281.

Mckee, A. C., & Daneshvar, D. H. (2015). The neuropathology of traumatic brain injury. *Handbook of Clinical Neurology, 127,* 45-66.

Michael, J., Wolf, T., Letesson, C., Butterfill, S., Skewes, J., & Hohwy, J. (2018). Seeing it both ways: Using a double-cuing task to investigate the role of spatial cuing in Level-1 visual perspective-taking. *Journal of Experimental Psychology: Human Perception and Performance*, *44*(5), 693–702.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology*, *41*(1), 49-100.

Monsell, S. (1996). Control of mental processes. *Unsolved mysteries of the mind: Tutorial essays in cognition* (pp. 93–148). Hove, UK: Erlbaum.

National Statistics Office. (2019) Indices of Deprivation. *Ministry of Housing, Communities & Local Government*. Accessed from: https://data.london.gov.uk/dataset/indices-of-deprivation

Ng, S. Y., & Lee, A. Y. W. (2019). Traumatic brain injuries: pathophysiology and potential therapeutic targets. *Frontiers in cellular neuroscience*, *13*, 528.

Roberts, A. C. (1996). Executive and cognitive functions of the prefrontal cortex-Introduction. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences, 351*(1346), 1389-1395.

Roca, M., Parr, A., Thompson, R., Woolgar, A., Torralva, T., Antoun, N., Manes, F., & Duncan, J. (2010). Executive function and fluid intelligence after frontal lobe lesions. *Brain*, *133*(1), 234-247.

Savoy, R. L. (2001). History and future directions of human brain mapping and functional neuroimaging. *Acta Psychologika*, *107*(1-3), 9-42.

Shallice, T. (1988). *From neuropsychology to mental structure*. Cambridge University Press.

Shallice, T., & Cipolotti, L. (2018). The prefrontal cortex and neurological impairments of active thought. *Annual Review of Psychology*, *69*, 157-180.

Sharp, D. J., Beckmann, C. F., Greenwood, R., Kinnunen, K. M., Bonnelle, V., De Boissezon, X., ... & Leech, R. (2011). Default mode network functional and structural connectivity after traumatic brain injury. *Brain*, *134*(8), 2233-2247.

Sharp, D. J., Scott, G., & Leech, R. (2014). Network dysfunction after traumatic brain injury. *Nature Reviews Neurology*, *10*(3), 156-166.

Spearman. C. (1904). General intelligence, objectively determined and measured. American Journal of Psychology; 15: 201–93.

Spearman. C. (1927). The Abilities of Man, their Nature and Measurement.

Stuss, D.T., & Alexander, M. P. (2007). Is there a dysexecutive syndrome? Philosophical transactions of the royal society B: Biological Sciences, 362 (1481), 901-915.

Stuss, D. T. (2011). Traumatic brain injury: relation to executive dysfunction and the frontal lobes. *Current opinion in neurology*, *24*(6), 584-589.

Stuss, D. T. (2011). Functions of the frontal lobes: relation to executive functions. *Journal of the International Neuropsychological Society. 17*, 759-765

Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2013). *Using multivariate statistics* (Vol. 6, pp. 497-516). Boston, MA: Pearson.

Wilson, B. A., Alderman. N., Burgess. P. W., Emslie. H., & Evans, J. J. (1996). *Behavioural assessment of the dysexecutive syndrome.* St Edmunds UK: Thames Valley Test Company.

Wood, R. L. & Liossi, C. (2007). The relationship between general intellectual ability and performance on ecologically valid executive tests in a severe brain injury sample. *Journal of the International Neuropsychological Society, (13)* (1), 90-98.

Złotogórska-Suwińska, A., & Putko, A. (2019). Visual Perspective-Taking: Are Theory of Mind and Executive Function Involved?. *Polish Psychological Bulletin*, *50*(2).

Appendices:

Appendix 1: Details of search strategy development for systematic review

Search Iteration	***String 1: Terms related to Brain Injury***	***String 2: Terms Related to Neuropsychological Assessment***	***String 3: Terms related to Zoo Map***
Search Strategy 1 *(originally developed to apply criteria for ABI studies)*	((Brain Injur* adj3 (Acquir* OR Trauma* OR ABI OR TBI) OR Head Injur* OR Brain Tumo* OR Stroke OR Meningitis OR Septicemia OR encephalitis OR Contusion OR Diffuse Axonal Injury OR Anoxic Injury OR Hypoxic Injury).tw	((Cognit* adj2 (Assess* OR test* OR task OR eval* OR measur*) OR (Neuro* adj2 (assess* OR test* OR task OR eval* OR measur*) (Cognit* adj2* (assess* OR test* OR Task OR eval* OR measur*)).tw	(Zoo Map OR ZM OR Behavio?r* Assessment of the Dysexecutive Syndrome OR BADS)
Search Strategy 2 *(Following change to focus on TBI studies)*	((Brain Injur* adj3 (OR Trauma* OR TBI) OR Head Injur* OR Contusion OR Diffuse Axonal Injury OR Anoxic Injury OR haematoma OR hemorrhage Hypoxic Injury)).tw	((Cognit* adj2 (Assess* OR test* OR task OR eval* OR measur*) OR (Neuro* adj2 (assess* OR test* OR task OR eval* OR measur*) (Cognit* adj2* (assess* OR test* OR Task OR eval* OR measur*)).tw	(Zoo Map OR ZM OR Behavio?r* Assessment of the Dysexecutive Syndrome OR BADS)
Search Strategy 3 *(Final Strategy focussing solely on Zoo Map terms to increase sensitivity)*	Removed	Removed	Zoo Map OR ZM OR Behavio?r* Assessment of the Dysexecutive Syndrome OR BADS

Appendix 2: Summary of methodological quality

Methodological Quality Assessment of Included Studies

Property	**Study**	**Overall Rating**	**Area With Lowest Rating**	**Main Limitation**
Convergent Validity	Spikman et al., 2012	Doubtful	Design	No information provided on the measurement properties of comparison instruments for any population.
	Weddell & Wood, 2016	Doubtful	Design	Some information on the measurement properties of comparison instruments on any population, but conflicting evidence.
	Wood & Liossi, 2006	Very Good	N/A	N/A
	Wood & Liossi, 2007	Very Good	N/A	N/A
Ecological Validity	Hendry et al., 2016	Adequate	Statistical method	Used appropriate statistical analyses, but had low sample size for regression analysis.
	Weddell & Wood, 2016	Very Good	N/A	N/A
	Weddell & Wood, 2016	Doubtful	Design	No information provided on the measurement properties of comparison instruments (Frontal and Social Behaviour Scale) on any population.
	Wood & Liossi, 2006	Very Good	N/A	N/A
Known-Group Validity	Rakers et al., 2018	Adequate	Statistical Method	Did not fully report details of between group comparison, as focussed on intervention outcome. Can only score as “assumable that statistical methods were appropriate”.
	Spikman et al., 2012	Very Good	N/A	N/A

	Westerhof-Evers et al., 2019	Doubtful	Statistical method	Reporting insufficient as did not fully report details of between group comparison. It was therefore unclear whether parametric, or non-parametric test was used to determine results, and effect size could not be calculated.
Responsiveness to Change	Ownsworth et al., 2017	Adequate	Statistical method	Appropriate statistical test used, gave appropriate rationale for use of ZM version 1 over profile score, though methods used to highlight between-group differences rather than within-group differences.

Appendix 3: Rating per psychometric property

Convergent Validity

Study	**Lowest Quality Score/Category**	**Sample (n)**	**Comparator Measure**	**Results: Correlation between ZM and comparator measures (r=)**		**Rating (+/-/?)^**
				ZM Version 1	ZM Profile Score	
Spikman et al., 2012	Doubtful/Design	28	FEEST	N/A	0.33	-
Spikman et al., 2012	Doubtful/Design	28	Cartoon Test	N/A	0.44	-
Spikman et al., 2012	Doubtful/Design	28	Faux Pas (Detection)	N/A	0.45	-
Spikman et al., 2012	Doubtful/Design	28	Emotional Empathy Questionnaire	N/A	-0.03	-
Spikman et al., 2012	Doubtful/Design	28	Faux Pas (Empathy)	N/A	0.33	-
Weddell & Wood, 2016	Doubtful	71	AX	N/A	-0.22	-
Weddell & Wood, 2016	Very Good	71	Barratt	N/A	-0.21	-
Wood & Liossi, 2006	Very Good	59	Hayling A	N/A	0.13	-
Wood & Liossi, 2006	Very Good	59	Hayling B	N/A	0.15	-
Wood & Liossi, 2006	Very Good	59	Hayling C	N/A	0.36*	+
Wood & Liossi, 2006	Very Good	59	Brixton	N/A	0.09	+

Wood & Liossi, 2006	Very Good	59	Key Search	N/A	0.63*	-
Wood & Liossi, 2006	Very Good	59	WAIS-FSIQ	N/A	0.37*	+
Wood & Liossi, 2006	Very Good	59	WAIS-VIQ	N/A	0.33*	+
Wood & Liossi, 2006	Very Good	59	WAIS-PIQ	N/A	0.35*	+
Wood & Liossi, 2007	Very Good	118	WAIS-VIQ	N/A	0.35*	+
Wood & Liossi, 2007	Very Good	118	WAIS-VIQ	N/A	0.31*	+
Wood & Liossi, 2007	Very Good	118	WAIS-PIQ	N/A	0.34*	+
Wood & Liossi, 2007	Very Good	118	Hayling A	N/A	0.29*	+
Wood & Liossi, 2007	Very Good	118	Hayling B	N/A	0.26*	+
Wood & Liossi, 2007	Very Good	118	Hayling C	N/A	0.34*	+
Wood & Liossi, 2007	Very Good	118	Brixton	N/A	0.11	-
Wood & Liossi, 2007	Very Good	118	Key Search	N/A	0.59*	+
Wood & Liossi, 2007	Very Good	118	COWAT	N/A	0.16	-
Wood & Liossi, 2007	Very Good	118	Trail Making Task (A)	N/A	0.02	-

Wood & Liossi, 2007	Very Good	118	Trail Making Task (B)	N/A	0.15	-
Wood & Liossi, 2007	Very Good	118	Hayling A^{+}	N/A	0.18	+
Wood & Liossi, 2007	Very Good	118	Hayling B$^{=}$	N/A	0.18*	+
Wood & Liossi, 2007	Very Good	118	Hayling C^{+}	N/A	0.27*	+
Wood & Liossi, 2007	Very Good	118	Brixton^{+}	N/A	0.02	-
Wood & Liossi, 2007	Very Good	118	Key Search^{+}	N/A	0.52*	+
Wood & Liossi, 2007	Very Good	118	COWAT	N/A	0.03	-
Wood & Liossi, 2007	Very Good	118	Trail Making Task (B)	N/A	0.23*	+

Abbreviations: HBCT = Home Based Cooking Task, FEEST = Facial Expressions of Emotion-Stimuli and Tests, Barratt = Barratt Impulsiveness Scale, DEX-S/I = Dysexecutive Questionnaire-Self/ Other, F&SB = Frontal and Social Behaviour Scale, BDI = Beck Depression Inventory, HADS-A = Anxiety Scale from Hospital Anxiety and Depression Scale, AX = STAXI-II Anger Expression Index, WAIS-FSIQ/VIQ/PIQ = Wechsler Adult Intelligence Scale Full Scale IQ Score/Verbal IQ score/Performance IQ score

Notes: *P<.05, $^{+}$controlled for g, ^(+) sufficient, (?) indeterminate, (-) insufficient

Ecological Validity

Study	**Lowest Quality Score/Category**	**Sample (n)**	**Comparator Measure**	**Results: Correlation between ZM and comparator measures (r=)**		**Rating (+/-/?)^**
Hendry et al., 2016	Adequate/Statistical method	45	HBCT Total Errors	0.04	-0.06	-
Hendry et al., 2016	Adequate/Statistical method	45	HBCT Omissions	0.31*	-0.05	+
Hendry et al., 2016	Adequate/Statistical method	45	HBCT Additions	-0.27	-0.04	-
Hendry et al., 2016	Adequate/Statistical method	45	HBCT Sequence/Substitutions	-0.18	-0.17	-
Hendry et al., 2016	Adequate/Statistical method	45	HBCT Estimations	0.34*	-0.027	+
Hendry et al., 2016	Adequate/Statistical method	45	HBCT Commentary/Questions	0.32*	-0.05	+
Hendry et al., 2016	Adequate/Statistical method	45	HBCT Goal Achievement	0.04	-0.14	-
Hendry et al., 2016	Adequate/Statistical method	45	HBCT Dangerous Behaviour	0.11	-0.08	-
Hendry et al., 2016	Adequate/Statistical method	45	HBCT Completion Time	-0.16	0.28	-
Weddell & Wood, 2016	Very Good	71	DEX	N/A	-0.33*	+
Weddell & Wood, 2016	Doubtful/Study design	71	F&SB	N/A	-0.17	-
Wood & Liossi, 2006	Very Good	59	DEX-S	N/A	-0.17	+

Wood & Liossi, 2006	Very Good	59	DEX-I		N/A	-0.11	-

Known-Group Validity

Study	**Lowest Quality Score/Category**	**Sample (n, group)**			**Between group difference on ZM scores (direction and Cohen's D/T/Z value)[1]**		**Rating (+/-/?)**
		Group 1	**Group 2**	**Group 3**	**ZM Version 1**	**ZM Profile Score**	
Rakers et al. (2018)	Adequate/Statistical method	51, HC	47, mTBI	59, msTBI	Group 3 lower than 1 & 2, D=0.443*	N/A	+
Spikman et al. (2012)[1]	Very Good	55, HC	28, msTBI	N/A	Group 2 lower, D=0.57*	N/A	+
Westerhof-evers et al. (2019)	Doubtful/Statistical method	45, HC	63, msTBI	N/A	Group 2 lower, T/Z=2.11	N/A	-

Abbreviations: HC = Healthy Controls, mTBI = mild Traumatic brain injury, msTBI = moderate-severe TBI

Notes:*p ≤ .05, [1]Reporting of results for this study was unclear whether T or Z scores were reported.

Responsiveness to change

Study	**Lowest Quality Score/Category**	**Sample (n, group)**		**Mean within group difference on ZM scores, Cohen's d**		**Rating (+/-/?)**
		Group 1	**Group 2**	**Group 1**	**Group 2**	
Ownsworth et al. (2017)	Adequate/Study design	27, EBL	27, ELL	1.75, 0.298	0.66, 0.298	-

Abbreviations: EBL = Error-Based Learning, ELL = Errorless learning

Appendix 4: Details of psychometric properties analysed in review

Psychometric property	***Definition of property***	***Example in relation to Zoo Map***
Convergent validity	The degree to which the scores in a tool correspond to other instruments purporting to measure the same construct	Zoo Map compared to other EF measures
Ecological validity	The degree to which the results of an outcome measure accurately reflect the impact on real-world functioning	Zoo Map purports to represent planning behaviour, but does this represent the true challenges associated with planning in real-world setting
Known-Group validity	The degree to which an instrument produces different scores for groups know to vary on the variables being measured	Moderate-severe TBI compared to mild TBI
Responsiveness to Change	The ability of a tool to detect change in the construct to be measured over time	Zoo Map score calculated pre-intervention and following intervention

Appendix 5: Summary of GRADE quality of evidence ratings

Psychometric property	***Grade score***	***Reasons for downgrading***
Convergent validity	Could not be determined	N/A
Ecological validity	High	N/A
Known-Group validity	Moderate	Inconsistency (-1 point)
Responsiveness to Change	Low	Risk of bias -1 point Imprecision -1 point

Appendix 6: Prisma checklist for systematic reviews

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	8
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	9
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	10
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	12
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	12
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	13
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	12
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	13
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	13

Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	14
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	14
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	N/A
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	N/A

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	14
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	N/A
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	16
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	17
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	17,18,19
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	19
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	N/A
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	17, 18, 19
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	N/A
DISCUSSION			

Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	23,24
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	24
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	25
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	N/A

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

For more information, visit: **www.prisma-statement.org**.

Appendix 7: Approved research proposal

Link: https://osf.io/z6dv5

Appendix 8: Letter of ethical approval

Ymchwil Iechyd a Gofal Cymru
Health and Care Research Wales

NHS
Health Research Authority

Dr Jessica Fish
Institute of Health and Wellbeing
College of Medical, Veterinary and Life Sciences
University of Glasgow 1st floor
1055 Great Western Road
G12 0XH

Email: approvals@hra.nhs.uk
HCRW.approvals@wales.nhs.uk

13 February 2023

Dear Dr Fish

HRA and Health and Care Research Wales (HCRW) Approval Letter

Study title:	**The Assessment of Executive Function and Fluid Intelligence in Patients With Traumatic Brain Injury**
IRAS project ID:	**317718**
Protocol number:	**317718**
HRA reference:	**23/HRA/0185**
Sponsor	**University of Glasgow**

I am pleased to confirm that **HRA and Health and Care Research Wales (HCRW) Approval** has been given for the above referenced study, on the basis described in the application form, protocol, supporting documentation and any clarifications received. You should not expect to receive anything further relating to this application.

Please now work with participating NHS organisations to confirm capacity and capability, in line with the instructions provided in the "Information to support study set up" section towards the end of this letter.

How should I work with participating NHS/HSC organisations in Northern Ireland and Scotland?
HRA and HCRW Approval does not apply to NHS/HSC organisations within Northern Ireland and Scotland.

If you indicated in your IRAS form that you do have participating organisations in either of these devolved administrations, the final document set and the study wide governance report (including this letter) have been sent to the coordinating centre of each participating nation. The relevant national coordinating function/s will contact you as appropriate.

Appendix 9: Scree plot from principal axis factoring analysis