Performance validity testing in an NHS acquired brain injury sample

Anna Isherwood

D. Clin. Psy. Thesis (Volume 1), 2017

University College London

UCL Doctorate in Clinical Psychology

Thesis declaration form

I confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signature:

Name: Anna Isherwood

Date: 18th October 2017

Overview

This thesis explores the utility of established performance validity tests (PVTs) for assessing non-neurological aspects of test-taking behaviour in clinical populations. Part one is a systematic literature review examining the classification accuracy of a range of PVTs in dementia populations. Consideration is given to the impact of dementia severity and diagnosis on PVT specificity. Issues surrounding the use of cut-off scores on these measures are discussed. Part two is an empirical study exploring the base rate of failure on commonly-used PVTs in a sample of NHS acquired brain injury patients. The relationship between PVT failure and overall performance on cognitive testing is also assessed. Part three is a critical appraisal of the thesis which reflects on the broader clinical and conceptual challenges inherent in the field of performance validity testing as a whole.

Table of contents

Tables and Figures

Part 1: Literature Review

Part 2: Empirical Paper

List of abbreviations

AD	Alzheimer's disease
BPS	British Psychological Society
DS-SS	Digit span scaled score
FTD	Fronto-temporal dementia
PVT	Performance validity test
RDS	Reliable digit span
TBI	Traumatic brain injury
TOMM	Test of Memory Malingering
VD	Vascular dementia
WAIS	Wechsler Adult Intelligence Scale
WMT	Word Memory Test

Acknowledgements

I would like to say thank you to everyone who contributed their time and expertise to enable me to complete this thesis. Particular thanks go to my supervisors Sanjay Sunak and John King for their constructive commentary, lively discussions and patient guidance. I owe significant gratitude to Jessica Hooker and Kelly Llanfear for their contribution to the completion of the archive database, and to staff and patients in the neuropsychology departments at both Homerton University Hospital and St George's Hospital for supporting data collection. Much-deserved thanks go to my parents, who fostered my desire to learn and who have provided unwavering support throughout my academic career. Huge appreciation also goes to my husband, Conor. I am truly grateful for everything you have done to support me on my training journey. A final, special, thanks goes to Elisabeth for tolerating my absence. I dedicate this to you.

Part 1: Literature review

The assessment of performance validity in dementia: a systematic review

Abstract

Aims: Performance validity tests (PVTs) are included in neuropsychological test batteries to assess if results offer a reliable estimate of the individual's cognitive ability. Research exploring how individuals with dementia perform on these measures is limited. This review aims to synthesise the literature on the use of PVTs in dementia, with particular focus on the identified specificity of these measures across different severities and diagnoses.

Method: Systematic electronic searches were conducted on PsychINFO and MEDLINE databases to identify studies utilising validated PVTs with dementia samples. Methodological quality of the studies was assessed using the 'QualSyst' critical appraisal tool. PVT specificity data was subsequently extracted and pooled across studies to assess classification accuracy.

Results: Twenty-four studies were identified which investigated a total of 31 PVTs. Pooled specificity was examined for 11 PVTs. Only Vocabulary Minus Digit Span, Trail-Making Test Ratio and Coin in the Hand tests achieved adequate specificities of over 90%. PVTs with dementia profile algorithms, however, showed consistently higher classification accuracy. Results pointed to the potential influence of dementia severity and diagnosis on PVT failure.

Conclusions: The majority of PVTs used in UK clinical practice demonstrated inadequate specificity when used with clinical dementia populations. There is further evidence to suggest that PVTs may be sensitive to cognitive impairment. Adjusting cut-off scores has been found to improve specificity, but the concomitant reduction in sensitivity brings their ability to identify invalid performance into question. The clinical implications of these findings are discussed.

Introduction

'Dementia' is a broad term encompassing a number of progressive neurological conditions characterised by deterioration in multiple cognitive domains including memory, executive functions and language (Salmon & Bondi, 2009). Neuropsychological assessments are a performance-based method of assessing cognitive functioning frequently completed as part of diagnostic evaluations (Salmon & Bondi, 2009; MKhann et al., 2011; Sperling et al., 2011). They can provide important evidence to support the differential diagnosis of dementias with diverse cognitive profiles and to delineate between dementia and alternative causes of memory impairment (Harvey, 2012). They can also help clinicians identify strengths and weaknesses to inform clinical recommendations for treatment in addition to tracking longitudinal changes in cognition (Walter, Morris, Swier-Vosnos, & Pliskin, 2014). Given that neuropsychological assessment is a primary source of information in dementia diagnosis and management, it is of particular importance that the contributing test data offer a reliable estimate of a person's ability. This review will focus on the assessment of invalid performance and the application of conventional assessment tools to individuals with suspected or confirmed dementia diagnoses.

The concept of examinee 'effort'

Neuropsychological test data are prone to a range of biases which must be considered potential invalidating factors (Rudman, Oyebode, Jones, & Bentham, 2011). Test performance can be impaired for a number of reasons apart from neurological disease (Vickery, Berry, Inman, Harris, & Orey, 2001). One key factor is whether the examinee provided the clinician with a full and accurate impression of their symptoms and applied themselves adequately – or exerted maximum 'effort' -

during psychometric testing. It has been recognised that clinicians' ability to make *ad hoc* informal judgements of examinee 'effort' is poor (Faust, Hart, Guilmetter, & Arkes, 1988), indeed, it is recognised that the true extent of someone's motivation to apply themselves to testing is essentially unknowable (Suesse et al., 2015). Nonetheless, clinicians administering neuropsychological assessments have a responsibility to make determinations about test validity (Bush et al., 2005) and, as such, need to consider how such evaluations are made. In line with this, a 2009 document published by the British Psychological Society (BPS; McMillan et al., 2009), highlighted that addressing the issue of data quality is good practice in neuropsychological assessment. These guidelines emphasise that clinical assessments should routinely include valid and reliable indices sensitive to 'distortions of motivation', frequently referred to as 'effort tests'. In turn, this has led to a surge in the development of novel measures to meet this need.

The term 'effort testing' is ubiquitous, however, it is part of a nomenclature derived from medicolegal literature which frequently conflates 'poor effort' with malingering – the intentional production of exaggerated complaints, usually in the presence of external incentive (Slick, Sherman, & Iverson, 1999). This is misleading, and demonstrates a misconception about the properties of such tests, which can only observe and measure behaviour, not intent (Boone et al., 2002a). Using such terminology risks communicating over-simplistic categorisations: 'good effort' or 'poor effort' clients may be seen to have 'credible' or 'non-credible' performance which could, in turn, lead to potentially unhelpful interpretations regarding the source of such behaviour (for example that distortion in performance is intentional). To mitigate this issue, Bush et al. (2005) utilise the phrase 'investment in performing at capacity levels' rather than 'effort', which highlights a broader construction than

malingering alone, as there may be myriad reasons for reduced performance such as fatigue, side effects of medication or depressed mood (Bortnick, Horner, & Bachman, 2013; Walter et al., 2014), factors which are highly relevant in clinical settings. To reflect this position, this review will henceforth use the term 'performance validity' and associated 'performance validity tests (PVTs)' as per Larrabee (2012), who distinguishes the validity of ability task performance (as captured by the measures discussed below) from the accuracy of symptomatic complaint (as portrayed on symptom self-report measures, which are not examined in this review).

The measurement of performance validity

PVTs appear subjectively difficult to the examinee but are in fact measures designed to have a very low test ceiling. They purport to be insensitive to actual cognitive dysfunction, meaning that the majority of people applying themselves to the testing should achieve scores at or near the ceiling regardless of underlying neurological impairment (Tombaugh, 1997). Scores below chance on these measures are considered indicative of potential malingering when occurring in the presence of an external incentive (Slick et al., 1999). Anyone performing below ceiling and 'failing' the measures, however, would also raise questions about additional influences on performance and, concomitantly, the validity of the remaining neuropsychological data gathered in the assessment.

There are a range of PVTs available. Some are designed specifically for this purpose, such as the Test of Memory Malingering (TOMM; Tombaugh, 1996) or the Word Memory Test (WMT; Green, Allen, & Astner, 1996). These are referred to as 'standalone' PVTs and frequently take the form of recognition memory measures

using a 'forced choice' paradigm. There are also conventional psychometric tests which have established supplementary utility as PVTs. For example, Reliable Digit Span (RDS) can be computed from the Weschler Adult Intelligence Scale (WAIS) Digit Span subtest (Greiffenstein, Baker, & Gola, 1994), with the rationale that performance well below the 'normal' range on both forwards and backwards Digit Span tasks is relatively unusual in bona fide patients with documented brain damage (Iverson & Tulsky, 2003). These 'embedded' measures have the additional benefits of being less vulnerable to coaching and not adding to assessment times (McMillan et al., 2009). It has been recommended that clinicians take a 'multi-method, multi-test' approach to assessing performance validity, using a number of measures testing different cognitive domains (Larrabee, 2014).

Sensitivity and specificity: the psychometric foundations of performance validity testing

Test validity concerns the ability of a test to detect the presence or absence of a particular characteristic in the examinee. The validity of PVTs is evaluated using the psychometric concepts of sensitivity and specificity which indicate the likelihood of obtaining a true or a false positive respectively. The balance between the two is related to the cut-off score on the test: with PVTs, more stringent cut-offs are more likely to correctly identify those not performing at capacity levels (better sensitivity) but the resulting pool of individuals will likely include individuals who *are* performing at capacity levels (reduced specificity; Iverson & Brooks, 2011, p. 931). In clinical settings, the focus should be on maximising specificity over sensitivity to minimise the risk of false positive errors, as there are possible clinical, financial, occupational and emotional consequences for the patient associated with being labelled as having 'suspect effort' or 'invalid performance' (Greve & Bianchini,

2004). As such, within the PVT literature the trend has been to set a threshold for adequate test specificity at 90 per cent or greater (Vickery et al., 2001; Larrabee, 2008).

Measuring performance validity in dementia

The prolific body of literature demonstrating the efficacy of both standalone and embedded PVTs stems primarily from North American forensic and medicolegal contexts, where the examinees most frequently present with acquired brain injury and there are issues of potential secondary gain (for example, financial compensation or insurance claims). Little is known about their utility in clinical populations with confirmed or suspected diagnoses of dementia (Bortnick et al., 2013). Although many PVTs have been validated with cognitively impaired samples, these frequently exclude individuals with dementia, making it difficult to ascertain if reported cut-offs are appropriate for use in this population (Dean, Victor, Boone, Philpott, & Hess, 2009). A number of reasons have been proposed for this, for example the belief that older adults with dementia would rarely be incentivised to perform poorly (Kiewel, Wisdom, Bradshaw, Pastorek, & Strutt, 2012), though this perspective fails to recognise that a) there are myriad reasons (both explicit and implicit) why someone might be motivated to obtain a particular diagnostic status and b) invalid performance does not necessarily represent intentional underperformance. Although it is theoretically easy for most neurological patients to meet the task demands of PVTs it is unclear if this is the case in dementia, particularly in the more advanced stages. It may therefore represent a serious diagnostic error to conclude overall invalid performance based on PVT failure in this population. For these reasons, it is important to analyse the classification accuracy of PVTs in dementia populations across the spectrum of severity (Henry, Merten, Wolf, & Harth, 2010). Given the

growing incidence of dementia and the need for cognitive testing to support early diagnosis (Illife, Manthorpe, & Eden, 2003), in addition to professional guidelines highlighting the importance of performance validity testing for comprehensive neuropsychological assessment, this gap in the literature presents a challenge for clinicians.

The aim of the present literature review

To date, no systematic review has been published focusing specifically on performance validity measurement in dementia. A 2009 study by Dean and colleagues included an overview of the existing literature which indicated some promise for Digit Span indicators (Vocabulary Minus Digit Span, Iverson & Tulsky, 2003 and four-digit forward span time score, Babikian, Boone, Lu, & Arnold, 2006), the Medical Symptom Validity Test (MSVT; Green, 2004) and the Trail Making Test (TMT; Iverson, Lange, Green & Franzen, 2002). However, they emphasised the paucity of evidence in dementia samples. Since 2009, the PVT literature has burgeoned, and there is increased recognition that it is of clinical significance to identify which PVTs can be usefully employed in patients with possible dementia. The overall aim of the current review is therefore to synthesise current information regarding the performance of dementia patients on commonly-used PVTs. The key questions that this review seeks to address are as follows:

1) Which PVTs offer the greatest degree of classification accuracy (specificity) in a dementia population?
2) Should clinicians be using adjusted cut-offs to achieve adequate specificity on these PVTs in dementia populations?
3) Is PVT classification accuracy impacted by dementia severity or type?

Method

Search strategy

Initial searches were conducted on the electronic databases PsychInfo and Medline to identify relevant studies published up to August 2017. Three umbrella search term categories were identified (see Table 1), with additional search terms identified from key words in existing research studies. Terms were initially entered separately and then combined. Results were limited to human adult participants (over 18 years of age), written in the English language and peer-reviewed journals. It is recognised that utilising only published data may introduce a publication bias, however, this increased the likelihood that only high-quality data were included. Reference lists of included articles were also hand-searched to identify additional studies not picked up by the electronic search. This included a consultation of relevant review papers in this area (e.g. Dean et al, 2009).

Inclusion and exclusion criteria

The inclusion and exclusion criteria were guided by previous review papers in the PVT literature (e.g. Vickery et al., 2001; Sollman & Berry, 2011) to maximise the quality and specificity of the studies included. A flow diagram is provided (see Figure 1) to indicate where studies were eliminated from the final literature pool.

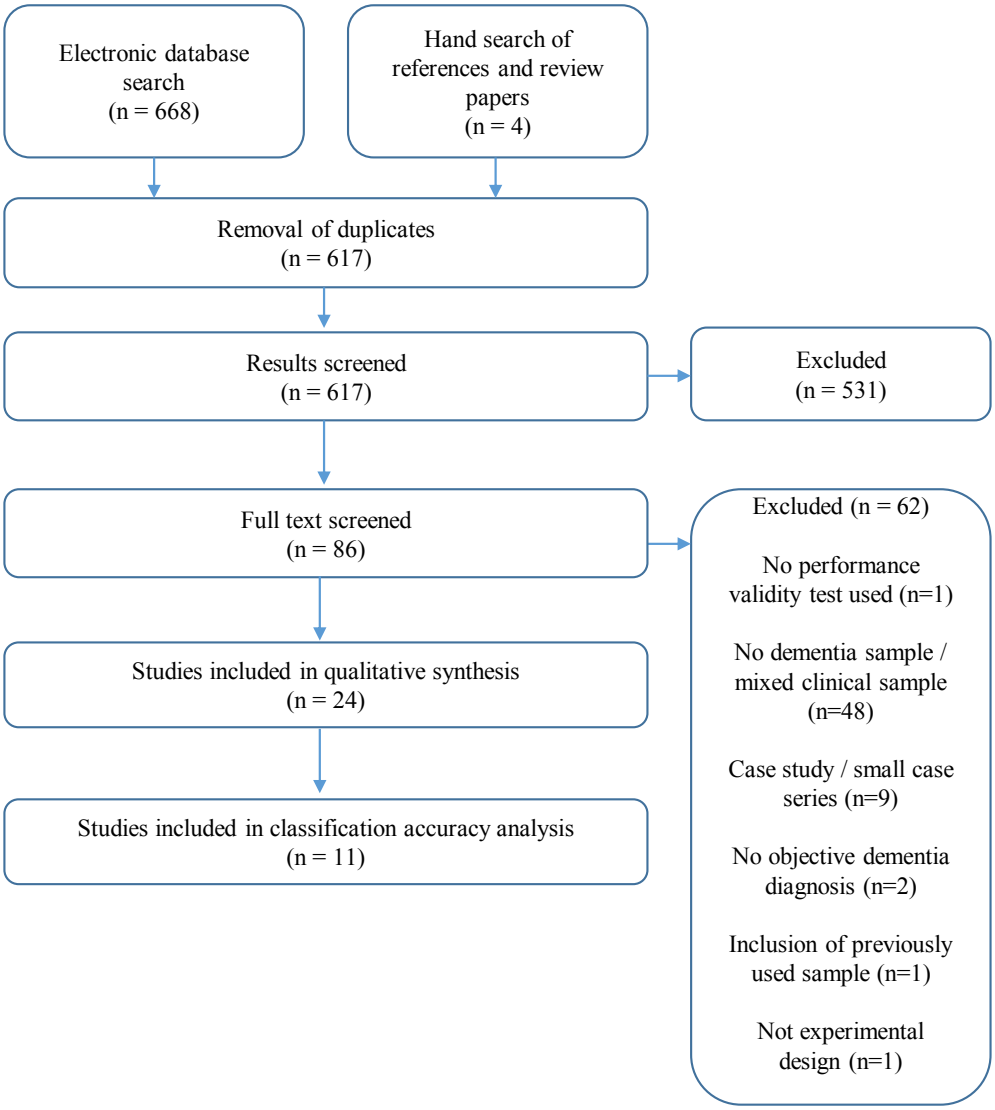
Measures used: Studies were selected if they investigated at least one established embedded or stand-alone PVT (see McMillian et al., 2009, pp 6-7 and Appendix 3 for information on commonly used PVTs). Note that measures are not described in full in this review to protect the fidelity of these instruments.

Table 1: *Electronic search strategy*

Search term category	Terms applied	Combined with
Performance validity test	Performance validity test*	OR
	Symptom validity test*	
	Effort test*	
	Validity indicator*	
Dementia	Dementia	AND
	Alzheimer?s disease	
	Vascular dementia	
	Fronto?temporal dementia / FTD	OR
	Memory disorder*	
	Mild cognitive impairment / MCI	
Symptom validity	Symptom validity	AND
	Performance validity	
	Malingering	OR
	Suboptimal effort	
	Response bias	

Notes: *Denotes truncation, looks for variants of words such as test, tests, testing

Figure 1: *Flow diagram illustrating the inclusion of studies in the review*



Study design: Both retrospective and prospective studies were included. Single case designs or case series were excluded. Some studies were found to use the same sample over a number of different studies: these data were included where the studies contributed unique information (for example, data on different measures), but not where the same outcomes were reported.

Participants: Selected studies were required to have a clinical dementia sample. The definition of 'dementia' was kept purposefully broad to capture the maximum number of studies. Alzheimer's disease (AD), vascular dementia (VD) and frontotemporal dementia (FTD) samples were included. Mild cognitive impairment (MCI) samples were not included, as evidence is mixed as to whether this constitutes a separate clinical entity to dementia or a symptomatic prodromal phase (Petersen et al., 1999; Dubois et al., 2010). Mixed samples where separate data were not provided for dementia alone were also excluded (for example, Barker, Horner, & Bachman, 2010; Novitski, Karantzoulis, & Randolph, 2012). Studies investigating other progressive conditions such as Huntington's disease, Parkinson's disease and multiple sclerosis were excluded.

Sample size: Sample sizes of less than 20 are generally considered 'small' in quantitative studies (Pallant, 2013; p 216). Small sample sizes are particularly problematic when investigating classification accuracy for measures with binary outcomes, as each individual represents a larger proportion of the sample (for example, one false positive out of a sample of 20 will represent a 5% reduction in specificity). As such, a cut-off of 20 participants was set to compromise between maximising study numbers and preventing the inclusion of data which may be skewed due to small samples.

Data collection and extraction

All studies identified were initially screened by title for relevance. Studies referencing performance validity in dementia populations were subject to detailed abstract screening. Full text articles were acquired for studies deemed appropriate, and these were subsequently analysed to ascertain if the above criteria were met.

Quality appraisal

The identified references were screened for quality using the 'QualSyst' critical appraisal tool (Kmet, Lee, & Cook, 2004), which enables the 'systematic, reproducible and quantitative' assessment of studies with disparate methodologies. Like all quality appraisal tools, Qualsyst represents the authors' subjective opinions of the elements that constitute study quality. It was also developed with a relatively small sample of test studies and includes limited investigation of inter-rater reliability (Kmet et al., 2004). However, in the absence of a 'gold standard' measure against which to compare studies (Katrak, Bialocerkowski, Massy-Westropp, Kumar, & Grimmer, 2004), it constitutes a useful means of supplementing the qualitative assessment of data quality.

The QualSyst method employs a 14-item checklist to assess the internal validity of the study design and analysis. Studies are scored depending on the degree to which certain criteria are met: a score of two indicates that the criteria are fully met; one is partially met and zero indicates that the criteria are not met. Summary scores can be calculated for each study by summing the total score gained across the items and dividing by the total possible score. A cut-off point of 0.75 was chosen as a relatively conservative indicator of quality. Some checklist items (Criteria 5, 6 and 7) were excluded as they were not relevant to any of the studies in the shortlist,

primarily because the studies were not interventional and the nature of the population meant that investigator blinding was not possible.

Evaluation of PVT classification accuracy

The majority of studies identified only allowed for the calculation of specificity, as using samples of 'bona fide' dementia patients theoretically means that there should be no 'invalid' performance to identify and thus no sensitivity values to compute. Where studies have identified subsamples of 'suspect effort' patients or used simulator samples, it follows that sensitivity could also be calculated. This analysis focused on specificity, as all studies provided this information and, clinically, this is the more informative value.

To evaluate the measures, specificity values were extracted from the studies or calculated based on the percentage of dementia patients appropriately passing the PVT. This requires an assumption that bona fide patients failing the test represent false positives, a limitation of the approach which is addressed in further detail in the discussion. Specificity data was then pooled across studies, collapsing where necessary across severity and diagnosis within study dementia samples. This method of assessing PVT classification accuracy across studies has been reported in previous meta-analyses (Vickery et al., 2001; Sollman & Berry, 2011; Schroeder, Twumasi-Ankrah, Baade, & Marshall, 2012). Using this method enables all studies to be included in the calculations, but it is recognised that it does not control for between-study heterogeneity and correlations (Schroeder et al., 2012). To account for these limitations, 95% confidence intervals were included. As specificity is a proportion, confidence intervals were calculated using the standard methods for proportions (Altman & Bland, 1994). Pooled specificity was only calculated where there were

two or more studies using the same measure and cut-off, which limited the amount of data available for the analysis.

Results

Overview

The review yielded 24 studies meeting the inclusion criteria. Thirteen studies were retrospective, using data from clinic or research archives. The remaining studies were prospective in that participants were enrolled in the research as part of their routine clinical assessments. No studies published prior to 1997 met the inclusion criteria. Twenty-two reported independent findings and two reported data from the same sample examining different PVTs (Howe, Anderson, Kaufman, Sachs, & Loring, 2007; Howe & Loring, 2009).

Study characteristics

Aims: The majority of studies cited the overall aim of investigating the performance or 'clinical utility' of a particular PVT (or multiple PVTs) in a dementia population. One study (Green et al., 2011) additionally aimed to cross-validate the MSVT and WMT cut-offs in a non-English speaking population.

Measures: Across the 24 studies, 31 PVTs were investigated: these are identified in Table 2. For a brief overview of the format and content of key measures the reader is referred to Table 1 and Appendix 3 in McMillan et al. (2009; pp. 6-7 and 21-24). The majority of measures were established tests of performance validity such as the TOMM (Tombaugh, 1997) whereas others were novel measures derived from existing validated tests (such as the Repeatable Battery for the Assessment of

Neuropsychological Status Effort Index [RBANS EI]; Silverberg, Wertheimer, & Fichtenberg, 2007). A combination of embedded (such as RDS; Greiffenstein et al., 1994) and standalone (such as WMT; Green et al., 1996) measures were utilised. The selected studies represented the majority of measures identified as 'commonly used' in UK clinical neuropsychological practice (McMillan et al., 2009; McCarter et al., 2009).

Samples: All studies included a dementia sample though clinical characteristics varied. The majority used 'mixed dementia types' (n = 19), within which AD was the most common presentation. The remainder (n = 5) specified the dementia subtype (most commonly AD). Comparison groups varied and included different severities of dementia (n = 9), different dementia subtypes (n = 6) and comparisons between dementia and different forms of acquired brain damage, most commonly traumatic brain injury (n = 7). A number of studies utilised age-matched normal controls (n = 7). One study included a large control group taken from the standardisation sample of the WAIS-III (Iverson & Tulsky, 2003) but did not include the clinical characteristics of this sample. Six studies included a comparison group of individuals categorised as 'poor' or 'suspect' effort (a 'known groups' sample), or simulator sample.

Settings and design: Studies were published between 1997 and 2016. The majority were American or Canadian studies, with two from the UK (Singhal, Green, Ashaye, & Gill, 2009; Rudman et al., 2011) and one from Germany (Merten, Bossnick, & Schmand, 2007). Traditionally, there is a North American bias in the PVT literature, as PVTs have been utilised there for longer and are a key element in medicolegal

Table 2: *Summary of studies included in the review with key findings*

Authors (year)	Total dementia sample size	Dementia sample type	Incentive to feign	PVTs assessed	Data included in classification analysis	Compares dementia group with KG/simulators	Compares dementia severities	Compares dementia subtypes	Key findings	Quality appraisal overall score
Dean et al (2009)	214	Mixed dementia types	No	Digit Span ACSS RDS Timed Digit Span Vocabulary Minus Digit Span Dot counting test TOMM T2 Warrington RMT (words) RFIT WMS-III logical memory RMI Finger tapping b-test Rey word recognition equation RAVLT equation Rey-O equation Rey-O/RAVLT equation	✓	✘	✓	✓	Majority of tests demonstrated high FP error rates Most specificities fell in the range of 30%-70% Lower MMSE scores associated with increased test failure Adjusting cut-offs to provide ≥90% specificity rendered several measures inappropriate for use in dementia	0.86
Walter et al (2014)	31	Mixed dementia types of moderate-severe severity	No	TOMM T2	✓	✘	✓	✘	Approximately 20% of moderate to severe dementia group failed TOMM T2 No significant difference between TOMM T2 scores between control and MCI groups	1.00

Authors (year)	Total dementia sample size	Dementia sample type	Incentive to feign	PVTs assessed	Data included in classification analysis	Compares dementia group with KG/simulators	Compares dementia severities	Compares dementia subtypes	Key findings	Quality appraisal overall score
Bortnick et al (2013)	128	Mixed dementia types	Unlikely	RBANS EI TMT RFIT (TOMM used as classification variable for effort groups)	✓	✓	✓	✓	Majority of measures demonstrated unacceptably high FP rates in patients classified as having 'adequate' motivation Specificities ranged from of 0% for RFIT combination to 99% for TMT ratio. Sensitivity values ranged from 0% for TMT ratio to 100% for RFIT (recall and combination) Patients with milder forms of dementia (MMSE >25) at higher risk of misclassification	0.86
Teichner & Wagner (2004)	21	Mixed dementia types	Not reported	TOMM T2 and retention	✓	✗	✗	✗	High misclassification rates for dementia sample - specificity 24% on TOMM T2 Specificity 100% for cognitively intact sample and 92.7% for cognitively impaired group	0.95
Tombaugh (1997)	37	Mixed dementia types	No	TOMM T2 and retention	✓	✗	✗	✗	Classification accuracy of dementia group using TOMM T2 significantly lower than other groups: 92% versus >97%	0.95
Merten et al. (2007)	20	AD	Not reported	ASTM WMT TOMM T2 and retention	✓	✗	✗	✗	Only 10% of Alzheimer's patients passed ASTM and WMT. 70% passed TOMM T2 All control participants passed all tests	0.86

Authors (year)	Total dementia sample size	Dementia sample type	Incentive to feign	PVTs assessed	Data included in classification analysis	Compares dementia group with KG/simulators	Compares dementia severities	Compares dementia subtypes	Key findings	Quality appraisal overall score
Rudman et al. (2011)	42	Mixed dementia types	Not reported	MSVT NV-MSVT TOMM T2 and retention Dot counting test RFIT Coin in the hand	✓	✘	✓	✘	Mild dementia group performed significantly better on all six measures that the moderate / severe group RFIT and TOMM were the most sensitive to severity of cognitive impairment Dot counting test (time) was the only measure that achieved 100% specificity No significant correlations between emotional functioning and measures of performance validity	0.95
Burton et al. (2015)	145	Mixed dementia types	No	RBANS EI RBANS ES	✓	✘	✓	✓	RBANS EI: 48% of total sample scored below cut-off. Increased severity of dementia associated with increased likelihood of scoring below cut-off RBANS ES: 14% of total sample scored below cut-off. ES not highly associated with dementia severity however, non-AD subsample had increased likelihood of scoring below cut-off	0.86
Dunham et al. (2014)	46	Mixed dementia types (including *n* = 1 with MCI)	No	RBANS EI RBANS ES (MSVT used as grouping variable)	✓	✓	✘	✘	RBANS ES demonstrated high specificity in dementia sample (81%), EI had low specificity (41%) Classification rates of ES relatively stable across different severities of cognitive impairment, EI rates declined as cognitive functioning decreased Comparable rates of sensitivity found for simulation sample	0.91

Authors (year)	Total dementia sample size	Dementia sample type	Incentive to feign	PVTs assessed	Data included in classification analysis	Compares dementia group with KG/simulators	Compares dementia severities	Compares dementia subtypes	Key findings	Quality appraisal overall score
Duff et al. (2011)	126	AD	Not reported	RBANS EI	✓	✗	✗	✗	Base rates of failure on EI ranged from 3% in cognitively intact OA to 33% in AD Years of education was related to EI performance in nursing home residents, MCI patients and AD patient samples	0.91
Kiewel et al. (2012)	142	AD	No	Digit span ACSS RDS LDSF (1 trial) LDSF (2 trials) Vocab-digit span	✓	✗	✓	✗	RDS demonstrated unacceptably high FP rates in moderate and severe AD (76% and 17% respectively) Classification accuracy overall decreased with increasing AD severity on all measures with exception of Vocab-Digit span (which was not tested in the severe group)	1.00
Greve et al. (2006)	22	Diagnosed memory disorder – suspected AD, VD or both	No	TOMM T1, T2 and retention	✓	✓	✗	✗	Memory disorder patients performed more poorly than TBI patients (specificity T2 is 82%) All three trials detected known malingerers with TBI with a low FP error rate of approximately 5%	0.86
Iverson & Tulsky (2003)	38	AD	Not reported	Digit span ACSS, LDSF, LDSB, Vocabulary-Digit Span ACSS	✓	✗	✗	✗	Digit span ACSS demonstrated specificity of 95% in AD Vocabulary-Digit Span ACSS demonstrated specificity of 97%	0.91

Authors (year)	Total dementia sample size	Dementia sample type	Incentive to feign	PVTs assessed	Data included in classification analysis	Compares dementia group with KG/simulators	Compares dementia severities	Compares dementia subtypes	Key findings	Quality appraisal overall score
Howe et al. (2007)	31	'Early' dementia (*n* = 13) 'Advanced' dementia (*n* = 18)	No	MSVT MSVT dementia profile algorithm	✗	✗	✓	✗	Using established symptom validity indices (IR, DR, CNS) specificity for early dementia was approximately 61% and for advanced dementia 17% With application of the dementia profile algorithm, specificity for early dementia increased to 92% and advanced dementia to 89%	0.95
Green et al. (2011)	65	Mixed dementia types	Not reported	WMT MSVT	✗	✗	✗	✗	Specificity of WMT and MSVT was 98.4% or higher in dementia patients with use of dementia profile analysis	0.95
Howe & Loring (2009)	31*	'Mild-Moderate dementia' (*n* = 13) 'Severe dementia' (*n* = 18)	No	MSVT	✗	✗	✓	✗	With use of dementia profile, FP rate was 5.8% (36/52 patients in total sample correctly classified)	0.95
Boone et al. (2002)	37	'Mild' dementia (*n* = 16) 'Moderate' dementia (*n* = 21)	Not reported	Dot Counting Test	✓	✓	✓	✗	Specificity in mild dementia = 75% and in moderate dementia = 33.4% Fairly robust in the context of mild dementia with relatively few incorrect identifications – provides advantages over memory-based PVTs	0.95

Authors (year)	Total dementia sample size	Dementia sample type	Incentive to feign	PVTs assessed	Data included in classification analysis	Compares dementia group with KG/simulators	Compares dementia severities	Compares dementia subtypes	Key findings	Quality appraisal overall score
Arnold et al. (2005)	31	AD (*n* = 12) Non AD dementia (*n* = 18)	Not reported	Finger tapping test	✗	✓	✗	✓	Men tapped faster than women across all groups, therefore groups divided by gender Dominant hand scores proved more sensitive to non-credible performance Across all groups, cut-off scores yielding highest sensitivity and specificity values were dominant hand ≤ 28 for women and ≤ 35 for men Useful in discriminating performance validity in AD but less effective in non-AD dementias due to motor component	0.91
Schroeder at al. (2012)	45	Mixed dementia	No	Coin in the hand test	✓	✗	✗	✓	11% of patients made 2 or more errors (89% specificity at this cut-off) No patient made > 4 errors 73% obtained perfect scores	0.86
Henry et al. (2010)	21	Mixed dementia	No	NV-MSVT	✗	✗	✗	✗	13/21 dementia patients failed 'A' criteria With addition of 'B' criteria none of the dementia patients were classified as having invalid performance	0.95
Loring et al. (2007)	50	Mixed dementia types	No	VSVT	✗	✗	✗	✗	At a cut off of <21 hard items, specificity was 62%	0.86
Schroeder & Marshall (2010)	22	Not specified	Not reported	Sentence repetition test	✗	✗	✗	✗	Specificity 77% for cut-off score of 9 and 64% for cut-off score of 10	0.86

Authors (year)	Total dementia sample size	Dementia sample type	Incentive to feign	PVTs assessed	Data included in classification analysis	Compares dementia group with KG/simulators	Compares dementia severities	Compares dementia subtypes	Key findings	Quality appraisal overall score
Loring et al. (2016)	178	'Early' AD	Not reported	RDS AVLT logistic regression AVLT recognition	✓	✘	✓	✘	RDS specificity at ≤7 was 66% and at ≤6 was 87% for 'early' AD AVLT indices yielded unacceptably high false-positive rates at a range of cut-offs Combining embedded PVT indicators lowered the false-positive rates	0.86
Zenisek et al. (2016)	183	Mixed dementia types	No	RDS	✓	✘	✘	✓	RDS specificity at ≤6 across dementia subtypes ranged from 73% - 83%. A criterion of ≤7 resulted in unacceptably low specificity. Those scoring below cut-offs performed worse on cognitive measures than those scoring above cut-offs.	1.00

Notes: KG = Known Groups; FP = false positive; AD = Alzheimer's disease; VD = Vascular Dementia; DLB = Dementia with Lewy Bodies; FTD = Fronttemporal Dementia; MCI = Mild Cognitive Impairment; OA = Older Adults; TBI = Traumatic Brain Injury; LD = Learning Disability; CVA = Cerebrovascular Accident; NOS = Not Otherwise Specified; MS = Multiple Sclerosis

Measures: Test of Memory Malingering (TOMM; Tombaugh, 1996); Digit Span Age Corrected Scaled Score (Digit Span ACSS; Babikian, Boone, Lu & Arnold, 2006); Reliable Digit Span (RDS; Greiffenstein et al., 1994); Timed Digit Span (Babikian et al., 2006); Vocabulary Minus Digit Span (Iverson & Tulsky, 2003); Dot Counting Test E-Score (Boone, Lu & Herzberg, 2002); Dot Counting Test (DCT; Lezak, 1995); Warrington Recognition Memory Test (Warrington RMT; Iverson & Franzen, 1994); Rey 15-item Test (RFIT; Free Recall Lezak, 1983, p619; Recognition Equation, Boone, Salazar, Lu, Warner-Chacon & Razani, 2002); Weschler Memory Scale-III Logical Memory Rarely Missed Index (WMS-III RMI; Killgore & DellaPietra, 2000); Finger Tapping dominant hand (Arnold et al., 2005); b-Test E-score (Boone et al., 2002); Medical Symptom Validity Test (MSVT; Green, 2004); Non-Verbal Medical Symptom Validity Tests (NV-MSVT; Green, 2008); Victoria Symptom Validity Test (VSVT; Slick, Hopp & Strauss, 1997); Rey Auditory Verbal Learning Test Effort Equation (RAVLT; Boone, Lu & Wen, 2005); Rey-Osterreith Effort Equation (Rey-O; Lu, Boone, Cozolino & Mitchell, 2003); Rey-Osterreith /RAVLT discriminant function (Rey-O/RAVLT; Sherman, Boone, Lu & Razani, 2002); Rey word recognition equation (Nitch, Boone, Wen, Arnold & Alfano, 2006); Longest Digit Span Forwards (LDSF; Babikian et al., 2006); Longest Digit Span Backwards (LDSF; Iverson & Tulsky, 2003); Repeatable Battery for the Assessment of Neuropsychological Status Effort Index (RBANS EI; Silverberg, Wertheimer & Fichtenberg, 2007); Repeatable Battery for the Assessment of Neuropsychological Status Effort Scale (RBANS ES; Novitski, Steele, Karantzoulis & Randolph, 2012); The Amsterdam Short Term Memory Test (ASTM; Schagen, Schmand, de Sterke & Lindeboom 1997); The Word Memory Test (WMT; Green, Allen & Astner, 1996); Trail Making Test (TMT; Iverson, Lange, Green & Franzen, 2002); Sentence Repetition Test (SRT; Strauss, Sherman & Spreen, 2006); The Coin in the Hand Test (Kapur, 1994); Auditory Verbal Learning Test (AVLT) logistic regression (Davis. Millis, & Axelrod, 2012); AVLT recognition (Binder, Villanueva, Howieson, & Moore, 1993)

*Same sample as used in Howe et al. (2007)

assessments (Suesse et al., 2015). Most of the data was derived from clinical practice and all studies were cross-sectional and between groups in design.

Appraisal of identified studies: An overview of the quality appraisal for each study is shown in Table 3. Scores ranged between 0.86 and 1.0 and no study was excluded on the basis of the appraisal. Overall, studies scored highly on sufficiently describing the study aims, use of appropriate measures and reporting of results and conclusions. The majority of studies considered potential confounding variables such as age, years of education, estimated premorbid intellect and mood. Common areas of weakness included the method of subject or comparison group selection (such as not explicitly stating whether participants had a known incentive to feign or not using clinical classification tools for the diagnosis of dementia); the description of subject characteristics (for example not specifying use of validated measures to assess dementia severity); use of an appropriate sample size (or vastly differing sample sizes between experimental and comparison groups) and the explicit description of analytic methods.

Description of findings

Classification accuracy of individual measures: It was possible to examine pooled specificity across 11 PVTs. Of these, six were purpose-designed 'standalone' measures: TOMM (Tombaugh, 1996), Rey Fifteen-Item Test (RFIT) free recall (Lezak, 1983, p619), RFIT combination equation (Boone, Salazar, Lu, Warner-Chacon, & Razani, 2002), Coin in the Hand (Kapur, 1994), Finger Tapping (Arnold et al., 2005) and the Dot Counting Test (DCT; Lezak, 1995). The remaining measures were 'embedded' PVTs: The Repeatable Battery for the Assessment of Neuropsychological Status Effort Index (RBANS EI; Silverberg et al., 2007), Digit

Table 3: *Results of quality appraisal*

	Dean et al. (2009)	Walter et al. (2014)	Bortnick et al. (2013)	Teichner & Wagner (2004)	Tombaugh (1997)	Merten et al. (2007)	Rudman et al. (2011)	Burton et al. (2015)	Dunham et al. (2014)	Duff et al. (2011)	Kiewel et al. (2012)	Greve et al. (2006)	Iverson & Tulsky (2003)	Howe et al. (2007)	Green et al. (2011)	Howe & Loring (2009)	Boone et al. (2002)	Arnold et al. (2005)	Schroeder at al. (2012)	Henry et al. (2010)	Loring et al. (2007)	Schroeder & Marshall (2010)	Loring et al. (2016)	Zenisek et al. (2016)
1. Question / objective sufficiently described?	2	2	2	2	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
2. Study design evident and appropriate?	2	2	2	2	2	2	2	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
3. Method of subject / comparison group selection *or* source information/input variables described and appropriate?	2	2	1	1	2	2	2	2	1	1	2	1	1	2	1	2	1	1	2	2	1	1	1	2
4. Subject (and comparison group, if applicable) characteristics sufficiently described?	2	2	2	2	2	1	2	2	2	2	2	1	2	2	2	2	2	1	2	2	1	1	1	2
8. Outcome and (if applicable) exposure measure(s) well defined and robust to measurement / misclassification bias? Means of assessment reported?	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
9. Sample size appropriate?	1	2	1	2	2	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
10. Analytic methods described / justified and appropriate?	1	2	2	2	2	1	2	1	2	2	2	2	1	1	2	2	2	2	1	2	2	2	2	2
11. Some estimate of variance reported for the main results?	2	2	1	2	2	2	2	2	2	1	2	2	2	2	2	2	2	2	1	2	2	2	2	2
12. Controlled for confounding?	1	2	2	2	2	2	2	1	1	2	2	2	2	2	2	1	2	2	2	1	1	1	1	2
13. Results reported in sufficient detail?	2	2	2	2	2	2	2	2	2	2	2	1	2	2	2	2	2	2	1	2	2	2	2	2
14. Conclusions supported by the results?	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
	0.86	1.00	0.86	0.95	0.95	0.86	0.95	0.86	0.91	0.91	1.00	0.86	0.91	0.95	0.95	0.95	0.95	0.91	0.86	0.95	0.86	0.86	0.86	1.00

Notes: Criteria 5, 6, and 7 from original QualSyst tool (Kmet et al., 2004) excluded as not relevant to current review

Span Age-Corrected Scaled Score (ACSS; Babikian et al., 2006), RDS (Grieffenstein et al., 1994), Vocabulary Minus Digit Span (Iverson & Tulsky, 2003) and Trail Making Test Ratio (TMT; Iverson, Lange, Green, & Franzen, 2002). The results of this analysis are shown in Table 4 and Figure 2. The Repeatable Battery for the Assessment of Neuropsychological Status Effort Scale (RBANS ES; Novitski et al., 2012), Medical Symptom Validity Test (MSVT; Green, 2004), Non-verbal MSVT (NV-MSVT; Green, 2008) and WMT (Green et al., 1996) were also assessed, however, as they were found to utilise specific algorithms to assess patients with dementia or suspected dementia (unlike other PVTs where traditional cut-offs are applied regardless of diagnosis), they are described separately below.

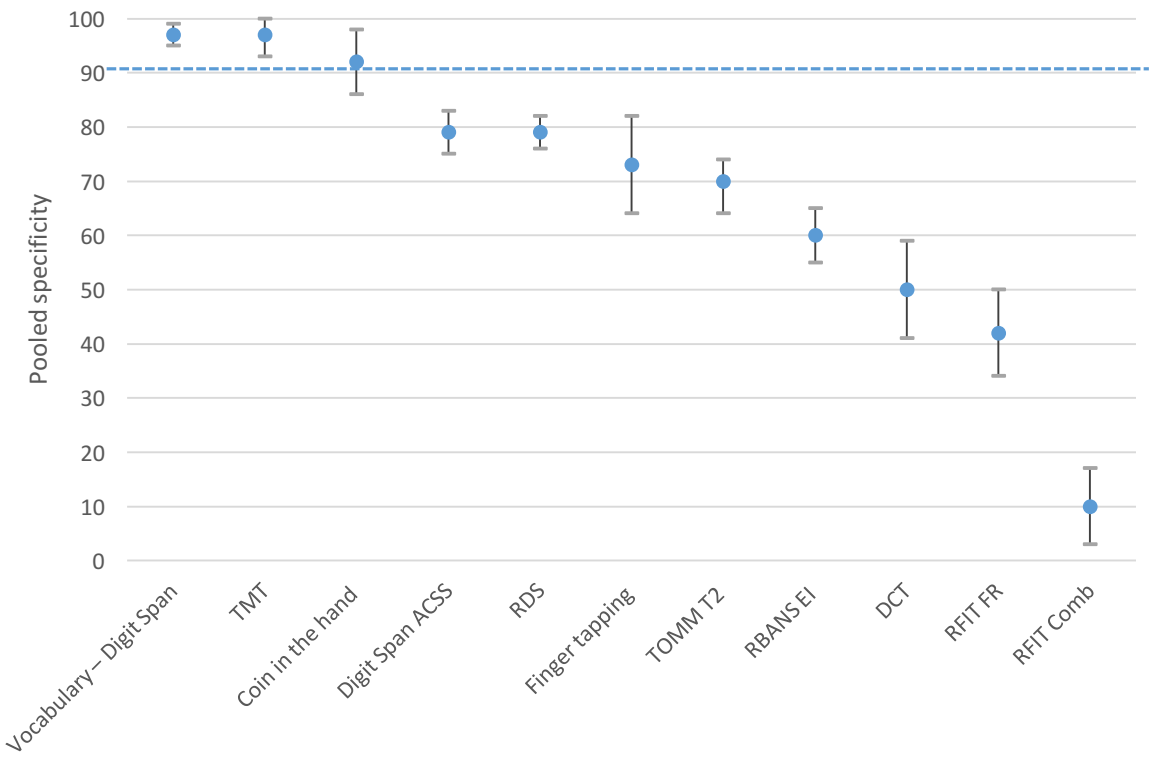
Only three of the measures identified, Vocabulary Minus Digit Span, the TMT Ratio and Coin in the Hand, met or exceeded the 90% threshold for appropriate specificity in a dementia population. All studies investigating Vocabulary Minus Digit Span demonstrated the same specificity of 97% (Iverson & Tulsky, 2003; Dean et al., 2009; Kiewel et al., 2012). The latter two studies accounted for 88% of the total pooled sample, however, between them included only seven patients with 'severe' dementia. Kiewel et al. (2012) highlight that the WAIS Vocabulary measure is rarely administered in more severe cases so it is possible that specificity rates were inflated as the full spectrum of dementia severity was not represented. The two studies examining the TMT Ratio similarly found consistently high specificity (95% in Merten et al., 2007; 99% in Bortnick et al., 2013). The Bortnick et al. (2013) paper, however, also reported sensitivity values of 0% for this measure, which, if replicated, would raise questions about the utility of this measure as a PVT as it is unable to identify invalid performance. The Coin in the Hand was explored across two studies again demonstrating consistently high specificity (88% in Rudman et al.,

Table 4: *Results of classification accuracy analysis*

Measure	**Cut-off**	**Total pooled *N***	**Specificity (%)**	**95% CI**
Vocabulary – Digit Span (*k* = 3)	> 5	304	97	95 – 99
TMT (*k* = 2)	Ratio of A:B completion time <1.5	62	97	93 - > 100
Coin in the hand (*k* = 2)	≤ 7	87	92	86 – 98
Digit Span ACSS (*k* = 3)	≤ 5	352	79	75 - 83
RDS (*k* = 4)	≤ 6	586	79	76 - 82
Finger tapping (*k* = 2)	Men ≤ 35 Women ≤ 28	86	73	64 - 82
TOMM T2 (*k* = 8)	< 45	315	70	64 – 74
RBANS EI (*k* = 4)	> 3	445	60	55 - 65
DCT (*k* = 2)	≥ 17 combination score	117	50	41 - 59
RFIT (*k* = 2)	Free recall < 9	130	42	34 - 50
RFIT (*k* = 2)	Combination equation < 20	73	10	3 - 17

Notes: *k* = number of studies; TMT = Trail Making Test; Digit Span ACSS = Digit Span Age Corrected Scaled Score; RDS = Reliable Digit Span; TOMM T2 = Test of Memory Malingering Trial 2; RBANS EI = Repeatable Battery of Neuropsychological Status Effort Index; DCT = Dot Counting Test; RFIT = Rey Fifteen Item Test

Figure 2: *Classification accuracy of individual measures with confidence intervals*



Notes: TMT = Trail Making Test; Digit Span ACSS – Digit Span Age Corrected Scaled Score; RDS = Reliable Digit Span; TOMM T2 = Test of Memory Malingering Trial 2; RBANS EI = Repeatable Battery of Neuropsychological Status Effort Index; DCT = Dot Counting Test; RFIT FR = Rey Fifteen Item Free Recall; RFIT Comb = Rey Fifteen Item Combination Equation.

2011, to 96% in Schroeder et al., 2012), however, the small pooled sample size (n = 87) means that these results should be interpreted with caution.

All other PVTs showed pooled specificity rates below 90%. The TOMM, which is the most widely studied PVT in this population (and the most widely used in UK clinical neuropsychological practice; McCarter et al., 2009), obtained an overall specificity on Trial 2 of 69% across eight studies. Values across studies ranged from 24% (Teichner & Wagner, 2004) to 79% (Walter et al., 2014) suggesting significant variability in findings but no indication that this measure could be used with confidence in this population. Similar inconsistency was noted in the RBANS EI, for which overall specificity was 60% with values ranging from 41% (Dunham, Shadi, Sofko, Denney, & Calloway, 2014) to 70% (Bortnick et al., 2013). This measure had a large overall pooled sample size (n = 445) across four studies. Given that one of the stated principles of the RBANS is to be a 'standalone core battery for the detection and characterisation of dementia in the elderly' (Randolph, 1998), it is of importance that the PVT embedded within the battery is one which is able to detect valid performance in a dementia population. The results of this analysis, however, suggest that use of this measure with conventional cut-offs could lead to high rates of false positive classification.

Data from two embedded measures calculated from WAIS Digit Span indices found comparable pooled specificity values. An overall classification accuracy of 79% was found for the Digit Span ACSS, with values ranging from 73% (Dean et al., 2009) to 95% (Iverson & Tulsky, 2003). Although again failing to reach the 90% mark, this measure does appear more robust to misclassification that the RDS, which achieved an overall value of 79% with a range from 70% (Dean et al., 2009) to 87%

(Loring et al., 2016). This may be related to the fact that Digit Span ACSS is adjusted for age, whereas RDS uses raw scores (Dean et al., 2009).

The RFIT had two indices that it was possible to assess for classification accuracy: Free Recall and Recognition. Small overall sample size for the latter (n = 73) means that the results should be interpreted with caution. Free Recall had a pooled specificity of 42% with a range from 28% (Bortnick et al., 2013) to 45% (Dean et al., 2009). The Recognition equation only achieved 10%, with a low of 0% (Bortnick et al., 2013) and a high of 14% (Dean et al., 2009). After the TOMM, the RFIT is the most commonly used PVT in clinical practice in the UK (McCarter et al., 2009): these results strongly imply that, if traditional cut-offs are used, this an inappropriate tool for use in populations with a potential diagnosis of dementia due to high risk of misclassification

Two studies investigated the DCT using the combination score criterion suggested by Boone, Lu & Herzberg (2002). Overall, pooled specificity was 50%, with no study demonstrating specificity over 51% (Dean et al., 2009; Boone et al., 2002). The Finger Tapping measure showed 73% specificity overall, although some studies demonstrated values approaching the 90% threshold (87% in Arnold et al., 2005). Again, small sample sizes for the latter measure are problematic.

There were a number of measures which were not included in the above analysis as results from only a single study were available (see Table 2 for identified studies). Of these, only Four Digits Timed (Dean et al., 2009) demonstrated a specificity above 90%, however, a small sample size suggests that additional data is needed to support these specificity values.

Key findings: The data indicate that the majority of measures do not provide adequate classification accuracy when conventional cut-off scores are used in dementia populations. This includes both standalone and embedded measures, and those tapping a range of cognitive domains including recognition memory, processing speed and motor speed. Three measures produced specificity above the 90% threshold (Vocabulary Minus Digit Span; TMT Ratio; Coin in the Hand) however, small sample sizes, limited inclusion of more severely impaired patients and issues with sensitivity suggest that these results be interpreted with caution.

Classification accuracy of PVTs adopting a 'dementia profile'

Four PVTs – the RBANS ES, WMT, MSVT and NV-MSVT - were not included in the classification accuracy analysis as they purport to offer specific criteria to allow the examiner to differentiate invalid performance due to 'effort' and that due to genuine memory impairment in the context of a potential differential diagnosis of dementia. Two studies examining the RBANS ES demonstrated markedly improved specificity when compared to the traditional RBANS EI embedded measure. Values ranged between 81% (Dunham et al., 2014) and 96% (Burton, Enright, O'Connell, Lanting, & Morgan, 2015) using a cut-off of less than 12. These data suggest that this measure holds some promise in this population, particularly as sample sizes were also adequate (n = 191). Some methodological issues were noted, however, such as failure to consider confounding factors such as age, education and IQ in both studies and, in the case of Burton et al. (2015), poor specification of analytic methods.

Similar high specificities between 89% and 100% were reported for the WMT, MSVT and NV-MSVT measures with the application of the dementia profile

algorithms, even in the case of advanced dementia (Howe et al., 2007; Howe & Loring, 2009; Henry et al., 2010; Green, Montijo, & Brockhaus, 2011). When the dementia profile algorithm is not utilised, specificity on the MSVT and NV-MSVT has been shown to be very poor: between 45% and 62% on the MSVT (Rudman et al., 2011; Howe et al., 2007) and 33% on the NV-MSVT (Rudman et al., 2011). WMT pass rates without use of the dementia profile were less than or equal to 10% for bona fide AD patients (Merten et al., 2007).

Key findings: The RBANS ES, WMT, MSVT and NV-MSVT dementia profile algorithms demonstrate consistently higher specificity than standard PVTs, with values approaching or exceeding 90% across the range of dementia severity. The use of these measures with conventional cut-offs in dementia populations is, however, not advised.

Adjusted cut-off scores and impact on classification accuracy

As highlighted by Walter et al. (2014), rigid application of conventional PVT cut-offs is likely inappropriate in a dementia population and, as evidenced by the analysis above, may lead to excessive false positive results. Given this, does adjusting the cut-offs for individuals with confirmed or suspected dementia leads to an improvement in specificity? A number of studies identified in this review addressed this issue.

For the TOMM, Teichner and Wagner (2004) were unable to reach the 90% specificity threshold even when the cut-off was reduced from the conventional cut-off of less than 45 to less than 40 (86% specificity). By contrast, Bortnick et al (2013) found that a threshold of less than 37 for Trail 2 produced a specificity of 90% whilst maintaining moderate sensitivity at 78%. Specificity was further

improved to 95% by lowering the cut-off to less than 28 (Dean et al., 2009). Investigating the RFIT, Bortnick et al. (2013) found that specificity above 90% could be achieved by adjusting the Free Recall cut-off to less than two and Recognition trial to less than three. The latter was corroborated by Dean et al. (2009), however, for the Free Recall these authors were unable to achieve a specificity above 90% even when the threshold was lowered to less than one.

The RBANS EI was found to reach the 90% specificity threshold at a cut-off of greater than seven (Bortnick et al., 2013). For the RBANS ES Burton et al. (2015) adjusted the cut-off to less than seven (as per an earlier study by Schroeder et al., 2012) and demonstrated improved specificity in both AD and non-AD groups, however, only specificity for AD patients reached the 90% threshold.

The two studies investigating DCT specificity (Boone et al., 2002a; Dean et al., 2009) already utilised an alternative criterion suggested by Boone et al. (2002b), as opposed to the traditional measure originally put forward by Lezak (1995). Dean et al. (2009) highlighted that adjusting the original Boone et al. (2002b) criterion from equal to or greater than 17 to greater than 42 enabled specificity to reach the 90% threshold. Rudman et al. (2011), however, identified the original Lezak (1995) index (which looks at the relative discrepancy between grouped and ungrouped dots reaction time) as more appropriate in dementia as it takes into account the potential reduction in processing speed in this population. Using this criterion, they achieved 100% specificity.

For Digit Span derived measures, Dean et al. (2009) found that the Digit Span ACSS needed to be reduced to less than three (from less than six) and the RDS to less than four (from less than seven) to achieve at least 90% specificity in their

dementia sample. Similarly, in an early AD sample Loring et al. (2016) had to reduce the RDS cut-off to less than or equal to five to reach a specificity of 97%. Whilst this would reduce the likelihood of false positive identification, the application of more conservative cut-offs impacts upon the sensitivity of these measures and thus severely limits their ability to detect invalid performance when it occurs (Kiewel et al., 2012).

Key findings: Adjusting PVT cut-offs has been shown to improve the specificity of some measures when applied in dementia populations. It is unclear, however, the extent to which the sensitivity, and thus utility, of these measures is sacrificed in the process of doing so.

Findings of known groups and simulator studies

To address the issue of sensitivity, five of the identified studies used a 'known groups' design, whereby clinical samples are compared to clinical groups with suspected invalid performance (usually determined by clinical consensus following consideration of a range of factors such as implausible patterns of response, or deficits disproportionate to functional impairment; Bortnick et al., 2013) or to 'simulators' (non-clinical samples coached to behave like individuals with invalid performance). The latter method enables researchers to address the problem of unequal sample sizes, which is a challenge when drawing 'suspect effort' participants from larger samples of interest. Both designs allow for the estimation of sensitivity as it is possible to ascertain how well the measure picks up on the individuals 'known' to have invalid performance.

A number of studies highlighted the 'trade off' between specificity and sensitivity in the RBANS EI, the Finger Tapping Test; RFIT Free Recall and

Combination; TMT Ratio and TOMM T2 (Dunham et al., 2014; Arnold et al., 2005; Bortnick et al; 2013; Greve et al., 2006), that is, with increased specificity, sensitivity to detect invalid performance dramatically reduced in these measures, for some measures down to 0% (TMT Ratio; Bortnick et al., 2013). The DCT was shown to balance sensitivity and specificity at 75%, but only in the case of mild dementia as specificity in moderate dementia at traditional cut-offs dropped to 33% (Boone et al., 2002). The RBANS ES appeared to be able to maintain moderate levels of both sensitivity and specificity (89% and 81%; Dunham et al., 2014), however, this measure is designed to account for the types of cognitive impairment frequently seen in dementia (that is, the decline of free recall before recognition in amnestic disorders), and would therefore likely produce higher false positive rates in a normal or non-dementia clinical population (Novitski et al., 2012). In addition, use of a coached simulator design in the Dunham et al. (2014) study is methodologically problematic, as it is unclear how these results would be reliably applied to clinical practice (Greve & Bianchini, 2004).

Key findings: Balancing sensitivity and specificity of PVTs in dementia populations is a significant challenge, and calls into question the clinical utility of these measures if they are unlikely to detect invalid performance when it is present. Using measures with dementia profile algorithms may mitigate this issue, however, further research using clinical populations is required.

Discrepancies in classification accuracy between different severities of dementia

Ten studies (Boone et al., 2002; Howe et al., 2007; Howe & Loring, 2009; Dean et al., 2009; Rudman et al., 2011; Kiewel et al., 2012; Bortnick et al., 2013; Walter et al., 2014; Burton et al., 2015; Loring et al., 2016) compared the

performance of a range of measures across different severities of dementia. The studies utilised different methods of categorising patients as mild, moderate and severe, for example Dean and colleagues (2009) use Mini-Mental State Examination (MMSE; Folstein, Folstein, & McHugh, 1975) scores of 21-30 for 'mild', 15-20 for 'moderate' and less than 15 for 'severe'. By contrast, Walter et al. (2014) utilised RBANS standard scores: 'Moderate to Severe' dementia was a score between 56 and 77, and these were compared to MCI patients (scores between 77 and 85) and cognitively intact controls (85 to 112).

In a large study covering 12 PVTs, Dean et al. (2009) noted that their 'mild' sample failed an average of 36% of PVTs, with proportions increasing to 47% for the 'moderate' sample and 83% for 'severe'. Worst performing measures in this study were RFIT, TOMM, b-Test, RAVLT equation, Rey-Osterrieth equation and Rey-Osterreith/RAVLT equation, which classified none of the more severely impaired patients. Other studies similarly found statistically significant differences in classification accuracy between milder and more severe groups. This was the case for the TOMM T2 (Walter et al., 2014; Rudman et al., 2011); RFIT (Rudman et al., 2011), Coin in the Hand (Rudman et al., 2011); DCT (Boone et al., 2002; Rudman et al., 2011), MSVT (Rudman et al., 2011) and NV-MSVT (Rudman et al., 2011). RBANS EI scores were similarly found to be correlated with dementia severity, where higher EI scores (those suggestive of invalid performance) were associated with increased dementia severity (Burton et al., 2015). The same study found no significant correlation with dementia severity for the RBANS ES, however, of note is that the authors used an adjusted cut-off score to assess the association between RBANS ES failure rates and dementia severity but used the conventional cut-off for

the RBANS EI. As such, it is unclear if applying a more stringent criterion for the RBANS EI would have impacted on the relationship with severity.

In a study of five WAIS-III Digit Span indices, Kiewel and colleagues (2012) found that RDS, Digit Span ACSS, longest digits forward 1 (LDF1) and longest digits forward 2 (LDF2) all produced unacceptable false positive error rates with increasing dementia severity. The ACSS was found to be useful in 'mild' dementia (95% specificity), as was Vocabulary Minus Digit Span (also 95% specificity), but only the latter held for 'moderate' severity patients (100% specificity versus 88% for the Digit Span ACSS). The Vocabulary Minus Digit Span index performed similarly in the Dean et al. (2009) study, being the only one of the 12 PVTs investigated which reached the 90% specificity threshold across all dementia severities. Both authors concede, however, that subsamples for patients with 'severe' impairments were small ($N = 7$ in Dean et al., 2009) or non-existent (Kiewel et al., 2012) as the Vocabulary subtest is rarely administered in cases of more severe impairment. The Loring et al. (2016) study further demonstrated that using an RDS cut-off score of less than or equal to six only produced adequate specificity (over 90%) in individuals with MMSE scores of 23 or more, but their 'early AD' sample did not include individuals with more severe impairment. Overall, it thus remains unclear if Digit Span indices are appropriate for use as PVTs in this population.

Bortnick et al. (2013) corroborated the Rudman et al. (2011) findings, demonstrating 0% specificity on the RFIT Combination Equation across all dementia severities. This study also found that three of the four tests explored, RBANS EI, TMT Ratio and RFIT Recall, had best specificity when patients had MMSE scores of between 21 and 25 but not at scores over 25. This raises an additional query as to

whether individuals with milder forms of dementia are at increased risk of being misclassified by these measures.

Key findings: Most measures examined demonstrated diminishing classification accuracy with increasing dementia severity, however, the potential for misclassification is likely to be problematic across the spectrum of the condition. The RBANS ES (dementia specific algorithm) was shown to be robust to severity of dementia, however, it is unclear if this is a product of the authors' use of an adjusted cut-off score.

Discrepancies in classification accuracy across dementia subtypes

Five studies investigated differences in PVT classification accuracy across different diagnoses of dementia (Arnold et al., 2005; Dean et al., 2009; Schroeder et al., 2012; Bortnick et al., 2013; Burton et al., 2015; Zenisek et al., 2016). Schroeder et al. (2012) was the only study to look at the Coin in the Hand Test, and found that the number of test errors were not related to dementia subtype (dementia not otherwise specified [NOS], AD, VD, FTD and alcohol-related dementia). Conversely, in a study comparing 'AD' to 'non-AD' dementia, Burton et al. (2015) found significant group differences in RBANS EI performance, with failure rates higher for individuals with non-AD dementia. Failure rates were comparable between the groups on the RBANS ES. The RBANS EI findings were not corroborated by Bortnick et al. (2013), who found comparable specificity rates on this measure between AD, VD and mixed dementia diagnoses, with marginally higher rates in individuals with dementia NOS. The same authors highlighted that the RFIT Free Recall and Recognition had inadequate specificity across the diagnoses, whereas the TMT Ratio met or exceeded the 90% specificity threshold across all

groups. It should be borne in mind that small sample sizes were an issue for this study.

Differences in specificity were also found when comparing AD, VD and FTD (Dean et al., 2009). Worst performing measures across all dementia subtypes were RFIT (Free Recall and Recognition); RAVLT equation, Rey-O equation and REY-O/RAVLT. Only Vocabulary Minus Digit span was found to afford specificity above 90% in all three patient groups. Three Digits Timed and Four Digits Timed had 100% specificity in AD and VD, but both had poor performance in FTD (33% and 67% respectively). Conversely, VD patients achieved only 43% specificity with traditional cut-offs on the Finger Tapping test, whereas both AD and FTD patient groups had 100% specificity on this measure. Again, many of the subgroups were small and the authors further note that there was a lack of demographic equivalence between them, for example FTD patients having significantly higher MMSE scores than AD patients. However, the Finger Tapping data replicates the results of an earlier study by Arnold et al. (2005), which highlights the utility of this measure in AD but not non-AD populations.

The RDS was explored in four dementia subgroups by Zenisek et al. (2016): AD, VD, Dementia with Lewy bodies (DLB) and FTD. At the conventional cut-off of less than or equal to six, the RDS did not achieve specificity over 90% in any of the groups however, differences between diagnoses were noted with values ranging from 73% in FTD to 85% in DLB. Again, sample sizes for all but the AD group were small.

Key findings: Overall, the data point to the potential influence of pathology on PVT performance. The results should, however, be viewed as preliminary due to small sample sizes.

Discussion

In summary, classification accuracy analysis of 11 PVTs demonstrated that in dementia patients without clear external incentives to underperform, most measures had inadequate specificity (under the 90% threshold). This included both standalone and embedded PVTs which are commonly used in UK clinical neuropsychological practice, such as the TOMM, RFIT, DCT, RBANS EI, Digit Span ACSS, RDS, WMT, MSVT and NV-MSVT. Only Vocabulary Minus Digit Span, TMT Ratio and Coin in the Hand achieved an appropriate level of specificity, suggesting that, of the measures examined, these PVTs are the most likely to detect valid performance in a dementia population. However, methodological issues with these studies mean that these results should be interpreted with caution. Where PVTs have been adapted or developed for use in dementia, specificity rates have improved. The RBANS ES (Novitski et al., 2012) and dementia profiles of the WMT, MSVT and NV-MSVT (Green et al., 2011; Howe et al., 2007; Howe & Loring, 2009; Henry et al., 2010) all demonstrated promise for use as PVTs in this population, with specificity values approaching or exceeding the 90% threshold.

The findings of this review have revealed an inverse relationship between PVT classification accuracy and dementia severity, with almost all measures demonstrating significantly worse specificity in more advanced dementia. This raises a broader question about whether the tests are measuring what they purport to

measure. The fundamental assumption of a PVT is that it will be insensitive to cognitive dysfunction (Tombaugh, 1997): this review has demonstrated that this is not necessarily the case across a range of measures examining multiple cognitive domains, with the possible exception of the aforementioned PVTs which have been specifically adapted for this population. In routine clinical practice, the need for accurate PVTs likely diminishes with increasing dementia severity as obvious functional or behavioural impairments reduce the likelihood of patient misclassification (Loring et al., 2016; Bortnick et al., 2013). Nonetheless, these results strongly imply that PVTs are less robust to cognitive impairment than generally assumed.

PVT specificity was further shown to vary related to dementia subtype. Different dementia subtypes differ in their cognitive profiles; for example, compared to subcortical VD patients, AD patients are typically more impaired in episodic memory and less impaired in semantic memory, executive functioning, attention and visuospatial and perceptual skills (Graham, Emery & Hodges, 2004). It makes intuitive sense, therefore, that specificity will be lower for measures which tap the cognitive domain affected by a particular dementia subtype. For example, it has been suggested that the motor cortex is relatively spared in the early stages of AD, which may have contributed to the high pass rate on the Finger Tapping measure in this population versus VD and non-AD patients (Arnold et al., 2005; Dean et al., 2009). One might also query if the poor performance of FTD patients compared to AD and VD patients on the Three- and Four Digits Timed tasks (Dean et al., 2009) is related to pathology in the frontal or anterior temporal lobes and concomitant effects on executive processes, however, this would require further investigation. More generally, Rudman et al. (2011) found that new learning ability was the strongest

predictor of failure across the TOMM, RFIT, MSVT and NV-MSVT. It is perhaps unsurprising, therefore, that dementia populations, who frequently exhibit difficulties in this domain (Aretouli & Brandt, 2010) would fail these measures as all have a learning and recall component. This points to the need to consider the appropriateness of the measure in light of the individual's subjective cognitive complaints and in the context of the neuropsychological assessment as a whole.

Clinical implications

The conclusions of this review have a number of implications for clinical practice. Firstly, the use of the aforementioned measures with traditional cut-offs is not recommended in dementia (or suspected dementia) populations due to the high risk of making a false positive error (classifying performance as invalid when it is not). Although adjusting cut-offs generally leads to an improvement in specificity, the results of known-groups and simulator studies indicate that applying more stringent thresholds will likely have a knock-on effect on test sensitivity (the ability to correctly identify invalid performance). Although advice regarding PVT cut-offs is generally to focus on specificity "while letting the sensitivity chips fall where they may" (Greve & Bianchini, 2004; p. 536), making dramatic modifications to conventional thresholds to account for the particular challenges of testing a dementia population is likely to significantly limit their clinical utility. Using measures with profile algorithms designed to account for the particular difficulties observed in dementia appears to offer the most reliable means of assessing performance validity in this population, with preliminary evidence for the RBANS ES suggesting that sensitivity is not hugely compromised (Dunham et al., 2014). Of note is that this measure has been found to be unsuitable for use in cognitively intact individuals

(Novitski et al., 2012), which limits its utility to situations where there is a suspected or confirmed diagnosis of dementia.

The fact that many PVTs were found to be sensitive to cognitive impairment raises questions about the use of PVTs more generally: if specificity levels are problematic in severe dementia then can the same be said for other neurological populations such as severe brain injury or intellectual disability? These data appear to illustrate that there comes a point at which the PVT stops measuring performance validity and begins to measure cognitive dysfunction, which clearly invalidates the test (Walter et al., 2014). These findings underscore the importance of robust biopsychosocial formulation in the context and information should be triangulated from multiple sources (for example results of PVTs, cognitive test scores, patient self-report, consideration of psychological state and co-morbid diagnoses) prior to reaching an opinion on an individual's performance validity (McMillan et al., 2009).

Methodological critique of the studies

A number of methodological issues have been raised through examination of the data contributing to this review. Some of these weaknesses are inherent to the study of dementia populations, such as the use of mixed etiology groups and lack of consistency in diagnostic categorisation (Kiewel et al., 2012). As it was not possible to randomly assign patients to groups internal validity was reduced across all studies (Sollman & Berry, 2011). Many studies further had small overall sample sizes and compared groups which were unequal in size. For example, Dean et al. (2009) included some subgroups with only one participant, which clearly limits how these findings might be generalised to the population as a whole. Suggested cut-offs based

on these small sample sizes may also be less stable, and so less confidence can be placed on accurate classification should they be used (Greve & Bianchini, 2004).

Seven studies also failed to adequately control for confounding variables such as age, estimated premorbid education or mood. Individual differences should always be considered when interpreting the results of neuropsychological measures, whether they are PVTs or assessments of particular cognitive domains (Duff et al., 2011). For example, it has been shown that scores on the RBANS EI are influenced by age and education (Duff et al., 2011), and that depressed older adults are more likely to be classified as 'suspect effort' on this measure (Barker et al., 2010). As such, the fact that a third of studies reviewed failed to establish demographic equivalence between groups is problematic.

Incomplete information regarding sensitivity of the identified measures is a major limitation of these studies, as only five studies utilised a design which would allow for an estimate of sensitivity to be produced, and these data were largely based on non-dementia comparison groups. As has been highlighted, PVTs have limited practical utility if they do not capture invalid performance when it occurs: the focus may be on specificity however, in reality, any PVT should offer a balance between the two (Bortnick et al., 2013). Further known-groups studies using dementia subsamples would address this issue and help to extend our understanding of the impact of adjusting cut-off scores for use in this population.

A key conceptual issue with this review rests on the assumption that PVT failures represent 'false positives' as the clinical groups are 'bona fide' dementia patients. PVT failures may well be 'false positives', but only if one assumes that the PVT is singular in scope. Is it possible, however, that there are other circumstances

which might produce 'false positives' over and above how the person applies themselves on the test? In reality, a range of factors may influence PVT failure in dementia samples, some of which have been illustrated in this review, for example severe cognitive impairment and dementia subtype. Other factors, however, must also be taken into consideration with this population (and, indeed, with all populations who present for neuropsychological assessment), including physical health issues (such as chronic health problems, polypharmacy and sensory impairments; Storandt, 1994) and factors affecting engagement (for example psychiatric distress or fatigue; Walter et al., 2014). Indeed, as Suesse et al. (2015) emphasise, there may be no such thing as a verifiable false positive on PVTs and, as such, they opt to refer to 'unexplained failures'. Again, this points to the critical importance of using PVT results as just one contributing element in the broader clinical formulation to ensure that inappropriate conclusions are not drawn.

Limitations of the review

The results of this review should be considered somewhat tentative for a number of reasons. Firstly, the results of the classification accuracy analysis were based on a small number of studies which did not cover the full range of PVTs used in clinical practice. It furthermore did not look to examine the classification accuracy of measures used in combination. Clinical guidelines (Bush et al., 2005; McMillan et al., 2009) emphasise the importance of using two or more PVTs in combination to reduce the risk of false positives (Larrabee, 2014). Future studies investigating optimal combinations of PVTs for use in a dementia population would provide clinically useful information (Vickery et al., 2001). It is also acknowledged that the weighted means approach presents some methodological issues, for example poor

control of extraneous variables. Ideally, specificity data should also be analysed alongside sensitivity data as there is an inverse relationship between the two, however, this was not possible with this dataset due to the issues outlined above. Nonetheless, it is argued that the generation of pooled specificity can provide clinicians with the most likely estimate of the classification accuracy of these measures until alternative methods of analysis are available (Deville et al., 2002).

A key issue to also consider is that even with the calculation of specificity and sensitivity values, these indicators cannot provide a full impression of the accuracy of clinical decision-making using a particular PVT at a particular cut-off (Vickery et al., 2001). To fully inform the clinical application of PVTs with a given population in a given setting, one must also consider the potential base rate of invalid performance in the population being assessed (McMillan et al., 2009). It is only with this information that one can begin to determine what proportion of a test's classifications are likely to be accurate: these values are referred to as the positive and negative predictive powers of a test (Sollman & Berry, 2011). To the author's knowledge, base rate information regarding the prevalence of invalid performance in clinical dementia settings has not been established, and indeed the PVT literature as a whole frequently fails to examine the influence of base rates when estimating the accuracy of prediction models (Rosenfeld, Sands & Van Gorp, 2000). Classification accuracy statistics and base rate data are interdependent, and it is recommended that without it clinicians should be cautious about interpreting performance validity data, and where possible acknowledge limitations in assessment methods by providing probability estimates or possible error rates (Rosenfeld et al., 2000). Understanding the base rate of invalid performance in clinical neuropsychological practice is therefore an important next step in validating some of the findings from this review.

References

Altman, D. G., & Bland, J. M. (1994). Diagnostic tests. 1: Sensitivity and specificity. *BMJ: British Medical Journal*, *308*(6943), 1552.

Aretouli, E., & Brandt, J. (2010). Episodic memory in dementia: Characteristics of new learning that differentiate Alzheimer's, Huntington's, and Parkinson's diseases. *Archives of clinical neuropsychology*, *25*(5), 396-409.

Arnold, G., Boone, K. B., Lu, P., Dean, A., Wen, J., Nitch, S., & McPherson, S. (2005). Sensitivity and specificity of finger tapping test scores for the detection of suspect effort. *The Clinical Neuropsychologist*, *19*(1), 105-120.

Babikian, T., Boone, K. B., Lu, P., & Arnold, G. (2006). Sensitivity and specificity of various digit span scores in the detection of suspect effort. *The Clinical Neuropsychologist*, *20*(1), 145-159.

Barker, M. D., Horner, M. D., & Bachman, D. L. (2010). Embedded indices of effort in the repeatable battery for the assessment of neuropsychological status (RBANS) in a geriatric sample. *The Clinical Neuropsychologist*, *24*(6), 1064-1077.

Binder, L. M., Villanueva, M. R., Howieson, D., & Moore, R. T. (1993). The Rey AVLT recognition memory task measures motivational impairment after mild head trauma. *Archives of Clinical Neuropsychology*, *8*(2), 137-147.

Boone, K. B., Lu, P., Back, C., King, C., Lee, A., Philpott, L. & Warner-Chacon, K. (2002a). Sensitivity and specificity of the Rey Dot Counting Test in patients with suspect effort and various clinical samples. *Archives of Clinical Neuropsychology*, *17*(7), 625-642.

Boone, K. B., Lu, P., & Herzberg, D. S. (2002). *The Dot Counting Test Manual*. Los Angeles, CA: Western Psychological Services.

Boone, K. B., Lu, P., & Wen, J. (2005). Comparison of various RAVLT scores in the detection of noncredible memory performance. *Archives of Clinical Neuropsychology*, *20*(3), 301-319.

Boone, K. B., Salazar, X., Lu, P., Warner-Chacon, K., & Razani, J. (2002). The Rey 15-item recognition trial: A technique to enhance sensitivity of the Rey 15-item memorization test. *Journal of clinical and Experimental Neuropsychology*, *24*(5), 561-573.

Bortnik, K. E., Horner, M. D., & Bachman, D. L. (2013). Performance on standard indexes of effort among patients with dementia. *Applied Neuropsychology: Adult*, *20*(4), 233-242.

Brooks, B. L., Sherman, E. M., Iverson, G. L., Slick, D. J., & Strauss, E. (2011). Psychometric foundations for the interpretation of neuropsychological test results. In *The Little Black Book of Neuropsychology* (pp. 893-922). Springer US.

Burton, R. L., Enright, J., O'Connell, M. E., Lanting, S., & Morgan, D. (2015). RBANS Embedded Measures of Suboptimal Effort in Dementia: Effort Scale Has a Lower Failure Rate than the Effort Index. *Archives of Clinical Neuropsychology*, *30*(1), 1-6.

Bush, S. S., Ruff, R. M., Tröster, A. I., Barth, J. T., Koffler, S. P., Pliskin, N. H., Reynolds, C. R. & Silver, C. H. (2005). Symptom validity assessment: Practice issues and medical necessity: NAN Policy & Planning Committee. *Archives of Clinical Neuropsychology*, *20*(4), 419-426.

Davis, J. J., Millis, S. R., & Axelrod, B. N. (2012). Derivation of an embedded Rey Auditory Verbal Learning Test performance validity indicator. *The Clinical Neuropsychologist*, *26*(8), 1397-1408.

Dean, A. C., Victor, T. L., Boone, K. B., Philpott, L. M., & Hess, R. A. (2009). Dementia and effort test performance. *The Clinical Neuropsychologist*, *23*(1), 133-152.

Devillé, W. L., Buntinx, F., Bouter, L. M., Montori, V. M., De Vet, H. C., Van der Windt, D. A., & Bezemer, D. P. (2002). Conducting systematic reviews of diagnostic studies: didactic guidelines. *BMC Medical Research Methodology*, *2*(1), 1.

Dubois, B., Feldman, H. H., Jacova, C., Cummings, J. L., DeKosky, S. T., Barberger-Gateau, P., & Gauthier, S. (2010). Revising the definition of Alzheimer's disease: a new lexicon. *The Lancet Neurology*, *9*(11), 1118-1127.

Duff, K., Spering, C. C., O'Bryant, S. E., Beglinger, L. J., Moser, D. J., Bayless, J. D. & Scott, J. G. (2011). The RBANS Effort Index: Base rates in geriatric samples. *Applied neuropsychology*, *18*(1), 11-17.

Dunham, K. J., Shadi, S., Sofko, C. A., Denney, R. L., & Calloway, J. (2014). Comparison of the repeatable battery for the assessment of neuropsychological status Effort Scale and Effort Index in a dementia sample. *Archives of Clinical Neuropsychology*, *29*(7), 633-641.

Faust, D., Hart, K., Guilmette, T. J., & Arkes, H. R. (1988). Neuropsychologists' capacity to detect adolescent malingerers. *Professional Psychology-Research and Practice, 19*(5), 508-515.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, *12*(3), 189-198.

Graham, N. L., Emery, T., & Hodges, J. R. (2004). Distinctive cognitive profiles in Alzheimer's disease and subcortical vascular dementia. *Journal of Neurology, Neurosurgery & Psychiatry*, *75*(1), 61-71.

Green, P. (2004). *Medical Symptom Validity Test (MSVT) for Microsoft Windows: User's Manual*. Paul Green Pub.

Green, P., Allen, L. M., & Astner, K. (1996). *The Word Memory Test: A user's guide to the oral and computer-administered forms*, US Version 1.1. Durham, NC: CogniSyst.

Green, P. (2008). *Green's Non-Verbal Medical Symptom Validity Test (NV-MSVT) for Microsoft Windows. User's Manual 1.0*. Edmonton, Canada: Green's Publishing.

Green, P., Montijo, J., & Brockhaus, R. (2011). High specificity of the Word Memory Test and Medical Symptom Validity Test in groups with severe verbal memory impairment. *Applied Neuropsychology, 18*(2), 86-94.

Greiffenstein, M. F., Baker, W. J., & Gola, T. (1994). Validation of malingered amnesia measures with a large clinical sample. *Psychological Assessment*, *6*(3), 218.

Greve, K.W., & Bianchini, K.J. (2004). Setting empirical cut-offs on psychometric indicators of negative response bias: A methodological commentary with recommendations. *Archives of Clinical Neuropsychology, 19*, 533–541.

Greve, K. W., Bianchini, K. J., & Doane, B. M. (2006). Classification accuracy of the Test of Memory Malingering in traumatic brain injury: Results of a known-groups analysis. *Journal of Clinical and Experimental Neuropsychology, 28*(7), 1176-1190.

Harvey, P. D. (2012). Clinical applications of neuropsychological assessment. *Dialogues in clinical neuroscience, 14*(1), 91.

Henry, M., Merten, T., Wolf, S. A., & Harth, S. (2010). Nonverbal Medical Symptom Validity Test performance of elderly healthy adults and clinical neurology patients. *Journal of Clinical and Experimental Neuropsychology*, *32*(1), 19-27.

Howe, L. L., Anderson, A. M., Kaufman, D. A., Sachs, B. C., & Loring, D. W. (2007). Characterization of the Medical Symptom Validity Test in evaluation of clinically referred memory disorders clinic patients. *Archives of Clinical Neuropsychology*, *22*(6), 753-761.

Howe, L. L., & Loring, D. W. (2009). Classification accuracy and predictive ability of the Medical Symptom Validity Test's dementia profile and general memory impairment profile. *The Clinical Neuropsychologist, 23*(2), 329-342.

Iliffe S., Manthorpe, J. & Eden, A. (2003). Sooner or later? Issues in the early diagnosis of dementia in general practice: a qualitative study. *Family Practice*, *20,* 376–381.

Iverson, G. L. & Brooks, B. L. (2011). Improving accuracy for identifying cognitive impairment. In Scott, J. G., & Schoenberg, M. R. (Eds). *The Little Black Book of Neuropsychology* (pp. 923-950). Springer US.

Iverson, G. L., & Franzen, M. D. (1994). The Recognition Memory Test, digit span, and Knox Cube Test as markers of malingered memory impairment. *Assessment*, *1*(4), 323-334.

Iverson, G. L., Lange, R. T., Green, P., & Franzen, M. D. (2002). Detecting exaggeration and malingering with the Trail Making Test. *The Clinical Neuropsychologist*, *16*(3), 398-406.

Iverson, G. L., & Tulsky, D. S. (2003). Detecting malingering on the WAIS-III: Unusual Digit Span performance patterns in the normal population and in clinical groups. *Archives of Clinical Neuropsychology*, *18*(1), 1-9.

Kapur, N. (1994) The Coin-in-the-Hand test: a new "bed-side" test for the detection of malingering in patients with suspected memory disorder. *Journal of Neurology, Neurosurgery, and Psychiatry*, 57, 385–386.

Katrak, P., Bialocerkowski, A. E., Massy-Westropp, N., Kumar, V. S., & Grimmer, K. A. (2004). A systematic review of the content of critical appraisal tools. *BMC medical research methodology*, *4*(1), 1.

Kiewel, N. A., Wisdom, N. M., Bradshaw, M. R., Pastorek, N. J., & Strutt, A. M. (2012). A retrospective review of Digit Span-related effort indicators in probable Alzheimer's disease patients. *The Clinical Neuropsychologist*, *26*(6), 965-974.

Scott Killgore, W. I., & DellaPietra, L. (2000). Using the WMS-III to detect malingering: Empirical validation of the Rarely Missed Index (RMI). *Journal of Clinical and Experimental Neuropsychology*, *22*(6), 761-771.

Kmet, L. M., Lee, R. C., & Cook, L. S. (2004). Standard quality assessment criteria for evaluating primary research papers from a variety of fields. *Accessed online at* http://www.biomedcentral.com/content/supplementary/1471-2393-14-52-s2.pdf (25/02/2016)

Larrabee, G. J. (2008). Aggregation across multiple indicators improves the detection of malingering: Relationship to likelihood ratios. *The Clinical Neuropsychologist*, *22*(4), 666-679.

Larrabee, G. J. (2012). Performance validity and symptom validity in neuropsychological assessment. *Journal of the International Neuropsychological Society*, *18*(04), 625-630.

Larrabee, G. J. (2014). False-positive rates associated with the use of multiple performance and symptom validity tests. *Archives of Clinical Neuropsychology*, *29*(4), 364-373.

Lezak, M. D. (1983). *Neuropsychological Assessment* (2nd ed.). New York: Oxford University Press.

Lezak, M. D. (1995). *Neuropsychological Assessment* (3rd ed.). New York: Oxford University Press.

Loring, D. W., Larrabee, G. J., Lee, G. P., & Meador, K. J. (2007). Victoria Symptom Validity Test performance in a heterogenous clinical sample. *The Clinical Neuropsychologist*, *21*(3), 522-531.

Loring, D. W., Goldstein, F. C., Chen, C., Drane, D. L., Lah, J. J., Zhao, L., & Larrabee, G. J. (2016). False-positive error rates for reliable digit span and auditory verbal learning test performance validity measures in amnestic mild

cognitive impairment and early Alzheimer disease. *Archives of Clinical Neuropsychology*, *31*(4), 313-331.

Lu, P. H., Boone, K. B., Cozolino, L., & Mitchell, C. (2003). Effectiveness of the Rey-Osterrieth Complex Figure Test and the Meyers and Meyers recognition trial in the detection of suspect effort. *The Clinical Neuropsychologist*, *17*(3), 426-440.

MKhann, G. M., Knopman, D. S., Chertkow, H., Hyman, B. T., Jack, C. R., Kawas, C. H., & Mohs, R. C. (2011). The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*, *7*(3), 263-269.

McMillan, T.M., Anderson, S., Baker, G., Berger, M., Powell, G.E., and Knight, R. (2009) *Assessment of effort in clinical testing of cognitive functioning for adults*. British Psychological Society, pp. 1-27.

Merten, T., Bossink, L., & Schmand, B. (2007). On the limits of effort testing: Symptom validity tests and severity of neurocognitive symptoms in nonlitigant patients. *Journal of Clinical and Experimental Neuropsychology*, *29*(3), 308-318.

Nitch, S., Boone, K. B., Wen, J., Arnold, G., & Alfano, K. (2006). The utility of the Rey Word Recognition Test in the detection of suspect effort. *The Clinical Neuropsychologist*, *20*(4), 873-887.

National Health and Medical Research Council (2000). *How to Review the Evidence: Systematic Identification and Review of the Scientific Literature.* Canberra: National Health and Medical Research Council.

Novitski, J., Steele, S., Karantzoulis, S., & Randolph, C. (2012). The repeatable battery for the assessment of neuropsychological status effort scale. *Archives of Clinical Neuropsychology*, *27*(2), 190-195.

O'Bryant, S. E., Waring, S. C., Cullum, C. M., Hall, J., Lacritz, L., Massman, P. J., & Doody, R. (2008). Staging dementia using Clinical Dementia Rating Scale Sum of Boxes scores: a Texas Alzheimer's research consortium study. *Archives of Neurology*, *65*(8), 1091-1095.

Pallant, J. (2013). *SPSS survival manual*. McGraw-Hill Education (UK).

Petersen, R. C., Smith, G. E., Waring, S. C., Ivnik, R. J., Tangalos, E. G., & Kokmen, E. (1999). Mild cognitive impairment: clinical characterization and outcome. *Archives of Neurology*, *56*(3), 303-308.

Randolph, C. (1998). RBANS manual: Repeatable battery for the assessment of neuropsychological status. *San Antonio, TX: The Psychological Corporation*.

Rudman, N., Oyebode, J. R., Jones, C. A., & Bentham, P. (2011). An investigation into the validity of effort tests in a working age dementia population. *Aging & Mental Health*, *15*(1), 47-57.

Rosenfeld, B., Sands, S. A., & Van Gorp, W. G. (2000). Have we forgotten the base rate problem? Methodological issues in the detection of distortion. *Archives of Clinical Neuropsychology*, *15*(4), 349-359.

Rudman, N., Oyebode, J. R., Jones, C. A., & Bentham, P. (2011). An investigation into the validity of effort tests in a working age dementia population. *Aging & Mental Health*, *15*(1), 47-57.

Salmon, D. P., & Bondi, M. W. (2009). Neuropsychological assessment of dementia. *Annual review of psychology*, *60*, 257.

Schagen, S., Schmand, B., Sterke, S. D., & Lindeboom, J. (1997). Amsterdam Short-Term Memory Test: A new procedure for the detection of feigned memory deficits. *Journal of Clinical and Experimental Neuropsychology*, *19*(1), 43-51.

Schroeder, R. W., & Marshall, P. S. (2010). Validation of the Sentence Repetition Test as a measure of suspect effort. *The Clinical Neuropsychologist*, *24*(2), 326-343.

Schroeder, R. W., Peck, C. P., Buddin Jr, W. H., Heinrichs, R. J., & Baade, L. E. (2012). The Coin-in-the-Hand Test and dementia: More evidence for a screening test for neurocognitive symptom exaggeration. *Cognitive and Behavioral Neurology*, *25*(3), 139-143.

Schroeder, R. W., Twumasi-Ankrah, P., Baade, L. E., & Marshall, P. S. (2012). Reliable digit span: A systematic review and cross-validation study. *Assessment*, *19*(1), 21-30.

Sherman, D. S., Boone, K. B., Lu, P., & Razani, J. (2002). Re-examination of a Rey auditory verbal learning test/Rey complex figure discriminant function to detect suspect effort. *The Clinical Neuropsychologist*, *16*(3), 242-250.

Silverberg, N. D., Wertheimer, J. C., & Fichtenberg, N. L. (2007). An effort index for the Repeatable Battery for the Assessment of Neuropsychological Status (RBANS). *The Clinical Neuropsychologist*, *21*(5), 841-854.

Singhal, A., Green, P., Ashaye, K., Shankar, K., & Gill, D. (2009). High specificity of the Medical Symptom Validity Test in patients with very severe memory impairment. *Archives of Clinical Neuropsychology*, *24*(8), 721-728.

Slick, D., Hopp, G., & Strauss, E. (1997). *The Victoria Symptom Validity Test*. Odessa, FL: Psychological Assessment Resources.

Slick, D. J., Sherman, E. M., & Iverson, G. L. (1999). Diagnostic criteria for malingered neurocognitive dysfunction: Proposed standards for clinical practice and research. *The Clinical Neuropsychologist*, *13*(4), 545-561.

Sollman, M. J., & Berry, D. T. (2011). Detection of inadequate effort on neuropsychological testing: A meta-analytic update and extension. *Archives of Clinical Neuropsychology*, 26(8), 774-789.

Sperling, R. A., Aisen, P. S., Beckett, L. A., Bennett, D. A., Craft, S., Fagan, A. M., & Phelps, C. H. (2011). Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*, *7*(3), 280-292.

Storandt, M. (1994). General principles of assessment of older adults. In M. Storandt and G. R. Van den Bos (Eds.), *Neuropsychological assessment of depression and dementia in older adults: A clinician's guide* pp. 7–32. Washington, DC: American Psychological Association.

Strauss, E., Sherman, E. M., & Spreen, O. (2006). *A compendium of neuropsychological tests: Administration, norms, and commentary*. Oxford University Press, USA.

Suesse, M., Wong, V. W., Stamper, L. L., Carpenter, K. N., & Scott, R. B. (2015). Evaluating the Clinical Utility of the Medical Symptom Validity Test (MSVT): A Clinical Series. *The Clinical Neuropsychologist*, *29*(2), 214-231.

Teichner, G., & Wagner, M. T. (2004). The Test of Memory Malingering (TOMM): Normative data from cognitively intact, cognitively impaired, and elderly patients with dementia. *Archives of Clinical Neuropsychology*, *19*(3), 455-464.

Tombaugh, T. N. (1996). *Test of memory malingering: TOMM*. New York/Toronto: MHS.

Tombaugh, T. N. (1997). The Test of Memory Malingering (TOMM): Normative data from cognitively intact and cognitively impaired individuals. *Psychological Assessment*, *9*(3), 260.

Vickery, C. D., Berry, D. T., Inman, T. H., Harris, M. J., & Orey, S. A. (2001). Detection of inadequate effort on neuropsychological testing: A meta-analytic review of selected procedures. *Archives of Clinical Neuropsychology*, *16*(1), 45-73.

Walter, J., Morris, J., Swier-Vosnos, A., & Pliskin, N. (2014). Effects of severity of dementia on a symptom validity measure. *The Clinical Neuropsychologist*, *28*(7), 1197-1208.

Zenisek, R., Millis, S. R., Banks, S. J., & Miller, J. B. (2016). Prevalence of below-criterion Reliable Digit Span scores in a clinical sample of older adults. *Archives of Clinical Neuropsychology*, *31*(5), 426-433.

Part 2: Empirical Paper

Performance validity testing in an NHS acquired brain injury sample

Abstract

Aims: Performance validity tests (PVTs) were initially designed to assist the actuarial judgement of test-taking behaviour in medicolegal settings. Less is known about the performance of NHS patient populations on PVTs. The study aimed to establish the base rate of failure on commonly-used PVTs in a sample of acquired brain injury patients representative of those accessing NHS services. A secondary aim was to examine differences between PVT pass and fail groups on cognitive testing.

Method: Clinical neuropsychological assessment data was gathered across three NHS acquired brain injury services. Data included at least one embedded PVT (derived from the WAIS Digit Span) and one standalone PVT (the Test of Memory Malingering [TOMM]). Individuals passing and failing PVTs were compared in terms of demographic variables and scores on tests of cognitive function.

Results: The base rate of TOMM failure was 10%. Only 4% of patients failed the TOMM plus an additional embedded PVT. Individuals failing PVTs demonstrated reduced scores on cognitive testing.

Conclusions: Findings suggest that whilst failure on multiple PVTs is relatively rare, a substantial minority of patients in clinical settings will fail one PVT. Individuals failing PVTs were more likely to have lower scores on tests of cognitive function, however, the factors underpinning these supressed scores were unclear. More research is required to delineate the contributions of cognitive and psychological variables to PVT failure. It is thus recommended that PVTs be used with caution in clinical settings.

Introduction

Fundamental to the practice of neuropsychology is the assumption of a relationship between an individual's performance as measured by neuropsychological tests and a condition of the brain (Lezak, Howieson, & Loring, 2004). Cognitive testing is employed because it is understood that the resulting scores, when taken as part of a biopsychosocial neuropsychological formulation, will inform the clinician's understanding of the individual and their needs. As such, it is important for clinicians to critically consider the reliability of their measures to enable judgements regarding whether scores represent a valid estimate of the individual's functioning at the time of testing.

The quality of the contributing test data is impacted by broad range of factors, including the test-taking environment, the psychometric properties of the test, the proficiency of the examiner and myriad influences acting upon the examinee. As cognitive tests are performance-based assessment methods the individual's test-taking behaviour is of particular relevance, as distortions in performance could result in scores which are not representative of 'true' deficits (Millis, 2009). Examinee 'effort' is a behavioural factor gaining increasing prominence in the literature (McMillan et al., 2009; McCarter, Walton, Brooks, & Powell, 2009). Clinical observation alone has been shown to be an unreliable method of discriminating which examinees are exerting 'adequate effort' in assessments, despite studies indicating high levels of clinician self-rated confidence in their judgements (Heaton, Smith, Lehman, & Vogt, 1978; Dawes, Faust, & Meehl, 1989; Faust, Hart, Guilmette, & Arkes, 1998). Clinicians are therefore at risk of making false positive errors by concluding that supressed scores are due to brain damage when this is not the case (Hampson, Kemp, Coughlan, Moulin, & Bhakta, 2014). As

such, the empirical validation of examinee performance is crucial for good practice (McMillan et al., 2009).

The assessment of performance validity

Research in the area of performance validity assessment initially focused on forensic and medicolegal samples (Merten, Bossink, & Schmand, 2007). These populations have clear issues of prospective secondary gain, which presented a challenge for clinicians wishing to obtain an unbiased impression of the individual's deficits in order to distinguish between potential malingerers (individuals producing false or grossly exaggerated symptoms motivated by external incentives; Slick, Sherman, & Iverson, 1999; American Psychiatric Association, 2000) and legitimate claimants. The presence of superordinate moderating variables (such as the potential for secondary gain) has been demonstrated to exert a significant impact on an individual's clinical presentation across a range of domains including cognitive, physical, sensory and psychiatric symptoms. A meta-analysis by Belanger, Curtiss, Demery, Lebowitz, & Vanderploeg (2005) noted that patients with mild traumatic brain injury (mTBI) involved in litigation proceedings were more likely to remain symptomatic or deteriorate over time compared to non-litigants who recovered on average by three months post-injury. The presence of litigation was found to be the key discriminating factor influencing the patients' reported symptomatology.

The practice of empirical performance validity testing has emerged in this context, motivated by increased recognition of the frequency of suboptimal performance in these settings and the concomitant need to provide a more stringent means of identifying this behaviour when it occurs (Strauss, Sherman, & Spreen, 2006, p. 1145). Measures have thus been developed which can be used to assist

clinical judgements regarding the reliability of the test data, known variously as 'effort tests', 'performance validity tests' (PVTs), 'symptom validity tests' (SVTs) or measures of 'response bias'. All purport to be sensitive to test-taking 'effort' whilst remaining insensitive to acquired cognitive impairment. PVTs appear subjectively difficult to the examinee but have been devised such that even individuals with globally reduced cognitive ability are mostly able to perform at or near ceiling (Tombaugh, 1997). Normative data from *bona fide* clinical groups, litigating samples and identified malingerers has contributed to the formation of 'cut-off' scores, that is, identification of the lowest possible score that the individual must obtain to 'pass' the test. Scoring below this threshold (and therefore below most clinical patients but similarly to those simulating impairment or engaging in probable malingering; Holdnack, Millis, Larrabee, & Iverson, 2013) would raise questions about performance across the test battery as it is assumed that this indicates a non-neurological influence on performance (Locke, Smigielski, Powell & Stevens, 2008).

It should be borne in mind that whilst test failure may reflect a purposeful intention to deceive through underperformance, it may equally be related to other factors, for example abnormal arousal, iatrogenic symptoms, somatoform disorders, or that examinees have a benign lack of interest in psychometric testing (Bunnage, Eichinger, Pearce, Duckworth & Newson, 2008; McMillan et al., 2009). The ubiquitous term 'effort testing' is therefore something of a misnomer which reflects the origins of these measures in the medicolegal field. In a clinical context, to characterise an individual as 'malingering', or indeed as exerting 'poor effort', on the basis of PVT failure alone may represent a serious false positive error with significant negative consequences (for example denial of benefits or treatment). Neuropsychological tests are only able to measure behaviour and not intent (Boone

et al., 2002), which is essentially unknowable (Suesse, Wong, Stamper, Carpenter, & Scott, 2015). To reflect a broader understanding of this construct which may be more applicable to clinical practice, the terms 'performance validity' and associated 'performance validity tests' (PVTs) will be utilised in this paper as per Larrabee (2012). PVTs are differentiated from 'symptom validity tests' (SVTs) in that they pertain specifically to the validity of cognitive test performance, as opposed to the validity of self-reported symptomatology which is outside the scope of the current paper.

Performance validity tests (PVTs)

A range of measures have been designed and validated for the assessment of performance validity. Some have been developed specifically for this purpose, for example the Test of Memory Malingering (TOMM; Tombaugh, 1996). These frequently take the form of forced-choice recognition memory measures administered as stand-alone tests within a larger neuropsychological battery, where scoring significantly below chance is viewed as indicative of 'non-credible' performance, but where thresholds well above this level may still indicate 'invalid' performance (Tombaugh, 1996; Frederick & Speed, 2007; Boone, 2007). In validation studies, the TOMM has been shown to be relatively insensitive to a range of physical and affective disorders in addition to age and level of education (Tombaugh, 1996; Rees, Tombaugh, Gansler, & Moczynski, 1998; Rees, Tombaugh, & Boulay, 2001; Ashendorf, Constantinou, & McCaffrey, 2004; Iverson, Le Page, Koehler, Shojania, & Badii, 2007).

Using 'embedded' performance validity measures is a second approach. These measures are derived from existing neuropsychological tests that have

additional value in identifying possible 'non-credible' performance, meaning that they serve a 'double duty' and therefore do not increase testing time in addition to being less vulnerable to coaching (McMillan et al., 2009; Heilbronner et al., 2009). One example is the Reliable Digit Span (RDS), derived originally from the Wechsler Adult Intelligence Scale - Revised (WAIS-R; Wechsler, 1981) Digit Span subtest (Greiffenstein, Baker, & Gola, 1994). This is the sum of the longest string of digits repeated without error over two trials in both forward and backward conditions, where a score of less than or equal to six is generally accepted as indicating suboptimal performance (Schroeder, Twumasi-Ankrah, Baade & Marshall, 2012). In a study of 17 embedded and three stand-alone indices, Miele, Gunner, Lynch, & McCaffrey (2012) found that compared to individuals passing the RDS, examinees failing the RDS were over 11 times more likely to be identified as 'suboptimal effort' (characterised via failure on a stand-alone PVT), suggesting that it is a useful predictor of invalid performance. These authors also demonstrated, however, that embedded measures overall had reduced sensitivity (ability to detect true positives) and specificity (ability to detect true negatives) as compared to stand-alone PVTs. This is perhaps unsurprising given that embedded measures are derived from tests originally designed to assess ability as opposed to 'effort' (Schutte & Axelrod, 2012). With this in mind, use of a combination of embedded and stand-alone measures (a 'multi-method, multi-test' approach) has been recommended for the assessment of performance validity in neuropsychological evaluation by both the British Psychological Society (BPS; McMillan et al., 2009) and the American Academy of Clinical Neuropsychology (AACN; Heilbronner et al., 2009).

What can PVT failure tell us about neuropsychological test data?

Research indicates that the presence of PVT failure undermines the confidence that can be placed in the remaining test data. For example, Fox (2011) conducted an archival study of 220 cases and found that when just one PVT was failed, there was no correlation between performance on cognitive tests and documented brain damage. Similarly, Green, Rohling, Lees-Haley, & Allen (2001) found that 53% of the variance in neuropsychological test data from 904 compensation claimants could be explained by 'effort', compared to only 11% attributed to years of education and 4% to age. This was corroborated by Victor, Boone, Serpa, Buehler, & Ziegler (2009): significantly lower Full Scale IQ was noted in 'non-credible' versus 'credible' patients (that is, those failing versus passing PVTs), which was thought to be secondary to higher levels of response bias as education levels were not significantly different across the groups. In a UK compensation-seeking sample, Moss, Jones, Fokias, & Quinn (2003) likewise found no evidence of a relationship between head injury severity and IQ and memory indices in individuals failing an established PVT. PVT failure thus appears to capture an element of test-taking behaviour which is linked to achievement on cognitive testing.

The use of PVTs in clinical settings

With an improved understanding of how superordinate factors impact upon neuropsychological test data, it is increasingly recognised that examination of performance validity is necessary in all clinical assessments which aim to understand brain-behaviour relationships, contribute to diagnoses or formulate treatment recommendations (Kemp et al., 2008). In line with this, professional guidelines in

both the US (Bush et al., 2005; Heilbronner et al., 2009) and UK (McMillan et al., 2009) emphasise that performance validity testing should no longer be the sole domain of clinicians working in forensic or medicolegal settings, and advocate incorporating PVTs as standard in clinical neuropsychological practice. The direct application of existing knowledge regarding PVTs, however, presents a challenge to clinicians when much of the normative reference data has been derived from litigating populations. The use of PVTs in clinical settings places greater demand on a measure's predictive validity, as the clinician is attempting to distinguish between symptom amplification and clinical syndrome as opposed to between malingering and 'normalcy' (Merten, Bossink, & Schmand, 2007). Unfortunately, little data from clinical samples which could be directly applied to general clinical or NHS populations is so far available, and thus many clinicians in the UK, concerned about test reliability and potential for misclassification, continue to rely on clinical judgement as a means of assessing performance validity (McCarter et al., 2009; Hall, Worthington, & Venables, 2014; Suesse et al., 2015).

Improving classification accuracy

So far it has been highlighted that there are empirical means of assessing performance validity, and that this identification is key to establishing the reliability and validity of the neuropsychological data as a whole. However, existing PVTs are imperfect measures. Depending on the sensitivity and specificity levels of the test(s) employed, there is a greater or lesser risk of falsely identifying examinees as having invalid performance , or conversely missing examinees who are performing sub-optimally and so taking invalid results as valid. Of note is that the specificity of these measures is often set high in order to reduce the occurrence of false positive results, as there are greater consequences for the patient associated with providing a clinical

opinion of invalid performance than vice versa. As a result, sensitivity is often compromised (Greve & Bianchini, 2004). This is illustrated in a meta-analysis investigating the ability of five PVTs to discriminate between honest responders and dissimulators, which found an average specificity of almost 96% but a sensitivity of only 56% (Vickery, Berry, Inman, Harris, & Orey, 2001). To mitigate this issue, Larrabee (2003) advocates the use of a multivariate failure model, that is, failure on two or more PVTs to indicate probable invalid clinical presentation. The finding of good sensitivity and specificity for distinguishing 'credible' and 'non-credible' patients on this basis has been demonstrated by Victor and colleagues (2009).

Diagnostic accuracy can be further improved by the consideration of the prevalence, or 'base rate', of the characteristic of interest (that is,invalid performance) in the population being examined. There is an established literature indicating that low scores are a normal occurrence across a battery of neuropsychological tests (for example memory and intelligence tests). Prevalence rates are related to inter-individual variability (in factors such as level of intelligence and years of education) in addition to test inter-correlations and the number of tests administered (Brooks, Holdnack, & Iverson, 2011). In the context of this, Iverson and Brooks (2011) highlight that the primacy of deficit measurement as the means of analysing neuropsychological test batteries can lead to the (erroneous) attribution of low or unexpected test scores to a condition of the brain, making clinicians prone to drawing false positive conclusions. Understanding the base rate of low scores across batteries is therefore increasingly considered critical for facilitating the advanced interpretation of performance. This information can be used together with sensitivity and specificity data to allow for the calculation of clinically-relevant probability indices such as positive and negative predictive power (PPP and NPP), which reflect

the real-world assessment settings in which the results are used (Strauss et al., 2006, p.1149) and so provide a more psychometrically robust methodology for test analysis. PVTs are subject to the same vagaries of psychometric statistics as other neuropsychological tests: just as one low score on a battery of IQ tests cannot be considered proof of cognitive impairment, failure on one PVT may not be proof of invalid performance if we are unsure to what extent this pattern of results is common in a particular assessment setting. In other words, failure to acknowledge the influence of local base rates may result in unwarranted confidence that invalid performance is present when it is not or vice versa (Rosenfeld, Sands, & Van Gorp, 2000). Consideration of the frequency of failure on one or more measures in a multivariate battery of PVTs in a specific population or assessment setting is therefore a central factor in improving clinical classification accuracy.

Base rates of performance validity test failure

Medicolegal and forensic samples: Base rates of identified invalid neuropsychological performance vary hugely depending on clinical context. In settings with external incentives (such as litigation) studies have reported base rates from 30 to 50% (Mittenberg, Patton, Canyon, & Condit, 2002; Larrabee, Millis, & Meyers, 2009). A survey exploring the clinician-estimated base rate of malingering and symptom exaggeration in over 33,500 cases further reported evidence of this in 29% of personal injury evaluations, 30% of disability evaluations, 19% of criminal evaluations and 8% of medical evaluations, the highest estimated prevalence being in personal injury litigants with mTBI (Mittenberg et al., 2002).

Focusing specifically on base rates of PVT failure, Fox (2011) found that 35% of patients in a mixed clinical and litigating setting failed the Word Memory

Test (WMT; Green, 2003) or Computerised Test of Attention and Memory (CTAM; Fox, 2009). Similarly, a base rate of 42% failure on two or more stand-alone PVTs tests was reported by Miele and colleagues (2012) in a medicolegal setting. Moss et al. (2003) further demonstrated an overall TOMM failure rate of 31% in a UK sample of 78 patients assessed in connection with compensation claims, highlighting that the prevalence of invalid performance may be as frequent in UK medicolegal settings as indicated in the North American evidence base.

Clinical samples: It is perhaps not surprising that high base rates of invalid performance are found in forensic settings where there is clear potential for secondary gain. Significantly less data exists to inform our understanding of base rates of PVT failure in clinical populations. Here the majority of patients are non-litigating but nonetheless issues of secondary gain may still be present, for example where the individual is receiving disability benefits or discounted taxation rates. It is thus increasingly recognised that 'effort', symptom exaggeration and invalid performance may still be complicating factors of neuropsychological assessment in these contexts (McCarter et al., 2009).

A study by Locke and colleagues (2008) investigating base rates of TOMM failure in a treatment-seeking outpatient brain injury rehabilitation population is relevant in this regard. They found that almost 22% of the sample performed below cut-off on this measure, and additionally demonstrated a significant relationship between lower TOMM scores and lower cognitive test scores which they established was not secondary to the severity of the cognitive impairment, corroborating the Moss et al. (2003) findings in UK compensation claimants. Similarly, a retrospective analysis of WMT performance in 132 non-litigating NHS patients without clear external incentive found that 26% of patients failed when using the least stringent

cut-off on this measure. Failure rates rose to 37% using the most stringent cut-scores (Bunnage et al., 2008). A more recent UK NHS study by Hampson et al. (2014) examined the base rates of failure across seven PVTs (including the WMT and embedded measures from the Wechsler Adult Intelligence Scale, 3rd Edition [WAIS-III; Wechsler, 1997a] and Wechsler Memory Scale, 3rd Edition [WMS-III, Wechsler, 1997b]) and three patient populations: acute brain injury, community brain injury and epilepsy. Their findings demonstrated that a 'significant minority' of patients failed PVTs when using conventional cut-off scores: for example, failure rates on the WMT immediate and / or delayed recognition trials were 27%, 35% and 19% in the three groups respectively. Community brain injury participants were shown to have an overall higher base rate of failure across the different tests - they were also found to be more severely impaired than other participants based on their clinical history, suggesting that PVT failure may have been related to more significant cognitive impairment in this population. Indeed, when the authors re-analysed the WMT failure rates using adjusted cut-off scores (based on profile analysis comparing participant scores to individuals with identified genuine cognitive impairment) failure rates reduced. Finally, a UK study by Hall et al. (2014) also found a false positive rate of 18% using the WMT with identified non-malingering mTBI patients. This was correlated with reduced verbal memory scores in these patients, suggesting that performance below cut-off in these patients may have been indicative of verbal processing deficits in this group. These findings again underscore the importance of establishing local base rate data for specific clinical populations.

Base rates of failure in non-litigating patients are altered with the application of Larrabee's (2003) more conservative test failure criterion. Meyers and Volbrecht (2003) observed that no clinical, non-litigating patients (including patients

across a spectrum of injury severity) failed more than one PVT. Victor et al. (2009) later demonstrated that 6% of a sample of 66 'credible' patients failed two or more PVTs. Davis and Millis (2014), by contrast, found a failure rate of 15% using this threshold in a neurological no-incentive group administered seven or eight PVTs. Kemp et al. (2008) found a base rate of failure of 11% on two or more PVTs in a UK sample of non-litigating neurology patients with medically unexplained symptoms. They cite a range of potential mechanisms to account for 'suboptimal effort' in this subset, including biased information processing due to health beliefs or anxiety, somatoform symptoms and non-specific factors such as fatigue or pain. Again, this study indicates that factors beyond identifiable incentive may serve to compromise PVT performance. Taken in addition to the fact that no assessment setting can be characterised as truly incentive 'free', this highlights the complexity of this issue and the importance of considering performance validity in clinical, as well as litigious, contexts.

To summarise, it has been demonstrated that base rates of failure on one PVT in clinical populations are comparable to those found in studies conducted in medicolegal settings. Applying the two or more failure criterion unsurprisingly reduces the base rate of PVT failure, however, given that base rates of up to 15% have been found, in addition to evidence demonstrating an appreciable reduction in false positive error rates as a function of utilising this more conservative threshold (Larrabee, 2014), this represents an important consideration for clinical practice. Firstly, if an assumption is made that invalid performance rarely occurs outside of medicolegal settings, the empirical testing of this construct may be neglected and in the process of doing so, a potentially large source of test variance overlooked. In a survey of 130 UK-based practicing neuropsychologists, McCarter and colleagues

(2009) found that 26% endorsed the statement that symptom validity testing was "neither mandatory nor necessary since clinical cases rarely exaggerate" (p. 1060). As has been outlined above, performance validity testing should not solely be applied to identify symptom exaggeration or malingering. PVTs appear to capture a non-neurological dimension of performance (Bigler, 2012) and in doing so provide information on the validity of test data, thus enabling more robust formulations to be drawn up. Without administering measures to assist with analysis of test-taking behaviour, patients may be classified as impaired when this is not the case.

Secondly, where performance validity testing is conducted, failure of a single test or a series of tests may inaccurately be considered clinically significant when in fact little is known about how frequently failure might be expected in that population. The clinical consequences of this could be substantial. Depending on the assessment context, individuals given a false positive diagnosis of 'suboptimal effort' may be wrongly deprived of social entitlements, be subject to incorrect legal verdicts or provided with disadvantageous recommendations (Mossman, Wygant & Gervais, 2012; McMillan et al., 2009). As such, interpretation of performance validity test scores should be conducted with reference to available base rate data best suited to the population and assessment context in question. This would enable an indication of prevalence of 'invalid' scores among examinees with bona fide injuries, thus providing an estimate of the likelihood that such scores are false positives (Strauss, Sherman & Spreen, 2006, p. 1152).

Study aims

The aim of the present study is to extend the Locke et al. (2008) and Hampson et al. (2014) papers and investigate the base rate of PVT failure in a UK

sample of NHS patients with acquired brain injuries. A stricter criterion will be applied based on the Larrabee (2003; 2014) data supporting the practice of using two or more PVT failures as indicative of probable invalid presentation. This would inform clinicians as to the frequency of multiple test failure in this population and thus allow for more accurate estimation of the likelihood of false positives. To the author's knowledge, this will be the first study to investigate this issue in the UK, and would be in line with the research needs identified by the BPS (McMillan et al., 2009, p12). It is also the first identified study to consider TOMM performance in this context: the TOMM has been identified as the PVT most commonly used in UK clinical practice (McCarter et al., 2009) this would therefore represent an important addition to the evidence-base.

A secondary aim of this study is to ascertain if there are differences between pass and fail groups in terms of performance on cognitive testing. Considering the US data from Fox (2011) highlighting that PVT failure invalidates expected brain-behaviour relationships, in addition to research indicating a strong relationship between test failure and reduction in overall neuropsychological test scores (for example Moss et al., 2003; Locke et al., 2008; Victor et al., 2009), establishing this finding in an NHS sample would further support the relevance of performance validity testing in UK clinical settings as an empirical means of assessing the validity of test data.

Hypotheses

It is hypothesised that a proportion of this NHS sample will fail multiple PVTs (including embedded and stand-alone measures). It is likely that this base rate will fall below 20%, as previous research in clinical, non-litigating populations

(Locke et al., 2008; Bunnage et al., 2008; Hampson et al., 2014) has looked at the prevalence of failure on individual PVTs as opposed to failure on multiple measures, which is a more conservative threshold. Given the evidence from existing studies applying this criterion, a base rate of around 10 to 15% might be anticipated. Base rates of failure on single PVTs in this sample will be provided for comparison.

It is further hypothesised that groups categorised on the basis of passing or failing PVTs will differ in terms of their performance on cognitive tests. Based on previous research (Moss et al., 2003; Locke et al., 2008; Victor et al., 2009; Fox, 2011), it is anticipated that the 'fail' group will exhibit reduced test scores in comparison to the 'pass' group. Performance will be compared on primary WAIS indices (Wechsler, 2008) IQ to explore difference in general intellectual functioning. Scores on additional battery measures such as Stroop Test (Golden, 1978), Verbal Fluency (from the Delis-Kaplan Executive Function System; Delis, Kaplan & Kramer, 2001) and Modified Card Sorting Test (Nelson, 1976) will be compared where these measures have been included as part of the flexible clinical assessment battery. It is hypothesised that the groups will not differ significantly on any demographic variables such as age, years of education, gender and injury severity.

Method

Settings

This study was approved by the London City and East NHS Research Ethics Committee and the local Research and Development Departments within the hosting NHS trusts (see Appendix 2). Data were gathered within two trusts, one in South London and the other in East London, both socioeconomically diverse areas

which was reflected in the patient population. The data were derived from three services to which the neuropsychology departments provide input and assessment: an inpatient neurorehabilitation unit; a community neurorehabilitation team and an outpatient service providing multidisciplinary neuropsychological assessment. Data from the first two services was gathered prospectively: all patients completed a 'flexible' battery of cognitive tests (where measures are selected to explore specific hypotheses rather than determined *a priori*; Bauer, 2014) with the inclusion of PVTs if not planned as part of the assessment. Data from the third service was retrospectively gathered from patient archives. All patients attending this latter service were tested using a standard (fixed) battery of neuropsychological measures, unless their presentation precluded the administration of specific tests. The data were anonymised on-site and patient details were not identifiable to the researcher. The archival data was shared with a fellow Trainee Clinical Psychologist who assisted with the completion of the database (see Appendix 1 for details). All analyses and write-up were conducted separately.

Participants

This study aimed to capture a patient sample which was representative of those accessing UK NHS adult neuropsychology services. As such, the inclusion criteria were kept purposefully broad to reflect the diversity of this population and consequently enhance the external validity of the study. All had a diagnosis of acquired brain injury corroborated by neurology reports, however, no parameters were set regarding type of injury or time since injury.

Patients were all over 18 years of age. Individuals with co-morbid mental or physical disorders were included where these disorders were not deemed to exert a

significant influence on testing (for example, individuals with acute psychiatric illness or visual impairment that would have precluded cognitive assessment were not included). A confirmed prior diagnosis of intellectual disability, degenerative neurological conditions (such as dementia or movement disorders) or functional neurological symptoms (such as conversion disorder) were used as a basis for exclusion from the study as these groups have been shown to score below cut-off more frequently on performance validity measures (Boone & Lu, 1999; Dean, Victor, Boone, & Arnold, 2008; Holdnack, Schoenberg, Lange, & Iverson, 2013; Davis & Millis, 2014; Meyers & Volbrecht, 2003). Patients who lacked capacity to consent to participation in the study (as assessed by their treating clinician) were also excluded.

Patients with identified external incentives were included. In this context, 'external incentives' encompassed a range of factors which might be relevant in UK clinical settings, such as disclosure of ongoing medicolegal proceedings or receipt of state benefit. Although it is acknowledged that the presence of external incentives is an important moderating factor in PVT performance (Belanger et al., 2005), it is likely that, given the service context, the majority of patients would be 'incentivised' to some degree. For example, most would be eligible to claim some form of government allowance (for example, free prescriptions or Disability Living Allowance), and all would be potential candidates for ongoing input from services. Excluding participants on the basis of identifiable incentive would therefore substantially reduce the applicability of the findings in clinical practice. The presence of external incentives was assessed via direct questioning of the client, examination of medical notes and / or confirmation with the individual's treating team.

Power analysis

Previous studies have investigated PVT failure in non-litigating samples on single (Locke et al., 2008; Bunnage et al., 2008; Hampson et al., 2014) and multiple measures (Meyers & Volbrecht, 2003; Kemp et al., 2008; Victor et al., 2009; Davis & Millis, 2014; Hall et al., 2014). For the base rate analysis it was assumed that the larger the sample, the greater the level of confidence in the base rate data. Power analysis for the group comparison was informed by prior work by Locke and colleagues (2008) based on available data from US non-litigating, treatment-seeking, acquired brain injury patients comparing TOMM pass and fail groups on tests of neuropsychological functioning. These authors found an average effect size of 0.98 (large effect; Cohen, 1992) for comparisons across tests in a neuropsychological battery (including WAIS-III, WMS-III, Stroop, Category Fluency and Wisconsin Card Sort Test). On the basis of this, power calculation was carried out using the "G*Power 3" computer program (Faul, Erdfelder, Lang, & Buchner, 2007), specifying alpha at 5% and desired power at 80%. The allocation ratio for participants in Group 1 (pass) versus Group 2 (fail) was set to 0.18 to account for a 15% estimated base rate of failure (i.e. N Group 1 = 85 / N Group 2 = 15). The required total sample size to detect significant group differences was estimated at 52 (Group 1 = 44, Group 2 = 8), which was felt to be within the resources of the current investigation. A study with this sample size would represent a valuable contribution to the field: a previous publication utilising different measures with a similar NHS population had comparable sample size of 47 participants (Hampson et al., 2014).

Recruitment procedure

Participants recruited prospectively were identified and approached to take part in the study by their treating clinicians. All participants had neuropsychological assessment planned as part of their clinical care. They were given the study

information sheet (see Appendix 3) at the time of their initial clinical interview (typically one week prior to the testing session) and then provided written consent (see Appendix 4) for their clinical data to be used for research prior to beginning cognitive testing. No incentives were provided to any of the participants. Assessment then took place as planned by the treating clinician, with the addition of PVTs if not already planned as part of the test battery. Written and oral feedback of test results was given to all patients as per usual service protocol. Demographic data from clinical interviews and medical records was also collected to provide information on age, gender, brain injury aetiology, time since injury, employment status and presence of external incentives.

Measures

All participants underwent comprehensive cognitive assessment which was sufficient to support clinical formulation (Suesse et al., 2015). As this was a naturalistic study, the cognitive test data gathered varied between participants. The measures chosen for analysis therefore reflect those which were available for the patients in this data set.

Performance validity measures

All participants completed the Test of Memory Malingering (TOMM; Tombaugh, 1996). Half of the participants also had data for an additional embedded PVT – the Digit Span Age-Corrected Scaled Score (DS-SS) or the Reliable Digit Span (RDS; Greiffenstein et al., 1994), both derived from the Wechsler Adult Intelligence Scale Digit Span subtest (Wechsler 1997; 2008). Digit span indices were selected as a) Digit Span was administered as standard to the majority of participants as part of the clinical assessment and b) it is a verbal measure, which contrasts to the

TOMM (a visual memory measure) thus meeting the stipulation of a 'multi-method, multi-test' approach to performance validity testing as recommended by the BPS (McMillan et al., 2009).

The TOMM is a standalone PVT which takes the form of a forced-choice visual recognition memory test. There are three trials, however, conventionally a score of less than 45 on Trial 2 is thought to be indicative of invalid performance (Tombaugh, 1996). At this cut-off, it has been found to correctly classify greater than 90% of neurologically impaired patients as 'not malingering' (traumatic brain injury, aphasia, cognitive impairment and dementia; Tombaugh, 1997), with a more recent review indicating a pooled sensitivity of 61% and specificity of 89% (Hall et al., 2014). The Retention trial was not administered for any patients, as in clinical practice time pressures frequently mean that abbreviated test procedures are utilised where possible and appropriate.

The Reliable Digit Span (RDS) was originally derived from the Digit Span subtest of the Wechsler Adult Intelligence Scale – Revised (WAIS-R; Wechsler, 1981): a measure of immediate memory span for auditory-verbal information (Iverson & Franzen, 1994). RDS is the sum of the longest forwards and backwards spans where both trials are correctly completed (Greiffenstein, et al., 1994). It has subsequently been applied to the Digit Span subtests of the WAIS-III and WAIS-IV (Wechsler, 1997 and 2008). A meta-analytic review of RDS validation studies (Schroeder et al., 2012) indicated that, at a cut-off of less than or equal to six, the RDS had a sensitivity rate of between 30 and 35% but a specificity of over 90% in most clinical groups studied, including moderate to severe traumatic brain injury. At a cut-off of less than or equal to seven global sensitivity was improved (58%) but specificity dropped to less than 90% across clinical groups. Given that, in clinical

practice, convention is to maximise specificity to reduce risk of false positives (Greve & Bianchini, 2004), a cut-off of less than or equal to six will be used for the purposes of this study.

The DS-SS is also derived from the Digit Span subtest of the WAIS-IV (Wechsler, 2008). Determination of the DS-SS follows standard administration and scoring procedures as laid out in the WAIS-IV manual (Wechsler, 2008). A cut-off of less than or equal to six has previously been used as an indicator of potential invalid performance, with a specificity of over 90% demonstrated in a sample of mTBI patients (Spencer et al., 2013). A more conservative cut-off of less than or equal to five was associated with a 5% increase in specificity and a concomitant 10% decrease in sensitivity. Given the heterogeneity of injury severity likely to be present within this sample, the more conservative less than or equal to five criterion was chosen as a threshold for classification. The DS-SS is less commonly used as an embedded PVT than Reliable Digit Span (RDS; Greiffenstein, et al., 1994; McCarter et al., 2009), however, it has equivalent classification accuracy (Jasinski, Berry, Shandera, & Clark, 2011; Young, Sawyer, Roper, & Baughman, 2012). DS-SS was used as the embedded measure for the archival data as RDS scores were not available.

Cognitive measures

To ascertain if there were differences on cognitive testing between individuals passing and failing PVTs, scores were gathered from the wider neuropsychological test battery. The primary measures used in the analysis were verbal and performance indices of the Wechsler Adult Intelligence Scale (WAIS), as these tests tap a number of cognitive skills to provide an estimate of global

intellectual functioning. The WAIS-III (Wechsler, 1997) and WAIS-IV (Wechsler, 2008) have been extensively standardised (WAIS-III *n* = 2,450; WAIS-IV *n* = 2,200) and shown to have high reliability. The majority of patients in the current sample had completed all subtests of the WAIS allowing for calculation of primaryindices and Full Scale IQ. Where shorter forms of the battery were utilised (for example, due to patient fatigue or time constraints), indices were pro-rated where possible as per the WAIS manual. Research has shown that even a two-subtest short form using only Vocabulary and Block Design produces good correlations with the Full Scale IQ of between 0.88 and 0.89, with longer short forms improving classification rates (Strauss, Sherman & Spreen, 2006, pp. 285-286).

Additional measures administered as part of the larger battery of tests included: Wechsler Memory Scale-IV Auditory and Verbal Memory (WMS-IV; Wechsler, 2009), Wechsler Memory Scale-III Verbal Working Memory, Immediate Memory (auditory and visual), Delayed Memory (auditory and visual) (WMS-III; Wechsler, 1997), Verbal Fluency (Delis, Kaplan, & Kramer, 2001), Modified Card Sorting Test (Nelson, 1976) and the Graded Naming Test (McKenna & Warrington, 1983).

Sample characterisation measures

Information on estimated premorbid functioning was gathered as part of sample characterisation. Both the Wechsler Test of Adult Reading (WTAR; Wechsler, 2001) and the Test of Premorbid Functioning (TOPF; Wechsler 2011) were used in this sample, as the TOPF superseded the WTAR following the publication of the WAIS-IV (the TOPF being co-normed with WAIS-IV, and WTAR with the WAIS-III; Brooks et al., 2011, p. 226). Both measures utilise an oral reading

paradigm that presumes the pronunciation of irregular words is relatively unaffected by neurological change (Brooks et al., 2011). They thus are statistical tools designed to assist determination of whether an individual's current test performance represents a decline from their previous level of ability (Whipple Drozdick, Holdnack, Weiss & Zhou, 2013. p 67). Information on affective status was also included in the sample characterisation. The majority of patients had completed the Personality Assessment Inventory (PAI) self-report measure, which includes clinical scales for anxious and depressive symptomatology (Morey, 1991; 2007).

Practical issues with the use of archival data

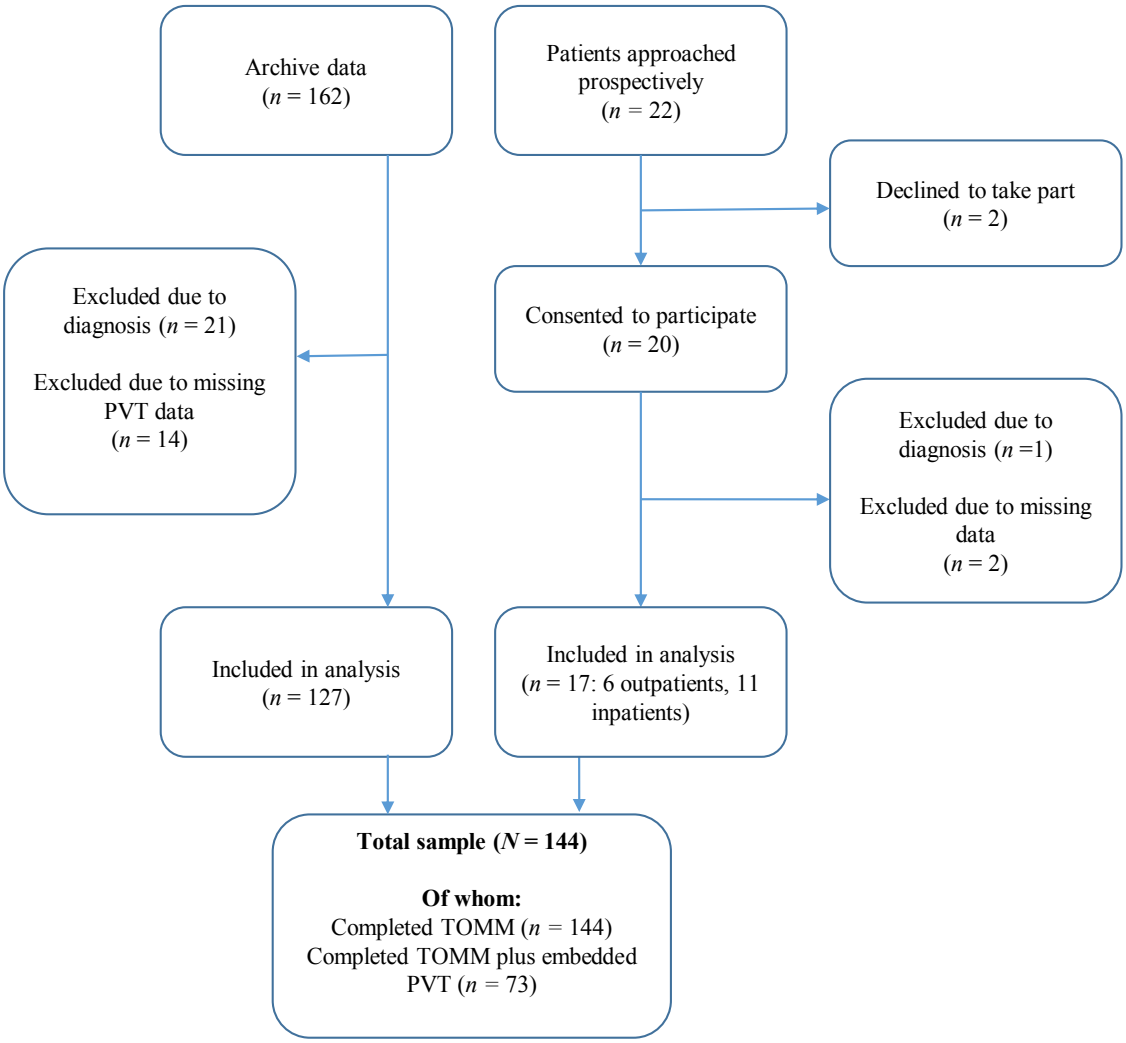
Different versions of tests were administered to patients before and after November 2011 (for example, WAIS-III versus WAIS-IV; WTAR versus TOPF), reflecting the availability of the most up-to-date test materials within the services. For the purposes of this study, full-scale IQ scores were analysed together across the different versions of the test. The potential lack of equivalence between the tests is acknowledged, as tests are refined over time to reflect updated conceptualisations of intelligence and re-standardised using new normative samples (Taub & Benson, 2013). There is also the issue of the Flynn Effect (the observed rise in IQ scores over time), which may mean that individuals tested later on older test versions have artificially inflated scores, as they are being compared to older norms (Flynn, 1984; Trahan, Stuebing, Fletcher, & Hiscock, 2014). Given that a) this analysis did not seek to compare between WAIS-III and WAIS-IV scores, and b) that versions have been shown to measure similar constructs (Taub & Benson, 2013) and have strong correlations between FSIQ indices (r=0.94; Holdnack, Schoenberg, Lange, & Iverson, 2013, p 218), the data were collapsed across groups but this was accepted as a potential limitation of the study.

Data analysis

A primary aim of this study was to provide information on the base rate of PVT failure in a naturalistic NHS clinical sample. The base rate (BR) of a condition of interest (in this case, below cut-off – or 'failed' - performance on PVTs) was calculated using the formula BR = number of cases with condition of interest / number of cases in the population (Gouvier, 1999). BRs were calculated for a) failure on the TOMM (as the most commonly-used standalone PVT), b) failure on any one PVT (TOMM or embedded PVT) and c) failure on two PVTs (TOMM and embedded PVTs). Pass and fail groups were analysed using SPSS 24 for Windows to examine differences in key demographic variables and current cognitive functioning, using WAIS Full Scale IQ as a measure of global ability. The PVT data had unequal sample sizes, hence non-parametric statistics were used throughout and a stringent criterion for significance was applied ($p<.01$).

As this is naturalistic patient data there is a significant degree of heterogeneity in the cognitive test scores available for each patient, as tests administered would be dependent on their presentation at assessment. The sample 'N' for each analysis is therefore indicated to clarify where data is missing. Figure 1 demonstrates how the final sample was derived from the prospectively gathered and archival data.

Figure 1*: Participant flow through the study*



Notes: PVT = performance validity test; TOMM = Test of Memory Malingering

Results

Sample demographics

A total of 144 sets of patient data were included in this study. This sample included males and females (66% male) with a broad range of acquired brain injuries. Traumatic brain injury accounted for 39% of the sample, other aetiologies included stroke / haemorrhage (35%), tumour (15%), infection / viral (2%), hypoxic injury (3%), epilepsy related (2%), encephalopathy (3%) and cysts (1%). Patients were a median 4 years post-injury, though there was a broad range (between 2 months and 45 years 4 months). Information on length of post-traumatic amnesia, duration of loss of consciousness or Glasgow Coma Score was not consistently available, therefore it was not possible to categorise TBI patients based on injury severity.

The sample ranged in age between 18 and 74 years (M = 44, SD = 14). The average estimated premorbid IQ was close to the general population mean (M = 103, SD = 12). Information on identifiable external incentives was available for 97% of the total sample (139 patients): 20% of the sample were found to have identifiable incentives, which included medicolegal claims, access to welfare benefits (such as Disability Living Allowance [DLA], Employment Support Allowance [ESA] or Personal Independence Payments [PIP]) and those in pursuit of medical retirement.

PVT performance: base rates

All patients (N = 144) completed the TOMM Trial 2. Fifteen patients scored below threshold on this measure (less than 45), giving a base rate of failure of 10% (M = 48.1, SD = 5.57, range: 21–50). Five patients scored below chance and 107 performed at ceiling. Scores on a second, embedded PVT were only available

for approximately half of the sample (n = 73). In this analysis, 'embedded' PVTs comprised either the DS-SS or RDS (with a cut-off of less than or equal to five for the DS-SS and less than or equal to six for RDS) subject to which measure was available for each patient. Of those who had data available for two PVTs, 15 patients scored below threshold on at least one (21%). Nine of these 15 patients failed on an embedded measure. Three patients failed two PVTs (4%). A summary of this information is provided in Table 1.

An exploratory analysis using a less-conservative cut-off of less than or equal to six on the DS-SS was also conducted, given previous research citing this as an appropriate threshold for the appraisal of performance validity in brain-injured populations (Spencer *et al*, 2013). Using this threshold, there was a base rate of 26% failure on one or more PVTs. Six patients failed two PVTs with the use of this cut-off (8%). Given the heterogeneity of this sample, and unknown severity of injury for the majority of patients, use of the less than or equal to five cut-off was deemed more appropriate for this analysis to reduce the likelihood of false positive errors.

Table 1: *Base rates of below cut-off performance on PVTs*

PVT variable	Sample N	Number of fails	Base rate of failure
Failed TOMM only	144	15	10%
Failed embedded PVT only	73	9	12%
Failed ≥1 PVT	73	15	21%
Failed 2 PVT	73	3	4%

Note: PVT = performance validity test; TOMM = Test of Memory Malingering

Demographic comparisons between PVT pass and fail groups

PVT pass and fail groups were compared to ascertain if there were any pre-existing demographic differences which may contribute to below-threshold performance. Analyses were conducted across two variables of interest: 1) individuals who passed or failed the TOMM only; 2) individuals who failed on one or more PVTs compared to those passing all PVTs. The two PVT fail group was not analysed separately due to the small sample size, however, further qualitative characterisation of this group is provided below to supplement this analysis (see below).

Non-parametric tests were used to account for uneven group size (chi squared test for categorical variables; Mann-Whitney U test for continuous variables) and a stringent criterion for significance was applied ($p<.01$). Demographic comparisons between the pass and fail groups indicated no significant differences in gender, age, time from injury to assessment, external incentives, employment status, estimated premorbid IQ and affective status (depression or anxiety). This information is summarised in Table 2.

Table 2: *PVT pass and fail groups: demographic comparisons*

		TOMM						≥1 PVT				
Variable	N	Pass group n	Fail group n	Pass	Fail	*p*	N	Pass group n	Fail group n	Pass	Fail	*p*
Gender	144	129	15	68% male	47% male	0.168	73	58	15	64% male	47% male	0.362
Age (years)	144	129	15	44 (18-74)	38 (27-65)	0.509	73	58	15	46.5 (21-74)	44 (27-65)	0.306
Time injury to ax (months)	140	126	14	13 (2-545)	35 (1-340)	0.025	72	57	15	9 (2-545)	23 (4-340)	0.031
External incentive	138	124	14	19% yes	29% yes	0.644	72	57	15	16% yes	20% yes	1.000
Employment	127	112	15	30% yes	27% yes	1.000	57	44	13	27% yes	23% yes	0.986
Estimated premorbid FSIQ	133	119	14	104 (64-132)	100 (63-115)	0.100	70	57	13	105 (64-132)	99 (63-121)	0.061
PAI anx	123	111	12	56 (8-93)	67.5 (37-86)	0.061	53	44	9	51.5 (8-81)	63 (37-81)	0.362
PAI dep	123	111	12	61 (5-105)	82 (32-95)	0.078	53	44	9	59 (5-101)	68 (32-98)	0.129

Notes: Median (range) scores are provided; **p<.01; ax = assessment; FSIQ = Full Scale IQ; PAI anx = Personality Assessment Inventory, anxiety subscale; PAI dep = Personality Assessment Inventory, depression subscale. Median and range scores are provided for the FSIQ and PAI variables. PAI *T* scores: <60 = no difficulty; 60-69 = mild to moderate difficulty; 70-82 = moderate difficulty; >82 = significant difficulty

Of note is that, although comparisons did not reach the threshold for significance, those failing PVTs had a consistently longer median duration between injury and the assessment (suggesting greater chronicity of brain injury-related difficulties) and scored higher on measures of affective status (suggesting a greater degree of symptomatology). Regarding affective status, individuals failing the TOMM were the only group to have median depressive symptom scores in the range for 'significant difficulty' as per the PAI scoring system (Morey, 1991; 2007). Scores across the groups for depression and anxiety were otherwise all in the mild to moderate range. Median premorbid FSIQ estimates of all groups fell within the 'Average' range as per the standard WAIS descriptors (Wechsler, 2008), with the full range of ability represented (range: 63 – 132).

PVT pass and fail groups: current cognitive functioning

Consistent with the second goal of this study, neuropsychological test results were compared between pass and fail groups across the variables of interest identified above. Median scores for WAIS-IV (FSIQ), Verbal Comprehension Index (VCI) and Perceptual Reasoning Index (PRI) are reported in Table 3 as primary measures of current cognitive function. Mann-Whitney U tests revealed that individuals failing PVTs (whether this was one or two fails) performed consistently worse on these measures than their counterparts who passed PVTs. As indicated in Table 3, the majority of analyses were significant at the $p<.01$ level. The differences in FSIQ between those passing all PVTs and those failing one or more PVTs remained significant at $p<.001$, with a large effect size noted ($U = 74$, $z = -4.04$, $r = -0.53$).

Table 3: *Comparison of median scores on primary cognitive measures across PVT groups*

		TOMM							≥1 PVT					
Variable	N	Pass group n	Fail group n	Pass group median score	Fail group median score	*p*	*r*	N	Pass group n	Fail group n	Pass group median score	Fail group median score	*p*	*r*
FSIQ	119	105	14	97	89	0.005**	-0.26	57	44	13	97	71	<.001***	-0.53
VIQ/VCI	67	59	8	98	73	0.030*	-0.26	61	48	13	101	91	0.001**	-0.41
PIQ/PRI	75	67	8	98	78	0.001**	-0.37	67	53	14	102	82.5	0.001**	-0.40

Notes: *p<.05, **p<.01, ***p<.001; FSIQ = WAIS Full-scale IQ; VIQ/VCI = WAIS Verbal IQ/Verbal Comprehension Index; PIQ/PRI = WAIS Performance IQ/Perceptual Reasoning Index.

To further explore differences in overall cognitive functioning between individuals failing 0, 1 or 2 PVTs, FSIQ scores were compared between the three groups. A Kruskal-Wallis Test revealed a statistically significant difference across the three groups (0 fail, $n = 44$; 1 fail, $n = 10$; 2 fails, $n = 3$; chi-square [2, 57] = 17.9, $p<.001$). Post-hoc testing to explore significant comparisons indicated that the FSIQ score of the 0 fail group ($Md = 97$) was significantly higher than that of both the 1 fail ($Md = 80.5$, $U = 74$, $z = -3.25$, $p = 0.001$, $r = -0.44$) and 2 fails ($Md = 51$, $U = .000$, $z = -2.88$, $p = 0.004$, $r = -0.42$). The 1 and 2 fail groups were significantly different at the $p<.05$ level ($U = .000$, $z = -2.54$, $p = .011$). These results imply that the strong effect sizes for the ≥1 PVT fail group demonstrated in the FSIQ comparison above are not driven purely by the presence of '2 PVT fail' individuals within that group, but that the failure of a single PVT alone is associated with significantly poorer scores on cognitive testing than if all PVTs are passed.

Examination of the median FSIQ scores for each group would suggest that failing more PVTs is associated with decreasing global cognitive ability. Although data from the '2 PVT fail' group should be viewed as exploratory due to small sample sizes, all patients in this group had lower scores on this measure than the lowest-scoring patients in the '0 fail' and '1 fail' groups, with scores falling into the 'Extremely Low' range (that is, lower than 98% of the general population). Only two of the three patients in this group had premorbid IQ estimates available, however, both fell in the 'Average' range (102 and 92), suggesting that they did not have low baseline FSIQ scores relative to others in the sample which might account for the findings. For comparison, the median FSIQ score of the '1 fail' group fell within the

'Low Average' range (*Md* = 80.5, range = 68-96), compared to estimated premorbid scores also in the 'Average' range (*Md* = 99, range: 63–121).

Comparing pass and fail groups on performance in specific cognitive domains

To explore whether PVT failure was associated with reduced performance in specific cognitive domains, comparisons were made between those individuals passing all PVTs and those failing one or more PVTs across a battery of neuropsychological tests. Data were missing for some variables, however, the minimum total *n* was 47 patients (minimum fail group *n* = 8). Failure on one or more PVTs was strongly associated with lower scores across the Working Memory, Processing Speed and General Ability indices of the WAIS-IV; the Verbal Working Memory and Visual Recognition subscales of the WMS-III; Category Fluency, the Graded Naming Test and Stroop test (see Table 4).

The analysis was repeated for individuals passing and failing the TOMM only (see Table 5). Due to variations in test batteries between patients there were fewer neuropsychological measures consistently available for comparison, however, there were larger sample sizes: total *n*'s ranged from 98 (Stroop) to 121 (WAIS PSI), with fail group *n*'s ranging from 10 (MCST) to 14 (WAIS indices). Results echoed the previous analysis: TOMM pass / fail groups differed significantly on all WAIS indices bar VCI, on the GNT and on the Stroop. Effect sizes, however, were all small to medium as opposed to medium to large.

Table 4: *Cognitive test scores for individuals passing and failing ≥1 PVT*

Test variable	N	Pass group n	Fail group n	Pass *Md* (range)	Fail *Md* (range)	*p*	*r*
WAIS-IV							
WMI	56	43	13	97 (74-135)	77 (60-100)	<.001***	-0.54
PSI	55	43	12	92 (65-146)	72.5 (50-94)	<.001***	-0.50
GAI	56	43	13	101 (69-144)	81 (52-105)	<.001***	-0.48
WMS-IV							
AM	53	43	10	97 (51-124)	84.5 (55-110)	.232	-0.16
VM	52	43	9	92 (63-138)	85 (76-87)	.013*	-0.34
WMS-III							
VWM	52	42	10	89.5 (69-126)	75 (63-88)	.001**	-0.46
IM	52	43	9	96 (65-129)	81 (61-98)	.086	-0.24
DM	51	42	9	92 (56-130)	82 (62-100)	.092	-0.24
AI	51	41	10	94 (42-123)	86 (56-108)	.301	-0.14
AD	51	41	10	92 (48-127)	83 (52-144)	.454	-0.10
AR	51	41	10	91 (45-118)	85.5 (49-98)	.058	-0.27
VI	47	38	9	92 (50-127)	80 (74-92)	.056	-0.28
VD	47	38	9	92 (63-140)	85 (75-95)	.085	-0.25
VR	46	37	9	90 (57-120)	80 (72-87)	.008*	-0.39
FAS							
Letter	53	43	10	8 (1-19)	6.5 (3-13)	.218	-0.17
Category	53	43	10	9 (2-16)	6 (1-10)	.001**	-0.46
Switch correct	53	43	10	9 (1-15)	7 (1-12)	.059	-0.26
Switch accuracy	53	43	10	10 (1-31)	8.5 (2-11)	.069	-0.25
MCST							
Categories	49	42	7	6 (1-6)	5 (1-6)	.135	-0.21
% perseverative errors	49	42	7	27.5 (0-100)	40 (33-68)	.169	-0.20
GNT (raw)	51	41	10	21 (1-28)	7 (4-17)	<.001***	-0.50
Stroop (raw)	40	29	11	85 (10-112)	70 (17-93)	.028*	-0.37

Notes: *p<.05, **p<.01, ***p<.001. WAIS-IV = Wechsler Adult Intelligence Test-IV (Wechsler, 2008); WMI = Working Memory Index; PSI = Processing Speed Index; GAI = General Ability Index; WMS-IV = Wechsler Memory Scale-IV (Wechsler, 2009); AM = Auditory Memory Index; VM = Verbal Memory Index; WMS-III = Wechsler Memory Scale-III (Weschler, 1997b); VWM = Verbal Working Memory; IM = Immediate Memory; DM = Delayed Memory; AI = Auditory Immediate; AD = Auditory Delayed; AR = Auditory Recognition; VI = Visual Immediate; VD = Visual Delayed; VR = Visual Recognition; FAS = DKEFS verbal fluency test (Delis, Kaplan, & Kramer, 2001); MCST = Modified Card Sorting Test (Nelson, 1976); GNT = Graded Naming Test (McKenna & Warrington, 1983).

Table 5: *Cognitive test scores for individuals passing and failing the TOMM only*

Test variable	N	Pass group n	Fail group n	Pass group *Md* (range)	Fail group *Md* (range)	*p*	*r*
WAIS-III / IV							
PSI	121	107	14	89.5 (56-146)	79 (50-144)	.002**	-0.29
WMI	120	106	14	97 (63-144)	87.5 (60-119)	.039*	-0.19
MCST							
Categories	108	98	10	6 (1-6)	5 (3-6)	.315	-0.1
% perseverative errors	108	98	10	22.5 (0-100)	21 (0-68)	.977	-0.002
GNT (raw)	108	96	12	20.5 (1-29)	15 (4-24)	.03*	-0.21
Stroop (raw)	98	87	11	87 (9-112)	66 (17-96)	.012*	-0.25

Note: *p<.05, **p<.01, ***p<.001. WAIS-III/IV = Wechsler Adult Intelligence Scale-III/IV (Wechsler, 1997b; 2008); WMI = Working Memory Index; PSI = Processing Speed Index; MCST = Modified Card Sorting Test (Nelson, 1976); GNT = Graded Naming Test (McKenna & Warrington, 1983); Stroop = Stroop Test (Golden, 1978).

Clinical characterisation of patients failing two PVTs

Qualitative analysis of the demographics and neuropsychological test performance of the three patients failing two PVTs was conducted to characterise the group and ascertain if there were clinical features which may have influenced performance.

Patient A had a diagnosis of severe TBI eight months prior to the assessment. Co-morbid depression was reported. An atypical pattern of dense retrograde amnesia was observed on self-report. On testing it was noted that auditory and visual recognition scores on the WMS were reduced in comparison to auditory and visual delayed free-recall (delayed scores both in Low Average range), which again is atypical. No premorbid IQ estimates were available, however, current intellectual functioning fell in the Extremely Low range (WAIS-IV FSIQ standard

score = 51, score exceeding less than 0.1% of peers). This patient was in the process of applying for DLA at the time of assessment. TOMM Trial 2 score was 38; DS-SS was 3.

Patient B had experienced a moderate TBI due to assault three years prior to the assessment. It was noted that the assessment was challenging due to low confidence on testing and the fact that English was a second language; as such the majority of the battery was not completed. A number of significant social welfare issues were reported. Estimated premorbid IQ was estimated to be in the average range. Current intellectual functioning was found to be in the Extremely Low range (WAIS-IV FSIQ standard score = 49, score exceeding less than 0.1% of peers). TOMM Trial 2 score was 27; DS-SS was 2.

Patient C had undergone neurosurgery to remove a brain tumour in childhood and subsequently developed cognitive difficulties and epilepsy. Co-morbid psychiatric symptomatology was present which likely impacted engagement on tasks. An 'apathetic' approach to testing was observed. Premorbid IQ estimates were in the average range, current intellectual functioning was in the Extremely Low range (WAIS-IV FSIQ standard score = 65, score exceeding less than 1% of peers). No external incentives were noted. TOMM Trial 2 score was 41; DS-SS was 4.

Discussion

This study aimed to establish the base rate of PVT failure in a post-acute NHS acquired brain injury sample, in addition to understanding whether failing PVTs was associated with reduced scores on cognitive assessment. Limited data exist regarding the performance of NHS patient samples on PVTs, despite professional bodies recommending that they be used routinely in clinical evaluations (McMillan et al., 2009). As such, garnering data from clinical settings is essential to increase understanding in this area and thus assist clinicians to make more accurate judgements regarding the nature of individual deficits.

Base rates of PVT failure

In this sample, there was a base rate of 10% failure on the TOMM; 21% failure on at least one PVT when two were administered (the TOMM plus one Digit Span-derived embedded PVT) and 4% failure on two PVTs. To the author's knowledge, this is the first UK clinical study of this size to establish base rates of failure on these measures. Previous research from the US using a comparable clinical sample found a base rate of 22% failure on the TOMM (Locke et al., 2008). The finding of only 4% of the sample failing two PVTs as per the 'multi-method, multi-test' approach advocated by professional practice guidelines (McMillan et al., 2009; Heilbronner et al., 2009) is lower than originally hypothesised. It is similar to Victor et al.'s (2009) results demonstrating a 6% base rate in 'credible' patients, but substantially less than other studies using non-litigating clinical samples (for example, 15% in Davis & Millis, 2014), including studies conducted in the UK (for example, 11% in Kemp et al., 2008).

The relationship between PVT failure and neuropsychological test performance

Where individuals with confirmed brain injury failed one or more PVTs, they were more likely to have lower scores on global measures of current cognitive functioning than those passing PVTs. This was also the case on specific tests tapping verbal working memory, visual memory, semantic fluency, object naming and processing speed. This effect was found in the context of pass and fail groups being broadly matched on all other demographic variables including estimated premorbid IQ. Effect sizes for the significant comparisons were moderate to large (as per Cohen, 1992), indicating that these group differences would be substantial enough so as to be clinically distinguishable (Bigler, 2014).

The use of PVTs is predicated on the assumption that 'passing' the tests places such negligible demands on cognition that they can be considered impervious to all but the most severe forms of central nervous system dysfunction. They are therefore purported to provide clinicians with additional information on non-neurological dimensions of performance. If this were the case, one might anticipate that performing below published cut-offs on these measures would lead to a pattern of generally compromised scores across the neuropsychological test battery, as any superordinate factors would presumably impact on all measures to some degree. The findings of this study suggest that group differences between people passing and failing PVTs do not occur in a uniform pattern across test batteries but appear to differentially impact specific tests. As such, an alternative hypothesis may be that PVTs are picking up on aspects of individual organic deficit in addition to – or perhaps instead of - the 'non-neurological' factors they were designed to assess.

The evidence-base is equivocal on the matter of whether PVT-predicted downgrading of neuropsychological performance can be related to individual cognitive impairment. The results of the current study echo those demonstrated by Locke et al. (2008), where inconsistent performance across a neuropsychological battery was also demonstrated in individuals passing and failing the TOMM. The authors hypothesised that this effect was, however, unlikely to be secondary to organic factors as failure rates were unrelated to proximal variables of cognitive impairment severity (such as injury severity, employment status and disability status). Conversely, studies investigating other forced-choice PVTs in clinical samples (Hall et al., 2014, Keary et al., 2013) have attributed poor performance of 'fail' groups to the cognitive demands of the PVT (for example verbal processing and working memory for the Word Memory Test and Victoria Symptom Validity Test respectively), which would imply a primary influence of organic factors. In the dementia literature, decreasing cognitive functioning has been correlated with poorer performance on the TOMM, which one study found to be one of the PVTs most sensitive to the severity of cognitive impairment (Rudman, Oyebode, Jones, & Bentham, 2011).

In cognitive neuroscience, functional Magnetic Resonance Imaging (fMRI) research has shown that any task over and above primary sensory stimulation will require 'effort' and therefore will engage cognitive processes, regardless of how trivial the task may objectively appear (see Bigler, 2014, for a review). Given the nature of the PVTs used in this study, one could assume that a degree of attentional and working memory capability would be required to attempt them. It has been established that the frontotemporal and limbic regions thought to underpin these functions have the greatest propensity for damage in TBI (Cowell, Bussey &

Saksida, 2006; Allen, Bigler, Larsen, Goodrich-Hunsaker, & Hopkins, 2007; West, Curtis, Greve, & Bianchini, 2010), which potentially lends further credence to the 'organic cause for PVT failure' hypothesis. Further research is needed, however, to establish a) whether performance on different PVTs varies with condition or lesion location and b) if it is possible to clearly delineate which neuropsychological measures are likely to be suppressed in relation to different PVTs.

With respect to point (b), the results of this study would suggest that failing the TOMM or Digit Span-derived PVTs is not associated with worse performance on the Modified Card Sorting Test (MCST; Nelson, 1976) or phonemic fluency (DKEFS; Delis, Kaplan, & Kramer, 2001), both of which are considered tests of higher-level executive functions which are sensitive to brain injury. Similarly, the two measures on which group differences were not demonstrated in the Locke et al. (2008) paper were the Wisconsin Card Sort Test (WCST; Grant & Berg, 1948; Heaton, 1981; Heaton, Chelune, Talley, Kay, & Curtiss, 1993) and the Category Test (Russell & Levy, 1987), which are again executive measures assessing complex skills such as concept formation, abstraction and cognitive flexibility. Examination of the median MCST scores in the current study (see Table 4) indicate that individuals in the pass group are performing overall in the average range (greater than the 45th per centile for categories; 25th-30th per centile for perseverative errors) compared to the low average to average range for the fail group (35th – 40th per centile for categories; 10th-15th for perseverative errors; norms from Obonsawin et al., 1999). One would anticipate that tests exerting greater demands on executive functions would be among those most vulnerable to both brain injury and any non-neurological factors influencing performance, but these data are not consistent with this. It is possible that, given small sample sizes for fail groups, this represents a

spurious finding: further research is therefore needed to explore the associations of specific tests with PVT performance, specifically those assessing higher-level executive functions most commonly impaired in brain injury.

Differences between the TOMM and embedded measures

The PVTs investigated in this study were chosen as together they met the demands of the 'multi-test, multi-method' approach advocated in professional guidelines for the use of PVTs (McMillan et al., 2009; Heilbronner et al., 2009). It was not a primary aim of the study to compare their relative utility, however, the data raised some pertinent issues related to their use in clinical populations. Firstly, larger effect sizes were demonstrated when the analyses combined Digit Span PVT and TOMM fails (the 'one or more PVT fails' group) than when TOMM fails were analysed separately (see Tables 4 and 5). This suggests that inclusion of Digit Span PVT fails strengthens the association between PVT failure and underperformance on specific neuropsychological tests. Given the arguments presented above, is it possible that Digit Span embedded PVTs are more likely to pick up on organic deficits than the TOMM? As highlighted by Bigler (2012), using embedded measures which were not explicitly designed to assess validity makes the issue of 'disentangling' true neuropsychological deficits from associated non-neurological elements all the more complex. It seems parsimonious to infer that where brain injured individuals are vulnerable to disruption in neural networks underpinning the cognitive functions needed to do the Digit Span task (that is, auditory attention and working memory), performance will be suppressed on both the task and the PVT related to the task. Indeed, examination of median scores on the Working Memory Index (of which the Digit Span test is a primary subtest) indicated a 10 point difference in standard scores between people who failed any one PVT (*Md* = 77;

range: 60-100) versus those who just failed the TOMM (*Md* = 87.5; range: 60-119). Even though cut-offs have been validated in brain injured populations (Spencer et al., 2013), these results suggest that caution should be exercised when interpreting patient performance on embedded validity measures and supports the assertion that these indices should not be used in the absence of standalone PVTs (Miele et al., 2012).

Impact of the number of PVTs failed

Previous research has indicated that a criterion of two or more PVT fails should be used to indicate invalid performance (see Larrabee, 2014 for a discussion of this issue). The current study demonstrated that, in clinical practice, very few patients will fail two PVTs but a substantial minority will fail at least one. These results are consistent with the literature, and again underscore the need for caution when interpreting the output of PVTs in clinical settings. Whilst it is possible that individuals without cognitive deficits may achieve 100% specificity on PVTs, it is unlikely that *bona fide* patients with neurological conditions will perform at this level. The majority of PVTs were developed with medicolegal populations in mind, where the aim was to identify intentionally feigned symptoms. The tests have since been adopted by clinical practitioners as a means of identifying amplified symptoms in the context of established clinical syndrome, which is likely to place far greater demands on the instrument's predictive validity (Merten et al., 2007). Clinicians must therefore adjust their conclusions accordingly whilst bearing in mind that within their pool of patients failing one PVT there will likely be a mix of true-valid and true-invalid performers (Martin et al., 2016). A number of authors assert that qualitative analysis of PVT fail scores is necessary to ameliorate this issue. Rigid application of cut-offs in clinical groups risks imposing artificial valid-invalid

dichotomies (Willis, Farrer, & Bigler, 2011), and it may be that considering individual performance on a spectrum from 'likely valid' to 'likely invalid' would be more appropriate. Of the 15 individuals in this study who performed below the published cut-off on the TOMM Trial 2, seven scored over 40. It is with this 'near pass' group that clinicians are at greatest risk of making false positive errors (Bigler, 2012), and further guidance on whether cut-off scores should be variable depending on the characteristics of specific patient subgroups is therefore warranted.

By contrast, the evidence base suggests that individuals failing two or more PVTs likely constitute 'true invalid' test-takers, even in clinical groups (Larrabee, 2014; Martin et al., 2016). Qualitative analysis of the three patients in the two-PVT fail group in this study brought to light a number of factors which may pose threats to valid performance. For example, developmental neurological compromise, co-morbid psychiatric diagnosis, borderline IQ and English as a second language have all been linked to failure on PVTs (Salazar, Lu, Wen, & Boone, 2007; Dean et al., 2008; Victor et al., 2009). In addition, two of the three were observed to have overt difficulties engaging with the assessment, and the third had an atypical amnestic presentation both in self report and on objective memory measures. Whilst it is not possible to discern from this limited information whether these represent 'true invalid' assessments, there appears to be reasonable convergence from PVTs, self-reports, behavioural observations and patterns of neuropsychological scores to query if the assessment results are fully representative of individual cognitive ability.

Methodological critiques and recommendations for future research

In the British Psychological Society (BPS) guidelines on the assessment of effort (McMillan et al., 2009), the need for further evidence on UK base rates of

cognitive impairment and PVT performance was identified. However, only a handful of studies have been published to meet this research goal (including Hall et al., 2013; Hampson et al., 2014; Suesse et al., 2015). This is the first study to investigate performance of a clinical acquired brain injury group on the TOMM, the PVT most commonly used in UK clinical neuropsychology practice (McCarter et al., 2009). The fact that this study has a comparatively large total sample size relative to previous studies in UK clinical populations (see Kemp et al., 2008; Hampson et al., 2014 and Hall et al., 2014 which had samples of 43, 47 and 48 patients respectively) suggests that some confidence can be placed in the base rate data. The sample comprised participants with a wide range of neurological diagnoses of varied severity which are representative of the referrals received by NHS acquired brain injury services. No systematic biases in demographic variables were identified, which enhances the ecological validity of this research. Given these points, it is hoped that these data are strongly applicable to the day-to-day practice of NHS clinical psychologists and neuropsychologists.

There are, however, a number of potential limitations to this study. Methodologically, although the diagnostic heterogeneity represents a strength of this study, there was no available means of quantitatively comparing the severity of injury or condition across the groups. This would have been a beneficial addition to the analysis as an index of likely cognitive impairment, as the 'dose-response' relationship between injury severity and neuropsychological outcome has been previously established (Dikmen, Machamer, Winn, & Temkin, 1995). Qualitative comparison of diagnoses between the pass and fail 'one or more' PVT groups demonstrated similar proportions of TBI patients (28% versus 33%) but a greater proportion of CVA/stroke patients in the pass group (43%) and a greater proportion

of tumour patients in the fail group (40%). Specific analysis of how individuals with different diagnoses can be expected to perform on the TOMM and embedded PVTs would be valuable, however, the small sample sizes of subgroups in this analysis precluded this.

As a general critique, all analyses had uneven sample sizes and correspondingly some samples for fail groups were small: this is a challenge within the PVT research as a whole, and was offset in this study as far as possible by the use of non-parametric statistics and more conservative criteria for significance. Nonetheless, this is a caveat to bear in mind when interpreting the data from the group comparisons, which should be viewed as preliminary.

A further criticism rests with the choice of PVTs. The professional guidelines advocate a 'multi-test, multi-method' approach which is useful as, in practice, we do not know how expressions of invalid performance might manifest across the course of an assessment battery or over testing in different cognitive domains. However, guidelines do not specify which combinations of tests will provide the greatest degree of classification accuracy. The use of multiple measures will only provide additional benefit if the results of each classification are independent from each other (Rosenfeld et al., 2000). More research establishing the inter-correlations between PVTs is therefore necessary to clarify which combinations of tests offer the greatest classification accuracy in clinical populations.

Regarding the study sample, the inclusion of individuals with identified external incentives may be a criticism of this study. Approximately 20% of the sample were known to have identifiable incentives, some of which included medicolegal claims. It has previously been established that individuals with

incentives, and particularly those with ongoing litigation proceedings, are more likely to fail PVTs (Bianchini, Curtis, & Greve, 2006). For this reason, studies using clinical samples frequently exclude patients on the basis of identified incentives. The decision to include incentivised patients in this study was intentional as the aim was to characterise a typical population presenting to NHS brain injury services, which will inevitably include a proportion of individuals with incentives, and it is of note that the proportion of patients with identified incentives did not differ significantly between pass and fail groups. It is also recognised that categorising individuals on the basis of incentive is somewhat arbitrary, as the true extent that someone is incentivised to do well on an assessment is essentially unknowable. Many patients in this sample would, for example, be eligible for ongoing treatment. One could also consider a range of psychological mechanisms, both conscious and unconscious, which may function to 'incentivise' the patient to perform in a specific way, for example a desire to be validated for perceived difficulties, or as a 'cry for help' (Locke et al., 2008).

In relation to this, one of the central critiques of this study, and indeed of the literature as a whole, is the lack of clarity regarding the source of the PVT failure. Whilst there are a number of arguments which would link the PVT fails in this study to potential organic deficit, based on these data there is no way of precisely delineating the involvement of factors over and above cognitive impairment. Mood factors were not explored in this analysis, however, it is possible that psychological variables are an important source of variance in the neuropsychological test data. For instance, it is noteworthy that, whilst comparisons did not reach statistical significance, individuals failing the TOMM had higher self-reported depressive symptomatology than any other group (PAI depression scores in the 'significant'

range). Studies have demonstrated that depressed patients who pass PVTs do not have suppressed scores on neuropsychological testing (Rohling, Green, Allen, & Iverson, 2001), but depressed patients who fail PVTs do demonstrate reduced scores on test batteries (Green et al., 2001). Thus there appears to be a meaningful interaction between depressive symptomatology and PVT failure which may provide a further context in which to understand the results of this study.

Concerning psychological variables, the questionnaire measures of trait anxiety and depression would not have identified test-related issues such as performance anxiety, which are more typically observed by examiners during testing. Working memory has been found to be down-regulated in anxiety (Ikeda, Iwanaga, & Seiwa, 1996; Robinson, Vytal, Cornwell, & Grillon, 2013), therefore there is an argument that reduced Digit Span scores, and by association the PVTs derived from them, may be a marker of in-the-moment affective state. One might query whether a degree of neurological compromise would also increase susceptibility to the impact of anxiety and lead to more pronounced underperformance (Waldstein, Ryan, Jennings, Muldoon, & Manuck, 1997), thus the pattern of results might be indicative of a combination of organic and psychological factors. The use of symptom self-report measures with inbuilt validity scales (such as the Personality Assessment Inventory; Morey, 1991; 2007) alongside PVTs could go some way to detecting the influence of psychogenic variables on PVT performance and would therefore be a useful focus for future research.

Clinical utility of the findings

This research highlights a number of practical and conceptual issues which are pertinent to clinical neuropsychological practice within NHS acquired brain

injury services. The study demonstrated that a significant minority of individuals with established diagnoses of brain injury will fail the TOMM and/or Digit Span-derived PVTs using established cut-offs, but only a very small percentage will fail two PVTs. This knowledge could be a useful heuristic in clinical practice even though the data cannot conclusively state the number of 'true invalid' performers identified. PVT failure could be viewed as a 'red flag' indicating that low scores on cognitive testing may not be directly attributable to a condition of the brain. However, assessments should not automatically be judged as 'invalid' on the basis of failed PVTs. Firstly, cognitive impairment could still be a potential explanation for PVT underperformance, and clinicians should examine whether there is convergent evidence of impairment on measures which tap cognitive domains overlapping with PVTs (for example visual recognition memory for the TOMM; verbal attention and working memory for the Digit Span measures) as this could indicate 'risk factors' for PVT failure. With this in mind, the use of embedded Digit Span validity measures in isolation is not recommended in a clinical setting. Results of these measures should be interpreted with caution in acquired brain injury groups as the nature of their condition means they are vulnerable to working memory deficits.

Secondly, although PVT failure may reduce confidence in the objective findings of neuropsychological test data, it should serve to stimulate further enquiry into the clinical facts of the case. Little is known about the psychological mechanisms underpinning the experience of validity test failure on a group level, however, using PVT performance as part of the broader clinical formulation could guide a more nuanced understanding of the patient's needs on a case-by-case basis.

This research has highlighted the challenges inherent in attempting to utilise PVTs in NHS clinical practice. Using PVTs will not enable clinicians to draw

definitive conclusions about performance validity, and indeed to do so would pose a significant risk of making false positive errors. There are a number of areas of research need which must be clarified before the clinical community can start applying PVTs and their associated cut-off scores with confidence. Until then, it is recommended that they are utilised with the aforementioned caveats in mind and always in the context of comprehensive assessment of potential biopsychosocial influences on performance.

References

Allen, M. D., Bigler, E. D., Larsen, J., Goodrich-Hunsaker, N. J., & Hopkins, R. O. (2007). Functional neuroimaging evidence for high cognitive effort on the Word Memory Test in the absence of external incentives. *Brain Injury*, *21*(13-14), 1425-1428.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Ashendorf, L., Constantinou, M., & McCaffrey, R. J. (2004). The effect of depression and anxiety on the TOMM in community-dwelling older adults. *Archives of Clinical Neuropsychology, 19*(1), 125-130.

Bauer, R. M. (2014). The flexible battery approach to neuropsychological assessment. In Vanderploeg, R. D. (Ed.) *Clinician's guide to neuropsychological assessment* (pp. 419-449). New York, NY: Routledge.

Belanger, H. G., Curtiss, G., Demery, J. A., Lebowitz, B. K., & Vanderploeg, R. D. (2005). Factors moderating neuropsychological outcomes following mild traumatic brain injury: A meta-analysis. *Journal of the International Neuropsychological Society, 11*(03), 215-227.

Bianchini, K. J., Curtis, K. L., & Greve, K. W. (2006). Compensation and malingering in traumatic brain injury: a dose-response relationship? *The Clinical Neuropsychologist*, *20*(4), 831-847.

Bigler, E. D. (2012). Symptom validity testing, effort, and neuropsychological assessment. *Journal of the International Neuropsychological Society*, *18*(04), 632-640.

Bigler, E. D. (2014). Effort, symptom validity testing, performance validity testing and traumatic brain injury. *Brain injury*, *28*(13-14), 1623-1638.

Boone, K. B., & Lu, P. H. (1999). Impact of somatoform symptomatology on credibility of cognitive performance. *Clinical Neuropsychologist, 13*(4), 414-419.

Boone, K. B., Lu, P., Back, C., King, C., Lee, A., Philpott, L. & Warner-Chacon, K. (2002). Sensitivity and specificity of the Rey Dot Counting Test in patients with suspect effort and various clinical samples. *Archives of Clinical Neuropsychology*, *17*(7), 625-642.

Boone, K. B. (2007) A reconsideration of the Slick et al. (1999) criteria for malingered neurocognitive dysfunction. In Boone, K. B. (Ed.) *Assessment of feigned cognitive impairment: A neuropsychological perspective*. (pp. 29-50). Guilford Press.

Brooks, B. L., Holdnack, J. A., & Iverson, G. L. (2011). Advanced Clinical Interpretation of the WAIS-IV and WMS-IV: Prevalence of Low Scores Varies by Level of Intelligence and Years of Education. *Assessment, 18*(2), 156-167.

Bunnage, M., Eichinger, C., Pearce, N., Duckworth, A. & Newson, M. (2008). Criterion validity of the Word Memory Test: An audit of a sample of patients assessed for clinical, not litigious, reasons. [Proceedings of the 36th Annual Meeting of International Neuropsychological Society, Hawaii, February 2008 Abstract]. *Journal of International Neuropsychological Society, 14*(Suppl. 1), 138-139.

Bush, S. S., Ruff, R. M., Tröster, A. I., Barth, J. T., Koffler, S. P., Pliskin, N. H., Reynolds, C. R. & Silver, C. H. (2005). Symptom validity assessment: Practice issues and medical necessity: NAN Policy & Planning Committee. *Archives of Clinical Neuropsychology*, *20*(4), 419-426.

Cohen, J. (1992). A Power Primer. *Psychological Bulletin, 112*(1), 155-159

Cowell, R. A., Bussey, T. J., & Saksida, L. M. (2006). Why does brain damage impair memory? A connectionist model of object recognition memory in perirhinal cortex. *Journal of Neuroscience*, *26*(47), 12186-12197.

Davis, J. J., & Millis, S. R. (2014). Examination of Performance Validity Test Failure in Relation to Number of Tests Administered. *Clinical Neuropsychologist, 28*(2), 199-214.

Dawes, R. M., Faust, D., Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, volume 243(4899), 1668-1674.

Dean, A. C., Victor, T. L., Boone, K. B., & Arnold, G. (2008). The relationship of IQ to effort test performance. *The Clinical Neuropsychologist*, *22*(4), 705-722.

Delis, D. C., Kaplan, E., & Kramer, J. H. (2001). *Delis-Kaplan Executive Function System.* San Antonio, TX: The Psychological Corporation.

Dikmen, S. S., Machamer, J. E., Winn, H. R., & Temkin, N. R. (1995). Neuropsychological outcome at 1-year post head injury. *Neuropsychology*, *9*(1), 80.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis for the social, behavioural and biomedical sciences. *Behaviour Research Methods*, 39, 175-191

Faust, D., Hart, K., Guilmette, T. J., & Arkes, H. R. (1988). Neuropsychologists' capacity to detect adolescent malingerers. *Professional Psychology-Research and Practice, 19*(5), 508-515.

Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological bulletin*, *95*(1), 29.

Fox, D. D. (1990; 2009). *Manual for the Computerized Test of Attention and Memory* (8th ed.). Glendale, CA.

Fox, D. D. (2011). Symptom Validity Test Failure Indicates Invalidity of Neuropsychological Tests. *Clinical Neuropsychologist, 25*(3), 488-495.

Frederick, R. I., & Speed, F. M. (2007). On the interpretation of below-chance responding in forced-choice tests. *Assessment, 14*(1), 3-11.

Golden, C. J. (1978). *Stroop Colour and Word Test: a manual for clinical and experimental uses*. Chicago, IL: Stoelting Co.

Gouvier, W. D. (1999). Base rates and clinical decision making in neuropsychology. In J. J. Sweet (Ed.), *Forensic neuropsychology* (pp. 27–38). New York: Taylor & Francis

Grant, D. A., & Berg, E. (1948). A behavioral analysis of degree of reinforcement and ease of shifting to new responses in a Weigl-type card-sorting problem. *Journal of experimental psychology*, *38*(4), 404.

Green, P., Rohling, M. L., Lees-Haley, P. R., & Allen, L. M. (2001). Effort has a greater effect on test scores than severe brain injury in compensation claimants. *Brain Injury, 15*(12), 1045-1060.

Green, P. (2003). *Green's Word Memory Test for Microsoft Windows*. Edmonton, Alberta: Green's Publishing Inc.

Greiffenstein, M. F., Baker, W. J., & Gola, T. (1994). Validation of malingered amnesia measures with a large clinical sample. *Psychological Assessment*, *6*(3), 218.

Greve, K. W., & Bianchini, K. J. (2004). Setting empirical cut-offs on psychometric indicators of negative response bias: a methodological commentary with recommendations. *Archives of Clinical Neuropsychology, 19*(4), 533-541.

Hall, V. L., Worthington, A., & Venables, K. (2014). A UK pilot study: The specificity of the Word Memory Test effort sub-tests in acute minimal to mild head injury. *Journal of neuropsychology*, *8*(2), 216-230.

Hampson, N. E., Kemp, S., Coughlan, A. K., Moulin, C. J. A., & Bhakta, B. B. (2014). Effort Test Performance in Clinical Acute Brain Injury, Community Brain Injury, and Epilepsy Populations. *Applied Neuropsychology-Adult, 21*(3), 183-194.

Heaton, R. K., Smith H. H., Lehman, R. A. W. & Vogt, A. T. (1978). Prospects for faking believable deficits on neuropsychological testing. *Journal of Consulting and Clinical Psychology*, *46,* 892-900.

Heaton, R. K. (1981). *A manual for the Wisconsin Card Sorting Test*. Western Psychological Services.

Heaton, R. K., Chelune, G. J., Talley, J. L., Kay, G. G., & Curtis, G. (1993). *Wisconsin Card Sorting Test (WCST) manual, revised and expanded.* Odessa, FL: Psychological Assessment Resources.

Heilbronner, R. L., Sweet, J. J., Morgan, J. E., Larrabee, G. J., Millis, S. R., Bianchini, K. J., Frederick, R. L. (2009). American Academy of Clinical Neuropsychology Consensus Conference Statement on the Neuropsychological Assessment of Effort, Response Bias, and Malingering. *Clinical Neuropsychologist, 23*(7), 1093-1129.

Holdnack, J. A., Schoenberg, M. R., Lange, R. T., Iverson, G. L. (2013). In Holdnack, J. A., Drozdick, L., Weiss, L. G. & Iverson, G.L. (Eds.). *WAIS-IV, WMS-IV, and ACS Advanced Clinical Interpretation.* (pp. 217-278). London: Academic Press.

Holdnack, J. A., Millis, S., Larrabee, G. J. & Iverson, G. L. (2013). Assessing performance validity with the ACS. In Holdnack, J. A., Drozdick, L., Weiss, L. G. & Iverson, G.L. (Eds.). *WAIS-IV, WMS-IV, and ACS Advanced Clinical Interpretation.* (pp. 331-365). London: Academic Press.

IBM Corp. Released 2013. *IBM SPSS Statistics for Windows, Version 22.0*. Armonk, NY: IBM Corp

Ikeda, M., Iwanaga, M., & Seiwa, H. (1996). Test anxiety and working memory system. *Perceptual and motor skills*, *82*, 1223-1231

Iverson, G. L. and Brooks, B. L. (2011) Improving accuracy in identifying cognitive impairments. In Schoenberg, M. R., & Scott, J. G. (Eds). *The Little Black Book of Neuropsychology: A Syndrome-Based Approach*. (pp 923-950). New York: Springer.

Iverson, G. L., & Franzen, M. D. (1994). The Recognition Memory Test, digit span, and Knox Cube Test as markers of malingered memory impairment. *Assessment*, *1*(4), 323-334.

Iverson, G. L., Le Page, J., Koehler, B. E., Shojania, K., & Badii, M. (2007). Test of memory malingering (TOMM) scores are not affected by chronic pain or depression in patients with fibromyalgia. *Clinical Neuropsychologist, 21*(3), 532-546.

Jasinski, L. J., Berry, D. T., Shandera, A. L., & Clark, J. A. (2011). Use of the Wechsler Adult Intelligence Scale Digit Span subtest for malingering detection: A meta-analytic review. *Journal of Clinical and Experimental Neuropsychology*, *33*(3), 300-314.

Keary, T. A., Frazier, T. W., Belzile, C. J., Chapin, J. S., Naugle, R. I., Najm, I. M., & Busch, R. M. (2013). Working memory and intelligence are associated with Victoria Symptom Validity Test hard item performance in patients with intractable epilepsy. *Journal of the International Neuropsychological Society*, *19*(3), 314-323.

Kemp, S., Coughlan, A. K., Rowbottom, C., Wilkinson, K., Teggart, V., & Baker, G. (2008). The base rate of effort test failure in patients with medically unexplained symptoms. *Journal of Psychosomatic Research, 65*(4), 319-325.

Larrabee, G. J. (2003). Detection of malingering using atypical performance patterns on standard neuropsychological tests. *Clinical Neuropsychologist, 17*(3), 410-425.

Larrabee, G. J. (2012). Performance validity and symptom validity in neuropsychological assessment. *Journal of the International Neuropsychological Society*, *18*(04), 625-630.

Larrabee, G. J. (2014). False-positive rates associated with the use of multiple performance and symptom validity tests. *Archives of clinical neuropsychology, 29*(4), 364-373.

Larrabee, G. J., Millis, S. R., & Meyers, J. E. (2009). 40 Plus or Minus 10, a New Magical Number: Reply to Russell. *Clinical Neuropsychologist, 23*(5), 841-849.

Lezak, M. D., Howieson, D. B., & Loring, D. W. (2004). *Neuropsychological assessment* (4th ed.). New York: Oxford University Press.

Locke, D. E. C., Smigielski, J. S., Powell, M. R., & Stevens, S. R. (2008). Effort issues in post-acute outpatient acquired brain injury rehabilitation seekers. *Neurorehabilitation, 23*(3), 273-281.

Martin, P. K., Schroeder, R. W., Wyman-Chick, K. A., Hunter, B. P., Heinrichs, R. J., & Baade, L. E. (2016). Rates of Abnormally Low TOPF Word Reading Scores in Individuals Failing Versus Passing Performance Validity Testing. *Assessment*, *1*, 13.

McCarter, R. J., Walton, N. H., Brooks, D. N., & Powell, G. E. (2009). Effort Testing in Contemporary UK Neuropsychological Practice. *Clinical Neuropsychologist, 23*(6), 1050-1066.

McKenna, P., & Warrington, E. K. (1983). *Graded naming test: Manual*. NFER-Nelson.

McMillan, T.M., Anderson, S., Baker, G., Berger, M., Powell, G.E., and Knight, R. (2009) *Assessment of effort in clinical testing of cognitive functioning for adults*. British Psychological Society, pp. 1-27.

Merten, T., Bossink, L., & Schmand, B. (2007). On the limits of effort testing: Symptom validity tests and severity of neurocognitive symptoms in nonlitigant patients. *Journal of Clinical and Experimental Neuropsychology, 29*(3), 308-318.

Meyers, J. E., & Volbrecht, M. E. (2003). A validation of multiple malingering detection methods in a large clinical sample. *Archives of Clinical Neuropsychology, 18*(3), 261-276.

Miele, A. S., Gunner, J. H., Lynch, J. K., & McCaffrey, R. J. (2012). Are Embedded Validity Indices Equivalent to Free-Standing Symptom Validity Tests? *Archives of Clinical Neuropsychology, 27*(1), 10-22.

Millis, S. R. (2009). Methodological Challenges in Assessment of Cognition following Mild Head Injury: Response to Malojcic et al. 2008. *Journal of Neurotrauma, 26*(12), 2409-2410.

Mittenberg, W, Patton, C., Canyock, E. M., & Condit, D. C. (2002). Base rates of malingering and symptom exaggeration. *Journal of Clinical and Experimental Neuropsychology, 24*(8), 1094-1102.

Morey, L. C. (1991). *Personality assessment inventory (PAI)*. John Wiley & Sons, Inc.

Morey, L. C. (2007). *Personality assessment inventory (PAI): professional manual*. PAR (Psychological Assessment Resources).

Moss, A., Jones, C., Fokias, D., & Quinn, D. (2003). The mediating effects of effort upon the relationship between head injury severity and cognitive functioning. *Brain Injury, 17*(5), 377-387.

Mossman, D., Wygant, D. B., & Gervais, R. O. (2012). Estimating the accuracy of neurocognitive effort measures in the absence of a "gold standard". *Psychological assessment*, *24*(4), 815.

Nelson, H. E. (1976). A modified card sorting test sensitive to frontal lobe defects. *Cortex*, *12*(4), 313-324.

Obonsawin, M. C., Crawford, J. R., Page, J., Chalmers, P., Low, G., & Marsh, P. (1999). Performance on the Modified Card Sorting Test by normal, healthy individuals: Relationship to general intellectual ability and demographic variables. *British journal of clinical psychology*, *38*(1), 27-41.

Rees, L. M., Tombaugh, T. N., Gansler, D. A., & Moczynski, N. P. (1998). Five validation experiments of the Test of Memory Malingering (TOMM). *Psychological Assessment*, *10*(1), 10.

Rees, L. M., Tombaugh, T. N., & Boulay, L. (2001). Depression and the Test of Memory Malingering. *Archives of Clinical Neuropsychology, 16*(5), 501-506.

Rey, A. (1964). *L'examen clinique en psychologie*. Paris: Presses Universitaires de France

Robinson, O. J., Vytal, K., Cornwell, B. R., & Grillon, C. (2013). The impact of anxiety upon cognition: perspectives from human threat of shock studies. *Frontiers in Human Neuroscience*, *7*.

Rohling, M. L., Green, P., Allen, L. M., & Iverson, G. L. (2002). Depressive symptoms and neurocognitive test scores in patients passing symptom validity tests. *Archives of clinical neuropsychology*, *17*(3), 205-222.

Rosenfeld, B., Sands, S. A., & Van Gorp, W. G. (2000). Have we forgotten the base rate problem? Methodological issues in the detection of distortion. *Archives of Clinical Neuropsychology*, *15*(4), 349-359.

Rudman, N., Oyebode, J. R., Jones, C. A., & Bentham, P. (2011). An investigation into the validity of effort tests in a working age dementia population. *Aging & Mental Health*, *15*(1), 47-57.

Russell, E. W., & Levy, M. (1987). Revision of the Halstead Category Test. *Journal of Consulting and Clinical Psychology*, *55*(6), 898.

Salazar, X. F., Lu, P. H., Wen, J., & Boone, K. B. (2007). The use of effort tests in ethnic minorities and in non-English-speaking and English as a second language populations. *Assessment of feigned cognitive impairment: A neuropsychological perspective*, 405-427.

Schroeder, R. W., Twumasi-Ankrah, P., Baade, L. E., & Marshall, P. S. (2012). Reliable Digit Span: A Systematic Review and Cross-Validation Study. *Assessment, 19*(1), 21-30.

Schutte, C. & Axelrod, B. N. (2012). Use of embedded cognitive symptom validity measures in mild traumatic brain injury cases. In Carone, D. A. & Bush, S. S. (Eds) *Mild traumatic brain injury: Symptom validity assessment and malingering*. Springer Publishing Company. (pp. 159 – 182)

Slick, D. J., Sherman, E. M., & Iverson, G. L. (1999). Diagnostic criteria for malingered neurocognitive dysfunction: Proposed standards for clinical practice and research. *The Clinical Neuropsychologist*, *13*(4), 545-561.

Spencer, R. J., Axelrod, B. N., Drag, L. L., Waldron-Perrine, B., Pangilinan, P. H., & Bieliauskas, L. A. (2013). WAIS-IV reliable digit span is no more accurate than age corrected scaled score as an indicator of invalid performance in a veteran sample undergoing evaluation for mTBI. *The Clinical Neuropsychologist*, *27*(8), 1362-1372.

Strauss, E., Sherman, E. M., & Spreen, O. (2006). *A compendium of neuropsychological tests: Administration, norms, and commentary*. Oxford University Press.

Suesse, M., Wong, V. W., Stamper, L. L., Carpenter, K. N., & Scott, R. B. (2015). Evaluating the Clinical Utility of the Medical Symptom Validity Test (MSVT): A Clinical Series. *The Clinical Neuropsychologist*, *29*(2), 214-231.

Taub, G. E., & Benson, N. (2013). Matters of consequence: An empirical investigation of the WAIS-III and WAIS-IV and implications for addressing the Atkins intelligence criterion. *Journal of Forensic Psychology Practice*, *13*(1), 27-48.

The Psychological Corporation (1997). *WAIS-III-WMS-II technical manual*. San Antonio: Author.

The Psychological Corporation (1997, 2002). *WAIS-III-WMS-III technical manual*. San Antonio, Tex: The Psychological Corporation.

Tombaugh, T. (1996). *Test of Memory Malingering*. Toronto, Ontario, Canada: Multi-Health Systems.

Tombaugh, T. N. (1997). The test of memory malingering (TOMM): Normative data from cognitively intact and cognitively impaired individuals. *Psychological Assessment, 9*(3), 260-268.

Tombaugh, T. N. (2003). The Test of Memory Malingering (TOMM) in forensic psychology. *Journal of Forensic Neuropsychology*, *2*(3-4), 69-96.

Trahan, L. H., Stuebing, K. K., Fletcher, J. M., & Hiscock, M. (2014). The Flynn effect: A meta-analysis. *Psychological bulletin*, *140*(5), 1332.

Vickery, C. D., Berry, D. T. R., Inman, T. H., Harris, M. J., & Orey, S. A. (2001). Detection of inadequate effort on neuropsychological testing: A meta-analytic review of selected procedures. *Archives of Clinical Neuropsychology, 16*(1), 45-73.

Victor, T. L., Boone, K. B., Serpa, J. G., Buehler, J., & Ziegler, E. A. (2009). Interpreting the Meaning of Multiple Symptom Validity Test Failure. *Clinical Neuropsychologist, 23*(2), 297-313.

Weschler, D. (1981). *WAIS-R manual.* San Antonio, TX: Psychological Corporation.

Wechsler, D. (1997a). *Wechsler Adult Intelligence Scale – Third Edition*. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1997b). *Wechsler Memory Scale – Third Edition*. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2001). *Wechsler Test of Adult Reading*. San Antonio, TX: The Psychological Corporation.

Wechsler D. (2008) *Wechsler Adult Intelligence Scale – Fourth Edition*. San Antonio, TX: Pearson Assessment

Wechsler D. (2009) *Wechsler Memory Scale – Fourth Edition*. San Antonio, TX: Pearson Assessment

Wechsler, D. (2011). *Test of premorbid functioning: UK version (TOPF UK)*. UK: Pearson Corporation.

West, L. K., Curtis, K. L., Greve, K. W., & Bianchini, K. J. (2011). Memory in traumatic brain injury: The effects of injury severity and effort on the Wechsler Memory Scale- III. *Journal of Neuropsychology*, *5*(1), 114-125.

Willis, P. F., Farrer, T. J., & Bigler, E. D. (2011). Are effort measures sensitive to cognitive impairment? *Military medicine*, *176*(12).

Whipple Drozdick, L., Holdnack, J. A., Weiss, L. G., Zhou, X. (2013). In Holdnack, J. A., Drozdick, L., Weiss, L. G. & Iverson, G.L. (Eds.). *WAIS-IV, WMS-IV, and ACS Advanced Clinical Interpretation.* (pp. 1-69). London: Academic Press.

Young, J. C., Sawyer, R. J., Roper, B. L., & Baughman, B. C. (2012). Expansion and re-examination of Digit Span effort indices on the WAIS-IV. *The Clinical Neuropsychologist, 26*(1), 147-159.

Part 3: Critical Appraisal

In conducting this research I have encountered a number of conceptual issues which I feel speak strongly to the clinical applications of my findings. These are touched on briefly in the empirical paper, however, I believe that they warrant extended discussion here.

Clinician attitudes to performance validity testing

All participants recruited prospectively into the study required comprehensive neuropsychological assessment as part of their clinical care, however, only a small minority of these had performance validity assessment planned as part of their clinical test battery. In fact, some clinicians reported informally that they would not have utilised standalone performance validity tests (PVTs) in most cases had they not been taking part in this research study. As such, one could conclude that performance validity testing was not considered an essential component to comprehensive assessment. Where the archival sample had more consistent testing with PVTs, this was primarily related to the long-standing research interests of the team. This anecdotal experience raises questions about the degree to which clinical guidelines regarding the use of PVTs are currently being followed. The British Psychological Society (BPS) professional guidelines on the assessment of effort in clinical settings were released in 2009 (McMillan et al., 2009), and stated unequivocally that PVTs should be given routinely as part of clinical assessment of cognitive function. As such, it was interesting to note that eight years on from the publication of this guidance there does not appear to have been a consistent shift in clinical practice at ground level.

These observations mirror the results of a 2009 survey of 130 practicing UK neuropsychologists by McCarter, Walton, Brooks & Powell (2009), which

indicated that only 16% of respondents routinely used PVTs in clinical assessments. Since 2009, the performance validity assessment field has expanded dramatically, to the extent that between 2011 and 2016, 25% of peer-reviewed articles published by *The Clinical Neuropsychologist* and *Archives of Clinical Neuropsychology* were related to this topic (Martin, Schroeder & Odland, 2015). Whilst it was outside the scope of this research project to examine the specific attitudes and beliefs of the clinicians involved towards performance validity testing, there have been a number of recently published articles which have explored this issue further.

In a large survey of neuropsychologists across six Western European countries (not including the UK), Dandachi-Fitzgerald, Ponds & Merten (2013) found inconsistencies between the acknowledgement of the occurrence of non-credible symptoms in clinical practice (average clinician prevalence estimates were around 10%) and the use of objective measures to assess this phenomenon. Fifty per cent of the 492 clinical neuropsychologists surveyed reported that, despite having technical knowledge of performance validity, they 'very rarely or never' included PVTs in their assessments, opting to rely on qualitative metrics such as discrepancies between records, self-report and behaviour, or finding that the severity of cognitive impairment conflicted with known aspects of the patient's condition. By contrast, two North American surveys (Martin et al., 2015; Schroeder, Martin & Odland, 2016) found a pronounced "paradigm shift" in clinical neuropsychological practice in the years since the publication of two consensus statements recommending the formal application of PVTs in both forensic and clinical assessments (Bush et al., 2005; Heilbronner et al., 2009). Nearly all neuropsychologists surveyed utilised both standalone and embedded validity measures and only 6% believed them to be 'optional or unnecessary' in clinical assessments (Martin et al., 2015). It is

interesting to note the contrast in beliefs and practice between North American and European clinicians regarding the use of PVTs, whereby North American clinical practice appears to have greater convergence between research and clinical findings. It is possible that this reflects the increased emphasis on medicolegal and forensic work in North American neuropsychology and a longer tradition of performance validity testing (Suesse, Wong, Stamper, Carpenter, & Scott, 2015).

My experiences are consistent with the European data from the Dandachi-Fitzgerald et al. (2013) study, as it would appear that, in day-to-day NHS clinical practice, clinicians may be more likely to rely on subjective judgement for the evaluation of performance validity than on psychometric tests. In line with this, the McCarter et al. (2009) survey of UK clinicians highlighted a prevailing belief that 'invalidity would be obvious' in patient presentation or test scores. This view is somewhat problematic as there is ample evidence from across the social sciences to suggest that clinician impressions can be unreliable. For example, a meta-analytic review comparing clinical versus mechanical (that is, formal or statistical) decision-making demonstrated the 'general superiority' of mechanical prediction across clinical settings: this effect held for both medical professionals and psychologists and for experienced and inexperienced clinicians (Grove, Zald, Lebow, Snitz & Nelson, 2000). As regards the judgement of performance validity, it has been demonstrated that the use of heuristics as decisional simplification strategies can influence clinician assessment (Guilmette, 2013). The 'representativeness' heuristic, for example, describes the situation whereby the probability of an event is determined based on past experience or assumptions (Tversky & Kahneman, 1974), which may lead to the exclusion of actual base rate information in clinical decision-making. Given that base rate data on performance invalidity in NHS settings has been notable by its absence,

it is perhaps unsurprising that clinicians are more likely to resort to their own practice-based evidence to support judgements in this regard. It is hoped that the empirical study outlined in this thesis will supplement the evidence-base and that others will continue to research along this vein to assist clinician judgement using PVTs with their patient populations.

Further barriers to PVT use included practical considerations such as administration time and cost of measures (McCarter et al., 2009; Dandachi-Fitzgerald et al., 2013). In addition, the McCarter et al. (2009) survey included concern amongst clinicians that PVTs are 'unreliable'. Whilst the empirical study in this thesis certainly raises issues in this regard, it could also be asserted that ongoing reliance on subjective assessment alone is almost certainly incompatible with evidence-based practice (Dandachi-Fitzgerald et al., 2013). In line with this, I would query whether the lack of explicit guidance for clinicians regarding a) how to interpret PVTs with clinical populations and b) management strategies for individuals failing PVTs act as significant obstacles to implementation. These issues will be discussed in the remainder of this chapter.

What are we actually testing when we use performance validity tests? What can test failure tell us?

A key narrative thread running through this thesis has been the importance of utilising PVTs as formal means of assessing the validity of neuropsychological data. The overarching message in professional guidelines is that to neglect this is to potentially "leave the door wide open to artificially supressed scores" (Green, 2003, p. 626). This perspective is based on the significant body of evidence focused on improving detection techniques, which has provided important information regarding

the validity of PVTs. Nonetheless, there remains a significant degree of tautology and confusion regarding the phenomenon in question when looking to apply the evidence-base outside of medicolegal or forensic practice. At a basic level, the nomenclature used within this field lacks clarity. My decision to utilise the term 'performance validity' throughout this thesis was a conscious one, motivated by recent shifts in thinking away from more pejorative terminology using phrases such as 'effort', 'negative response bias' and particularly 'malingering' (see Larrabee, 2012, for a discussion of this issue). In practice and in the literature these phrases are often used synonymously yet there are few accepted operationalised definitions of what the terms mean and how they might overlap or differ from each other. The very fact that they are used interchangeably and inconsistently is illustrative of the ambiguity which plagues this field. As a result of this, even after considerable time spent analysing and synthesising the literature, I am left with unanswered questions about how best to conceptualise diminished performance on PVTs and, in turn, what should be done with the information they add to clinical assessments.

PVTs were originally conceived in the 1980's as malingering tests to be used primarily in the medicolegal arena. The literature has evolved, and there is now a general consensus that malingering is only one source of atypical performance on PVTs (McMillan et al., 2009; Merten & Merckelbach, 2013). Scores below chance on these measures (for example, scoring under 25 on the 50-item Test of Memory Malingering [TOMM]; Tombaugh, 1996) continue to be accepted as indicative of malingering as this strongly suggests that the patient is voluntarily endorsing incorrect answers (Bush et al., 2005; Bigler, 2012). Cut-off scores on these measures are, however, are generally set significantly above chance levels, yet there is a tendency within the literature (and, in my anecdotal experience, within clinical

practice), to equate this 'failed' performance with commentaries regarding the degree of 'effort' applied by the examinee. The logic behind this is that PVTs are so straightforward that the only cognitive requirement needed to pass is a general level of engagement with the task (Bigler, 2014), but to what extent is this a useful heuristic for clinical practitioners?

Cognitive neuroscience research indicates that, aside from primary sensory stimulation, all other tasks require a degree of cognitive processing, no matter how simple the task may appear at a surface level (Bigler, 2014). This clearly creates an issue for proponents of PVT use, as it would make the task of teasing apart the organic from the non-organic symptomatology extremely challenging. Indeed, this issue was highlighted in both the systematic review and empirical paper, as it has been demonstrated that a) different forms of dementia with varying neuropathological bases differentially impact on PVTs tapping different cognitive processes and b) that people passing and failing PVTs do not appear to have globally suppressed scores on neuropsychological assessment. One the basis of these data, assuming assessment invalidity purely as a result of failed PVTs would not be empirically supportable as the scores could be picking up on features related to the organic basis of the individual's condition.

PVT failure may also have a psychological basis beyond intentional underperformance. One school of thought is that diminished performance on testing may be part of a medically unexplained syndrome such as a somatoform disorder, and that non-conscious processes are acting to impact responding on PVTs. The 'non-conscious' mechanisms are not well defined, but one suggestion is that anxiety may bias individuals to respond in a way which is congruent with their health beliefs (for example, that they are cognitively impaired as a result of a condition of the

brain; Kemp et al., 2008). Related to this, some authors have proposed that diminished PVT performance taps an aspect of illness behaviour (Bigler, 2014). It is pertinent to this discussion to consider the association between illness behaviour and the potential benefits that society provides to individuals assuming the sick role, particularly when the sickness is a result of a physical condition as opposed to psychological or emotional disorders (Bass & Halligan, 2014). The determinants of illness behaviour are complex and multifactorial, but it is not a necessary condition that this behaviour is underpinned by consciously mediated choice or intention to deceive as would be the case for a diagnosis of factitious disorder or malingering. Psychological frameworks are beginning to emerge which identify other pathways to a behavioural phenomenon which may present in a similar way. For example, Merckelbach and Merten (2012) propose a cognitive dissonance model whereby individuals may initially misreport symptoms, but to resolve the internal conflict evoked by this behaviour (for example where beliefs about their own honesty are challenged by their actions) they begin to 'deceive' themselves that fabricated experiences are genuinely felt. Such a perspective emphasises that there are no sharp demarcation lines between malingering, somatoform or medically unexplained symptoms as would be suggested by the DSM-V taxonomy, and that perhaps it would be more helpful to consider these issues along a continuum.

In examining this debate it is clear that there is no parsimonious account which would explain why people fail PVTs. Whilst Occam's razor may favour "uncooperativeness" or similar (Merten & Mercklebach, 2013), this cannot be assumed to be the case for all patients and to draw this conclusion is to fail to take into account the myriad reasons which may underpin diminished scores. Given the evidence covered here, it would seem that neuropsychological test performance is

best considered as a construct: something that is not directly observable but can be explored using PVTs as a tool. If PVTs are considered in this way, perhaps this overcomes the seemingly insurmountable problem of whether or not they enable clinicians to impute motivation and volition, conscious or unconscious processes, organic or non-organic symptoms or internal or external goals (Berry & Nelson, 2010). This is in line with the perspective offered by Rogers, Sewell and Gillard (2010), which echoes the idea that PVTs cannot claim to identify malingering (even if, like the TOMM, they purport to do so in their title), but rather can only aim to detect invalid symptoms "without any assumptions about…goals" (Rogers et al., 2010, p. 5). Instead, low scores on PVTs may act as prompts for clinicians to think more broadly about how they conceptualise the individual and as one piece in a formulation jigsaw. As highlighted by Dandachi-Fitzgerald et al. (2013), scientific progress in this area feels limited to the improvement of detection techniques at the expense of exploring conceptual issues which "provide meaning to non-credible symptoms" (p. 782): perhaps future research needs to shift focus to shed light on the latter.

What do we do with failed PVTs in clinical assessments?

Hand-in-hand with a lack of clarity regarding what PVTs can show us, there is a limited evidence base focusing on the management of patients thought to be displaying invalid performance. The BPS guidelines (McMillan et al., 2009) provide some practical recommendations for how to manage suspected 'poor effort' within the assessment session. These emphasise that clinicians should actively reconsider the planned assessment (for example, whether additional PVTs need to be incorporated) but to continue testing such that sufficient information is available to produce a general formulation. The TOMM manual (Tombaugh, 1996) offers

guidance regarding how to communicate test failures in reports to clinicians and patients. A much-cited paper by Carone, Iverson & Bush (2010) further tackles the issue of cognitive assessment feedback in the context of PVT failure, providing a comprehensive framework to support clinicians to discuss this issue transparently and therapeutically. This is pertinent given findings of the Dandachi-Fitzgerald et al. (2013) survey, where respondents were divided regarding how they communicated issues around performance validity to their patients.

Other authors have examined the impact of speaking openly with patients about invalid performance during testing. Suchy, Chelune, Franchow & Thorgusen (2012) found that after confrontation regarding 'non-valid' PVT performance, two thirds of patients in a non-forensic sample produced valid scores on subsequent re-examination, both on the PVT and subsequent memory testing. This suggests that by tackling the issue directly in the assessment context, the clinician increases the likelihood of obtaining valid results and may gain additional insights regarding the cause of the failed PVT, which would likely help mitigate some of the issues raised in the previous section regarding our understanding of what PVT failure may represent. Alongside this, however, Carone et al. (2010) acknowledge that clinicians may experience 'anticipatory fear' in feeding back potentially contentious information about performance validity to clients, which again leads one to question if this is a barrier to the systematic use of PVTs in clinical settings.

Aside from management *in vivo* during testing and considerations regarding the communication of PVT results to clients and professionals, there is little written regarding psychological approaches for people who fail PVTs. Of note is that the TOMM manual emphasises the utility of "…trying to determine the motivation underlying the exaggerations and then including this information as part

of the final diagnosis, along with suggested interventions" (Tombaugh, 1996; p. 22). There is no guidance given, however, on what management strategies may be particularly useful for this cohort of patients. Again, this is potentially related to the origins of this literature in the medicolegal field, where PVTs perform a more binary function to help clinicians establish whether or not there are non-organic aspects of an individual's presentation. Applying PVTs in the clinical setting where clinicians may continue to see the patient therapeutically beyond the assessment, however, requires a somewhat more nuanced view which follows in the tradition of collaborative, person-centred care. This is one which is based on sound formulation of the patient's presentation which does not seek to mitigate the subjective experience of the patient by attributing symptoms directly to lack of 'effort' or motivation, but to provide a foundation for alternative explanations and therefore treatment approaches (Bass & Halligan, 2014). This perspective has been summarised succinctly by Stone & Boone (2007), who stress that reducing performance validity to 'moral failing', as inferred by some labels applied to describe this phenomenon ('feigning', 'malingering', 'non-credible' and so on), is unhelpful, and detracts the clinician from examining clues as to *why* the behaviour is occurring.

Summary

This appraisal has covered three areas which, on completion of this project, I felt warranted further exploration. I began with a consideration of my experiences working in neuropsychology settings with clinicians, and reflected on the inconsistent application of PVTs in NHS clinical practice despite clear professional guidelines and a large evidence-base emphasising the utility of incorporating PVTs in neuropsychological assessments. Having further explored the lack of clarity around what PVTs can tell us about performance and the dearth of guidance

regarding how to manage and intervene with individuals failing PVTs, it is perhaps not surprising that clinicians elect not to incorporate these measures on a systematic basis. It would be helpful for future research to more fully address these issues and, as the evidence-base evolves, for this to be incorporated into clinical guidelines. The discipline of neuropsychology has a tradition of strong, scientifically-based practice which is an area of particular strength when compared to other specialities in healthcare (Schroeder et al., 2016). Given that the area of performance validity testing is receiving significant research attention at the current time, this is now a good opportunity to consider these emerging issues. I would agree with the position of Stone & Boone (2007) that performance validity is "fascinating…worthy of continued, collaborative and enthusiastic research" (p. 11), but that field now needs to evolve from understanding how we recognise performance invalidity towards understanding the behavioural phenomenon of 'invalid performance' in greater depth. By doing so, this represents an important therapeutic step to understanding our patients' needs.

References

Bass, C., & Halligan, P. (2014). Factitious disorders and malingering: challenges for clinical assessment and management. *The Lancet, 383*(9926), 1422-1432.

Berry, D. T., & Nelson, N. W. (2010). DSM-5 and malingering: A modest proposal. *Psychological Injury and Law*, *3*(4), 295-303.

Bigler, E. D. (2012). Symptom validity testing, effort, and neuropsychological assessment. *Journal of the International Neuropsychological Society, 18*(04), 632-640.

Bigler, E. D. (2014). Effort, symptom validity testing, performance validity testing and traumatic brain injury. *Brain injury*, *28*(13-14), 1623-1638.

Bigler, E. D. (2015). Neuroimaging as a biomarker in symptom validity and performance validity testing. *Brain imaging and behavior, 9*(3), 421-444.

Bush, S. S., Ruff, R. M., Tröster, A. I., Barth, J. T., Koffler, S. P., Pliskin, N. H., Reynolds, C. R. & Silver, C. H. (2005). Symptom validity assessment: Practice issues and medical necessity: NAN Policy & Planning Committee. Archives of Clinical Neuropsychology, 20(4), 419-426.

Carone, D. A., Iverson, G. L., & Bush, S. S. (2010). A model to approaching and providing feedback to patients regarding invalid test performance in clinical neuropsychological evaluations. The Clinical Neuropsychologist, 24(5), 759-778.

Dandachi-FitzGerald, B., Ponds, R. W., & Merten, T. (2013). Symptom validity and neuropsychological assessment: A survey of practices and beliefs of

neuropsychologists in six European countries. Archives of Clinical Neuropsychology, 28(8), 771-783.

Green, P. (2003). Welcoming a paradigm shift in neuropsychology. Archives of Clinical Neuropsychology, 18(6), 625-627.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: a meta-analysis. *Psychological assessment*, *12*(1), 19.

Guilmette, T. J. (2013). The role of clinical judgment in symptom validity assessment. Mild traumatic brain injury: Symptom validity assessment and malingering, 31-41.

Heilbronner, R. L., Sweet, J. J., Morgan, J. E., Larrabee, G. J., Millis, S. R., Bianchini, K. J., Frederick, R. L. (2009). American Academy of Clinical Neuropsychology Consensus Conference Statement on the Neuropsychological Assessment of Effort, Response Bias, and Malingering. Clinical Neuropsychologist, 23(7), 1093-1129.

Kemp, S., Coughlan, A. K., Rowbottom, C., Wilkinson, K., Teggart, V., & Baker, G. (2008). The base rate of effort test failure in patients with medically unexplained symptoms. *Journal of Psychosomatic Research, 65*(4), 319-325.

Larrabee, G. J. (2012). Performance validity and symptom validity in neuropsychological assessment. *Journal of the International Neuropsychological Society*, *18*(04), 625-630.

Martin, P. K., Schroeder, R. W., & Odland, A. P. (2015). Neuropsychologists' validity testing beliefs and practices: A survey of North American professionals. The Clinical Neuropsychologist, 29(6), 741-776.

McCarter, R. J., Walton, N. H., Brooks, D. N., & Powell, G. E. (2009). Effort Testing in Contemporary UK Neuropsychological Practice. *Clinical Neuropsychologist, 23*(6), 1050-1066.

McMillan, T.M., Anderson, S., Baker, G., Berger, M., Powell, G.E., and Knight, R. (2009) Assessment of effort in clinical testing of cognitive functioning for adults. British Psychological Society, pp. 1-27.

Merckelbach, H., & Merten, T. (2012). A note on cognitive dissonance and malingering. The Clinical Neuropsychologist, 26(7), 1217-1229.

Merten, T., & Merckelbach, H. (2013). Symptom validity testing in somatoform and dissociative disorders: A critical review. Psychological Injury and Law, 6(2), 122-137.

Rogers, R., Sewell, K. W., & Gillard, N. D. (2010). Structured interview of reported symptom professional manual. Odessa: Psychological Assessment Resources.

Schroeder, R. W., Martin, P. K., & Odland, A. P. (2016). Expert beliefs and practices regarding neuropsychological validity testing. The Clinical Neuropsychologist, 30(4), 515-535.

Stone, D. C., & Boone, K. B. (2007). Feigning of physical, psychiatric, and cognitive symptoms: Examples from history, the arts, and animal behavior. In Boone, K. B. (Ed.). (2007). *Assessment of feigned cognitive impairment: A neuropsychological perspective*. Guilford Press.

Suchy, Y., Chelune, G., Franchow, E. I., & Thorgusen, S. R. (2012). Confronting patients about insufficient effort: The impact on subsequent symptom validity and memory performance. *The Clinical Neuropsychologist*, *26*(8), 1296-1311.

Suesse, M., Wong, V. W., Stamper, L. L., Carpenter, K. N., & Scott, R. B. (2015). Evaluating the Clinical Utility of the Medical Symptom Validity Test (MSVT): A Clinical Series. *The Clinical Neuropsychologist*, *29*(2), 214-231.

Tombaugh, T. (1996). Test of Memory Malingering. Toronto, Ontario, Canada: Multi-Health Systems.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Science, 185(4157), 1124-1131.

Appendix 1:

Description of joint projects

Description of joint projects

The empirical study was conducted in part collaboration with a Salomons DClinPsy student, Jessica Hooker, whose thesis is due for completion in summer 2018. The current study focused on performance validity test (PVT) pass and fail groups in terms of differences in cognitive testing. My colleague's thesis will look to examine the results of symptom-validity tests (SVTs) in terms of group differences in self-reported affective and personality variables. SVT information is only available for the archival data (see Part 2, Method section), therefore only these data are shared between the projects. Both researchers completed separate applications for ethical and local research and development office approval. Completion of the archival database was done jointly. All analysis and write-up has been conducted separately.

Appendix 2:

Ethical approval

NHS
Health Research Authority
London - City & East Research Ethics Committee
Bristol Research Ethics Committee Centre
Whitefriars
Level 3, Block B
Lewins Mead
Bristol
BS1 2NT

Telephone: 01173421386

16 September 2015

Dr John King
Research Department of Clinical, Educational and Health Psychology
1-19 Torrington Place
London
WC1E 7HB

Dear Dr King

Study title:	**Effort test performance in an NHS acquired brain injury sample**
REC reference:	**15/LO/1376**
IRAS project ID:	**170258**

The Research Ethics Committee reviewed the above application at the meeting held on 03 September 2015. Thank you for attending to discuss the application.

We plan to publish your research summary wording for the above study on the HRA website, together with your contact details. Publication will be no earlier than three months from the date of this favourable opinion letter. The expectation is that this information will be published for all studies that receive an ethical opinion but should you wish to provide a substitute contact point, wish to make a request to defer, or require further information, please contact the REC Manager Mr Rajat Khullar, nrescommittee.london-cityandeast@nhs.net. Under very limited circumstances (e.g. for student research which has received an unfavourable opinion), it may be possible to grant an exemption to the publication of the study.

Ethical opinion

The members of the Committee present gave a favourable ethical opinion of the above research on the basis described in the application form, protocol and supporting documentation, subject to the conditions specified below. .

Conditions of the favourable opinion

The favourable opinion is subject to the following conditions being met prior to the start of the study.

1. *As agreed, reference to American forensic psychology should be revised to the term that lay readers in the UK are able to understand. This relates to PIS and application.*

You should notify the REC in writing once all conditions have been met (except for site approvals from host organisations) and provide copies of any revised documentation with updated version numbers. The REC will acknowledge receipt and provide a final list of the

approved documentation for the study, which can be made available to host organisations to facilitate their permission for the study. Failure to provide the final versions to the REC may cause delay in obtaining permissions.

Management permission or approval must be obtained from each host organisation prior to the start of the study at the site concerned.

Management permission ("R&D approval") should be sought from all NHS organisations involved in the study in accordance with NHS research governance arrangements.

Guidance on applying for NHS permission for research is available in the Integrated Research Application System or at http://www.rdforum.nhs.uk.

Where a NHS organisation's role in the study is limited to identifying and referring potential participants to research sites ("participant identification centre"), guidance should be sought from the R&D office on the information it requires to give permission for this activity.

For non-NHS sites, site management permission should be obtained in accordance with the procedures of the relevant host organisation.

Sponsors are not required to notify the Committee of approvals from host organisations.

Registration of Clinical Trials

All clinical trials (defined as the first four categories on the IRAS filter page) must be registered on a publically accessible database. This should be before the first participant is recruited but no later than 6 weeks after recruitment of the first participant.

There is no requirement to separately notify the REC but you should do so at the earliest opportunity e.g. when submitting an amendment. We will audit the registration details as part of the annual progress reporting process.

To ensure transparency in research, we strongly recommend that all research is registered but for non-clinical trials this is not currently mandatory.

If a sponsor wishes to request a deferral for study registration within the required timeframe, they should contact hra.studyregistration@nhs.net. The expectation is that all clinical trials will be registered, however, in exceptional circumstances non registration may be permissible with prior agreement from the HRA. Guidance on where to register is provided on the HRA website.

It is the responsibility of the sponsor to ensure that all the conditions are complied with before the start of the study or its initiation at a particular site (as applicable).

Ethical review of research sites

NHS Sites

The favourable opinion applies to all NHS sites taking part in the study taking part in the study, subject to management permission being obtained from the NHS/HSC R&D office prior to the start of the study (see "Conditions of the favourable opinion" below).

Summary of discussion at the meeting

Favourable risk benefit ratio; anticipated benefit/risks for research participants (present and future)

The Committee expressed concerns with regards to mental capacity of the participants to be included in the study. The Committee noted that the application mentions that those unable to consent will not be included in the study, however it is not clear how would the capacity be assessed and how would It be made absolutely sure that none of the participants lack capacity. Dr Sunak explained that the PIS to be used will be in standard formal at accessible format and that will allow the participant to engage with them and help them to make sure if they understand everything in the study. He added that they will ensure that they understand what are the consequences of participating as compared to their routine care and are aware of the tests and procedures in the study. Dr Sunak explained that they will also make sure that these participants fully understand their options of discontinuing and the consequences of the same.
The Committee acknowledged the explanation provided by Dr Sunak, however the Committee asked if they are absolutely certain that their contact with the participants would be sufficient to decide that they are competent to understand and take part in the study. Dr Sunak confirmed the same.

Other general comments

The Committee queried about the link with American forensic psychology as mentioned in the PIS and the application. Dr Sunak explained that the term is used differently in the US as compared to the UK. The Committee suggested that it should be revised to the term that lay readers in the UK are able to understand it. Dr Sunak agreed.

Approved documents

The documents reviewed and approved at the meeting were:

Document	*Version*	*Date*
Evidence of Sponsor insurance or indemnity (non NHS Sponsors only) [Insurance confirmation]	1	17 July 2015
IRAS Checklist XML [Checklist_23072015]		23 July 2015
Participant consent form	3	15 May 2015
Participant consent form	3	15 May 2015
Participant information sheet (PIS)	3	15 May 2015
Participant information sheet (PIS)	3	15 May 2015
Participant information sheet (PIS)	3	15 May 2015
REC Application Form [REC_Form_23072015]		23 July 2015
Research protocol or project proposal	1.2	24 March 2015
Summary CV for Chief Investigator (CI)	1	24 March 2015
Summary CV for student	1	01 April 2015
Summary, synopsis or diagram (flowchart) of protocol in non technical language	1.0	10 February 2015

Membership of the Committee

The members of the Ethics Committee who were present at the meeting are listed on the attached sheet.

There were no declarations of interest

The Committee is constituted in accordance with the Governance Arrangements for Research Ethics Committees and complies fully with the Standard Operating Procedures for Research Ethics Committees in the UK.

After ethical review

Reporting requirements

The attached document "After ethical review – guidance for researchers" gives detailed guidance on reporting requirements for studies with a favourable opinion, including:

- Notifying substantial amendments
- Adding new sites and investigators
- Notification of serious breaches of the protocol
- Progress and safety reports
- Notifying the end of the study

The HRA website also provides guidance on these topics, which is updated in the light of changes in reporting requirements or procedures.

User Feedback

The Health Research Authority is continually striving to provide a high quality service to all applicants and sponsors. You are invited to give your view of the service you have received and the application procedure. If you wish to make your views known please use the feedback form available on the HRA website: http://www.hra.nhs.uk/about-the-hra/governance/quality-assurance/

HRA Training

We are pleased to welcome researchers and R&D staff at our training days – see details at http://www.hra.nhs.uk/hra-training/

15/LO/1376	**Please quote this number on all correspondence**

With the Committee's best wishes for the success of this project.

Yours sincerely

pp Dr John Keen
Chair

E-mail: nrescommittee.london-cityandeast@nhs.net

Enclosures: *List of names and professions of members who were present at the meeting and those who submitted written comments*

"After ethical review – guidance for researchers"

Copy to: *Ms Smaragda Agathou, Joint Research Office,UCL*

NHS
Health Research Authority

London - City & East Research Ethics Committee
Bristol Research Ethics Committee Centre
Whitefriars
Level 3, Block B
Lewins Mead
Bristol
BS1 2NT

Telephone: 01173421386

06 November 2015

Dr John King
Research Department of Clinical, Educational and Health Psychology
1-19 Torrington Place
London
WC1E 7HB

Dear Dr King

Study title: **Effort test performance in an NHS acquired brain injury sample**
REC reference: **15/LO/1376**
IRAS project ID: **170258**

Thank you for your email of 26th October 2015. I can confirm the REC has received the documents listed below and that these comply with the approval conditions detailed in our letter dated 18 September 2015

Documents received

The documents received were as follows:

Document	*Version*	*Date*
Participant information sheet (PIS) [Participant Information Sheet_Homerton]	5	21 September 2015
Participant information sheet (PIS) [Participant Information Sheet_Homerton OPT IN]	5	21 September 2015
Research protocol or project proposal	1.3	21 September 2015

Approved documents

The final list of approved documentation for the study is therefore as follows:

Document	*Version*	*Date*
Evidence of Sponsor insurance or indemnity (non NHS Sponsors only) [Insurance confirmation]	1	17 July 2015
IRAS Checklist XML [Checklist_23072015]		23 July 2015
Participant consent form	3	15 May 2015

Accessible Participant consent form	3	15 May 2015
Accessible Participant information sheet (PIS)	3	15 May 2015
Participant information sheet (PIS) [Participant Information Sheet_Homerton]	5	21 September 2015
Participant information sheet (PIS) [Participant Information Sheet_Homerton OPT IN]	5	21 September 2015
REC Application Form [REC_Form_23072015]		23 July 2015
Research protocol or project proposal	1.3	21 September 2015
Summary CV for Chief Investigator (CI)	1	24 March 2015
Summary CV for student	1	01 April 2015
Summary, synopsis or diagram (flowchart) of protocol in non technical language	1.0	10 February 2015

You should ensure that the sponsor has a copy of the final documentation for the study. It is the sponsor's responsibility to ensure that the documentation is made available to R&D offices at all participating sites.

15/LO/1376 **Please quote this number on all correspondence**

Yours sincerely

Rajat Khullar
REC Manager

E-mail: nrescommittee.london-cityandeast@nhs.net

Copy to: *Ms Smaragda Agathou, Joint Research Office,UCL*

NHS
Health Research Authority

Dr John King
Research Department of Clinical, Educational and Health Psychology
1-19 Torrington Place
London
WC1E 7HB

Email: hra.approval@nhs.net

09 August 2017

Dear Dr King

Letter of HRA Approval for a study processed through pre-HRA Approval systems

Study title:	**Effort test performance in an NHS acquired brain injury sample**
IRAS project ID:	**170258**
Sponsor	**University College London**
Amendment number:	**NSA # 1**
Amendment date:	**13/07/2017**

Thank you for your request to bring the above referenced study, processed under pre-HRA Approval systems, under HRA Approval.

I am pleased to confirm that the study has been given **HRA Approval.** This has been issued on the basis of an existing assessment of regulatory compliance, which has confirmed that the study is compliant with the UK wide standards for research in the NHS.

The extension of HRA Approval to this study on this basis allows the sponsor and participating NHS organisations in England to set-up the study in accordance with HRA Approval processes, with decisions on study set-up being taken on the basis of capacity and capability alone.

Please note that the amendment submitted to bring this study under HRA Approval (referenced above) is also approved by issue of this letter. You should not expect anything further from the HRA regarding the amendment. If the submitted amendment included the addition of a new NHS organisation in England, the addition of the new NHS organisation is also approved and should be set up in accordance with HRA Approval processes (e.g. the organisation should be invited to assess and arrange its capacity and capability to deliver the study and confirm once it is ready to do so).

Participation of NHS Organisations in England

Please note that full information to enable set up of participating NHS organisations in England is not provided in this letter, on the basis that activities to set up these NHS organisations is likely to be underway already.

The sponsor should provide a copy of this letter, together with the local document package and a list of the documents provided, to participating NHS organisations in England that are being set up in accordance with HRA Approval Processes. It is for the sponsor to ensure that any documents provided to participating organisations are the current, approved documents.

For non-commercial studies the local document package provided to NHS organisations should include an appropriate Statement of Activities and HRA Schedule of Events. The sponsor should also provide the template agreement to be used in the study, where the sponsor is using an agreement in addition to the Statement of Activities. Participating NHS organisations in England should be aware that the Statement of Activities and Schedule of Events for this study have not been validated by the HRA, but the HRA expects that the sponsor provides these to participating NHS organisations. Any changes that are appropriate to the content of the Statement of Activities and Schedule of Events should be agreed in a pragmatic fashion as part of the process of assessing, arranging and confirming capacity and capability to deliver the study.

For commercial studies the local document package should include a validated industry costing template and the template agreement to be used with participating NHS organisations in England.

It is critical that you involve both the research management function (e.g. R&D office and, if the study is on the NIHR portfolio, the LCRN) supporting each organisation and the local research team (where there is one) in setting up your study. Contact details and further information about working with the research management function for each organisation can be accessed from www.hra.nhs.uk/hra-approval.

If subsequent NHS organisations in England are added, an amendment should be submitted to the HRA.

After HRA Approval

In addition to the document, *"After Ethical Review – guidance for sponsors and investigators"*, issued with your REC Favourable Opinion, please note the following:

- HRA Approval applies for the duration of your REC favourable opinion, unless otherwise notified in writing by the HRA.
- Substantial amendments should be submitted directly to the Research Ethics Committee, as detailed in the *After Ethical Review* document. Non-substantial amendments should be submitted for review by the HRA using the form provided on the HRA website, and emailed to hra.amendments@nhs.net.

- The HRA will categorise amendments (substantial and non-substantial) and issue confirmation of continued HRA Approval.

The HRA website also provides guidance on these topics and is updated in the light of changes in reporting expectations or procedures.

Scope

HRA Approval provides an approval for research involving patients or staff in NHS organisations in England.

If your study involves NHS organisations in other countries in the UK, please contact the relevant national coordinating functions for support and advice. Further information can be found at http://www.hra.nhs.uk/resources/applying-for-reviews/nhs-hsc-rd-review/.

If there are participating non-NHS organisations, local agreement should be obtained in accordance with the procedures of the local participating non-NHS organisation.

User Feedback

The Health Research Authority is continually striving to provide a high quality service to all applicants and sponsors. You are invited to give your view of the service you have received and the application procedure. If you wish to make your views known please use the feedback form available on the HRA website: http://www.hra.nhs.uk/about-the-hra/governance/quality-assurance/.

HRA Training

We are pleased to welcome researchers and research management staff at our training days – see details at http://www.hra.nhs.uk/hra-training/.

Your IRAS project ID is 170258. Please quote this on all correspondence.

Yours sincerely

HRA Assessment team

Email: hra.approval@nhs.net

Copy to: *Ms Smaragda Agathou, Joint Research Office, UCL*

St George's University of London

St George's University Hospitals NHS
NHS Foundation Trust

St George's Joint Research and Enterprise Office
Ground Floor, Hunter Wing, St George's University of London,
Cranmer Terrace, Tooting, London SW17 0RE

Date: 05/09/2017

Letter of access for research for Anna Isherwood– 15.0230 Effort test performance in brain injury

This letter should be presented to each participating organisation before you commence your research at that site. The participating organisation is: **St George's University Hospitals NHS Foundation Trust.**

In accepting this letter, each participating organisation confirms your right of access to conduct research through their organisation for the purpose and on the terms and conditions set out below. This right of access commences on **05/09/2017** and ends on **02/11/2017** unless terminated earlier in accordance with the clauses below.

You have a right of access to conduct such research as confirmed in writing in the letter of permission for research from **St George's University Hospitals NHS Foundation Trust.** Please note that you cannot start the research until the Principal Investigator for the research project has received a letter from us giving confirmation from the individual organisation of their agreement to conduct the research.

The information supplied about your role in research at the organisation has been reviewed and you do not require an honorary research contract with the organisation. We are satisfied that such pre-engagement checks as we consider necessary have been carried out. However the final decision rests with St George's, University of London. Evidence of checks should be available on request to the organisation.

You are considered to be a legal visitor to the organisations premises. You are not entitled to any form of payment or access to other benefits provided by the organisation or this organisation to employees and this letter does not give rise to any other relationship between you and the organisation, in particular that of an employee.

While undertaking research through the organisation(s) you will remain accountable to your substantive employer but you are required to follow the reasonable instructions of your line manager, **Dr Shai Betteridge** or those instructions given on their behalf in relation to the terms of this right of access.

Where any third party claim is made, whether or not legal proceedings are issued, arising out of or in connection with your right of access, you are required to co-operate fully with any investigation by **St George's University Hospitals NHS Foundation Trust** in connection with any such claim and to give all such assistance as may reasonably be required regarding the conduct of any legal proceedings.

You must act in accordance with **St George's University Hospitals NHS Foundation Trust** policies and procedures, which are available to you upon request, and the Research Governance Framework.

You are required to co-operate with **St George's University Hospitals NHS Foundation Trust** in discharging its/their duties under the Health and Safety at Work etc Act 1974 and other health and safety legislation and to take reasonable care for the health and safety of yourself and others while on **St George's University Hospitals NHS Foundation Trust** premises. You must observe the same standards of care and propriety in dealing with patients, staff, visitors, equipment and premises as is expected of any other contract holder and you must act appropriately, responsibly and professionally at all times.

If you have a physical or mental health condition or disability which may affect your research role and which might require special adjustments to your role, if you have not already done so, you must notify your employer and each organisation prior to commencing your research role at that organisation.

You are required to ensure that all information regarding patients or staff remains secure and *strictly confidential* at all times. You must ensure that you understand and comply with the requirements of the NHS Confidentiality Code of Practice and the Data Protection Act 1998. Furthermore you should be aware that under the Act, unauthorised disclosure of information is an offence and such disclosures may lead to prosecution.

You should ensure that, where you are issued with an identity or security card, a bleep number, email or library account, keys or protective clothing, these are returned upon termination of this arrangement. Please also ensure that while on the **St George's University Hospitals NHS Foundation Trust** premises you wear your ID badge at all times, or are able to prove your identity if challenged. Please note that the organisation does not accept responsibility for damage to or loss of personal property.

St George's University Hospitals NHS Foundation Trust may revoke this letter and any organisation may terminate your right to attend at any time either by giving seven days' written notice to you or immediately without any notice if you are in breach of any of the terms or conditions described in this letter or if you commit any act that we reasonably consider to amount to serious misconduct or to be disruptive and/or prejudicial to the interests and/or business of **St George's University Hospitals NHS Foundation Trust** or if you are convicted of any criminal offence. You must not undertake regulated activity if you are barred from such work. If you are barred from working with adults or children this letter of access is immediately terminated. Your employer will immediately withdraw you from undertaking this or any other regulated activity and you MUST stop undertaking any regulated activity immediately.

Your substantive employer is responsible for your conduct during this research project and may in the circumstances described above instigate disciplinary action against you.

No organisation will indemnify you against any liability incurred as a result of any breach of confidentiality or breach of the Data Protection Act 1998. Any breach of the Data Protection Act 1998 may result in legal action against you and/or your substantive employer.

If your current role or involvement in research changes, or any of the information provided in your Research Passport changes, you must inform your employer through their normal procedures. You

St George's University of London
St George's University Hospitals NHS
NHS Foundation Trust
St George's Joint Research and Enterprise Office
Ground Floor, Hunter Wing, St George's University of London,
Cranmer Terrace, Tooting, London SW17 0RE

must also inform your nominated manager in each participating organisation and JREO in this organisation.

Yours sincerely

Clinical Research Governance Officer
JREO

cc: HR department of the substantive employer
HR – St Georges

Homerton University Hospital NHS
NHS Foundation Trust

Research & Development
Chair: Dr Claire Gorman

Christine Mitchell-Inwang
Research & Development Manager
Christine.Inwang@homerton.nhs.uk

Homerton University Hospital
Research and Development
Yellow Roof Top Office
Homerton Row
London
E9 6SR

Tel: 020 8510 5134
Fax: 020 8510 7850
www.homerton.nhs.uk

Dr Sanjay Sunak
Clinical Neuropsychologist
Homerton University Hospital NHS Foundation Trust
Homerton Row
London E9 6SR

15th October 2015

Dear Dr Sunak,

Re: Effort test performance in an NHS acquired brain injury sample

R&D No: 1544

Thank you for sending all the relevant documents for Homerton University Hospital Trust Research and Development Approval of the above research study. As part of the Research and Development approval process we have conducted a site specific assessment for this study. I am happy to inform you that the Trust has approved the conduct of the study and that the Trust will indemnify against negligent harm that might occur during the course of this project.

The following main document/s has been received by R&D department as part of the approval process;

Protocol Version 1.2	Dated: 24/03/2015
Participant Information Sheet Version 3	Dated: 15/05/2015
Consent Form Version 3	Dated: 15/05/2015

All other document/s you have sent in as part of the process has also been received.

I would like to draw your attention to the following conditions of the approval of this research project with which you must comply. **Failure to do so may result in the Trust withdrawing R&D approval which allows you to conduct this research project at Homerton University Hospital NHS Foundation Trust.**

Untoward events - Should any untoward event occur it is **essential** that you complete a clinical incident form and write on the form '**R&D**'. Contact the R&D Office immediately and if patients or staff are involved in an incident you must also contact the Risk Manager on 020 8510 7649.

Status of Research - Inform us if your project is amended or if your project terminates early/requires an extension as well as informing the Research Ethics Committee. This is necessary to ensure that your indemnity cover is valid and also helps the office to maintain up-to-date records. A copy of any publications arising from the research should be sent to the

R&D Office for use in the R&D Annual Report. Please be reminded that this hospital should be acknowledged in any publication.

Research Information - You will be required to complete a project update as required by the R&D Office to ensure that we have up to date information so that we can send accurate reports to the DoH and research networks. The project update form will be emailed or sent to you by the R&D Office.

Research Governance - As part of research governance, all investigators accessing identifiable personal information are required to comply with current data protection requirements.

Intellectual Property - If you believe that protectable intellectual property may arise from your research, please contact the Christine Mitchell-Inwang, R&D Manager on ext 5134 who will advise you on the proper course of action.

Monitoring of Studies – You must comply with the Trust's legal responsibility as host of this research project to monitor and audit the research to ensure that the Research Governance Framework and Good Clinical Practice (GCP) if applicable is being adhered too. Monitoring questionnaires will be sent to you and random audit visits will also take place across the trust and will be conducted following at least a seven day notice period. **Failure to respond to any of these monitoring or auditing requests may result in the Trust withdrawing your R&D approval to conduct this research at Homerton University Hospital NHS Foundation Trust.**

Please note that all NHS and social care research is subject to the DoH *Research Governance Framework*. If you are unfamiliar with the standards contained in this document, you may obtain details from the Trust R&D Office or from the DoH website (www.dh.gov.uk).

Please do not hesitate to contact Christine Mitchell-Inwang, Research and Development Manager or me if you have any further questions.

Yours sincerely,

Dr Claire Gorman
Director Research & Development

Appendix 3:

Participant information sheets

/

Participant Information Sheet

Neuropsychological test performance in an NHS acquired brain injury sample

We invite you to take part in a research study

- Before you decide to take part, it is important for you to understand the research and what it involves
- Please read the following information carefully. Discuss it with friends and family if you wish
- You are free to decide whether or not to take part in this study. If you choose not to take part, this will not affect the clinical care you receive
- Please ask us if anything is not clear or if you would like more information

Important things you need to know

- We would like to investigate whether there are specific factors that influence how well people do on neuropsychological tests. To date, very few studies have looked at this issue in UK NHS patients: this project aims to develop our understanding of this area
- The study will fit into your planned assessment
- You will be asked to complete some additional tests (no more than 30 minutes on top of your planned assessment) and consent for your data to be used in the study. No further time commitment is required.
- All your data will be anonymised so not traceable back to you
- You can withdraw from the study at any point

Part 1: The study and your involvement

What is the purpose of the study?

This study aims to look at how people with acquired brain injuries perform on different types of neuropsychological tests. We are particularly interested in whether there are specific factors that influence how well people do on the tests. Most of the existing research in this area has been conducted in American medicolegal settings, where the patients are quite different to those being treated within the NHS. We hope that the findings will inform UK clinicians about factors that might affect performance on neuropsychological testing and so potentially impact upon the accuracy of test results in this client group.

This study forms the major part of a doctoral thesis in Clinical Psychology (DClinPsy) at University College London. The Researcher is Anna Isherwood. The study has been approved by the NHS Research Ethics Committee and the Research and Development Departments at Homerton University NHS Foundation Trust and St George's Healthcare NHS Trust.

Why have I been invited?

You have been invited because you are going to have a neuropsychological assessment as part of your clinical care. All participants will have been diagnosed with an acquired brain injury (traumatic brain injury, stroke, tumour or other). It is planned that over 50 individuals will take part in this study.

Do I have to take part?

It is up to you if you decide to join the study. We will describe the study and go through the information sheet. If you have any questions, please ask your clinician or the Research Student (details below) before the study begins. If you agree to take part we will ask you to sign a consent form. This will not affect the standard of care you receive.

What will happen to me if I take part?

If you consent to take part in the study, you will have a neuropsychological assessment as planned by your clinician. A neuropsychological assessment comprises different tasks that check your abilities like memory and concentration. As part of this study, you will be asked to undergo some additional brief tests as part of this assessment. We anticipate that these extra tests will add approximately 20-30 minutes

on to the total assessment time. There is no further time commitment required. You will be given feedback on your assessment results as normal within the service.

As part of the study you will also be asked to agree to your data (the scores on the different tests, and personal details from your medical records such as your age and diagnosis) being used for research purposes. This data will be anonymised and confidential (see Part 2, below) so will not be traceable back to you.

What are the possible disadvantages or risks of taking part?

Neuropsychological tests can be quite tiring. Some people can find them stressful. We will check in with you to make sure you are not feeling too tired, stressed or uncomfortable in any way. You are very welcome to take a break during testing and you are free to stop at any point.

What are the possible benefits of taking part?

Although we cannot promise that the study will help you directly, the information we get from this study will help to improve the assessment and treatment of people with acquired brain injury.

Will I be paid for taking part?

Your participation in this study is entirely voluntary.

Will my taking part in the study be kept confidential?

Yes. We will follow ethical and legal practice and all information about you will be handled in confidence.

If the information in Part 1 has interested you and you are considering taking part, please read the additional information in Part 2 before making any decision.

Part 2: Additional information

What will happen if I want to withdraw from the study?

You are free to withdraw at any time, and have the right to ask that any data you have supplied to that point be withdrawn or destroyed. This will not affect the standard of care you receive.

What if I have a concern or complaint?

If you wish to complain, or have any concerns about any aspect of the way you have been approached or treated by members of staff you may have experienced due to your participation in the research, National Health Service or UCL complaints mechanisms are available to you. Please ask your clinician or the Research Student, Anna Isherwood, if you would like more information on this.

In the unlikely event that you are harmed by taking part in this study, compensation may be available. If you suspect that the harm is the result of the Sponsor's (University College London) or the hospital's negligence then you may be able to claim compensation. After discussing with your research doctor, please make the claim in writing to Dr John King, who is the Chief Investigator for the research and is based at University College London. The Chief Investigator will then pass the claim to the Sponsor's Insurers, via the Sponsor's office. You may have to bear the costs of the legal action initially, and you should consult a lawyer about this.

Will my taking part in this study be kept confidential?

Yes. Only your clinician and the Research Student (Anna Isherwood) will have access to any identifiable data during this study. Following your assessment, any identifiable information (such as your name) will be anonymised and replaced with a code so you data will not be traceable to you. All physical data (such as paper test forms) will be managed and stored in accordance with local NHS Trust data protection policy. This means that it will be kept within a locked filing cabinet and only accessed with prior approval from the lead clinician on-site. In accordance with UCL policy all personal and/or sensitive personal data (as defined by the Data Protection Act 1998) will be securely destroyed at the conclusion of the research. Non-identifiable data and other records not containing person identifiable data may be retained for a longer period at the discretion of the Chief Investigator.

Who is organising and funding the research?

This study is organised and funded by University College London.

Who has reviewed the study?

All research in the NHS is looked at by an independent group of people, called a Research Ethics Committee (REC), to protect your interests. This study has been reviewed and given a favourable opinion by City and East Research Ethics Committee (REC reference: 15/LO/1376). This study has also been peer-reviewed by senior academic colleagues in the Research Department of Clinical, Educational and Health Psychology at UCL.

Further information and contact details

For further information about the research project:
Anna Isherwood
Trainee Clinical Psychologist
Research Department of Clinical, Educational and Health Psychology
University College London
1-19 Torrington Place
London
WC1E 7HB
Tel: 020 7679 1897

Dr Sanjay Sunak
Principal Clinical Neuropsychologist
Integrated Medical and Rehabilitation Services Division
St Leonard's Hospital
Nuttall Street
London
N1 5LZ
Tel: 020 7683 4489

If you wish to make a complaint about the study:
Patient Advice and Liaison Service (PALS)
Homerton University Hospital NHS Foundation Trust
Homerton Row
London
E9 6SR
020 8510 7315

Dr John King
Senior Lecturer
Research Department of Clinical, Educational and Health Psychology
University College London
1-19 Torrington Place
London
WC1E 7HB
Tel: 020 7679 5993

Neuropsychological test performance in an NHS acquired brain injury sample

Information Sheet

Why are we doing this study?

Neuropsychological tests look at your different thinking skills (like memory and concentration)

We want to understand more about the different factors that might influence results on these tests

This will help to improve the accuracy of our assessments in the future

Why me?

- You have an acquired brain injury
- You need a neuropsychological assessment as part of your care

  OR
- You have expressed an interest in taking part

What happens if I say "YES"?

You will have a neuropsychological assessment. This involves:

1) Answering questions about yourself…

2) Doing tests of your thinking skills…

Pen and paper tests 

Tests on a computer 

Puzzles 

Altogether this will take about 2-3 hours

If you are having an assessment already as part of your clinical care, the extra tests will add 20-30 minutes on to the total length of your assessment.

What happens to my information?

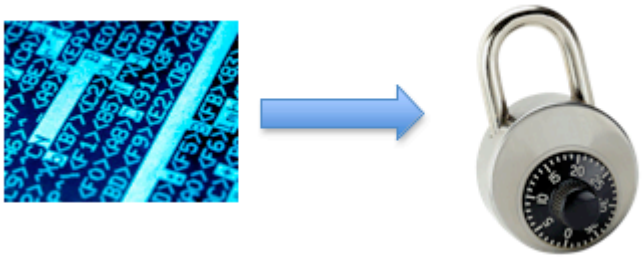
We ask that you agree for your data to be used for research. This data includes:

1) Test scores from your assessment
2) Personal details (like age and diagnosis) from your medical records

Only your clinician and the Researcher will be able to see this data.

Your information will then be anonymised. This means that your details are turned into a code and kept secret. If the research is published no one will be able to identify you.

All the data will be kept locked away.



Are there any risks?

The tests are not painful. You won't be asked to do anything that puts you at risk.

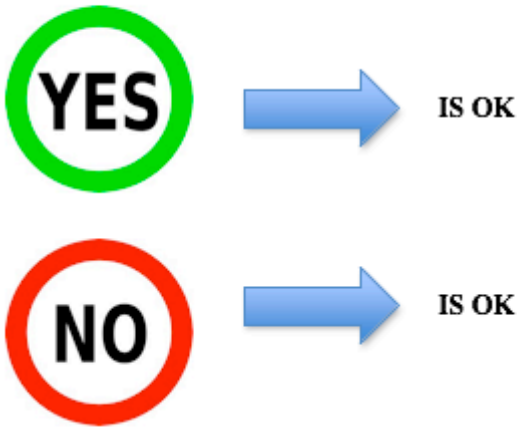
Sometimes the tests can be tiring. Other people might find them stressful.

Let us know if you are uncomfortable in any way. You can take a break or stop the testing at any time.

If you stop the testing this will not affect your clinical care in any way.

What are my rights?

You DO NOT have to take part in this study.



Ask us if you have any questions!

Appendix 4:

Consent forms

Homerton University Hospital NHS
NHS Foundation Trust

Study Number:

Participant Identification Number for this trial:

CONSENT FORM

Title of Project: Neuropsychological test performance in an NHS acquired brain injury sample

Name of Researcher: Anna Isherwood

Please initial box

1. I confirm that I have read the information sheet dated..... (version.....) for the above study. I have had the opportunity to consider the information, ask questions and have had these answered satisfactorily. ☐

2. I understand that my participation is voluntary and that I am free to withdraw my participation or my data at any time without giving any reason, without my medical care or legal rights being affected. ☐

3. I understand that relevant sections of my medical notes and data collected during the study may be looked at by individuals from University College London, from regulatory authorities or from the NHS Trust, where it is relevant to my taking part in this research. I give permission for these individuals to have access to my records. I understand that personal data will be handled in accordance with the UCL Information Governance Policy and the Data Protection Act. ☐

4. I agree to take part in the above study. ☐

Name of Participant	Date	Signature
Name of Person taking consent	Date	Signature

When completed: 1 for participant; 1 for researcher site file; 1 (original) to be kept in medical notes.
Participant Consent Form (A). Version 4. August 2015.
Page 1 of 1.

Neuropsychological test performance in an NHS acquired brain injury sample

Consent form

Please read this with the Participant Information Sheet

OR

Yes **No**

I have been supported to understand the information sheet dated ..... (version.....).

I understand what I have to do to be part of the study.

I agree to my personal details and neuropsychological data being used for the study. This means that sections of my medical notes may be looked at by the Researcher.

I understand that all my data will be made anonymous and kept confidential.

I understand that I can say NO if I want to and that this is OK.

I know that I can say STOP at any time and this will not affect my treatment.

I have been able to ask questions.

Signature of participant_____

Date_____

I confirm that I have explained the nature of the study as detailed in the information sheet in terms which, in my judgement, have been understood by the participant.

Signature of person taking consent_____

Date_____