JESSICA L. HOOKER BSc Hons

AN EXPLORATION INTO RESPONSE VALIDITY

Section A: Assessment of Response Validity: A Systematic Review of Clinician Beliefs and Practices

Word Count: 7944 (plus 291 additional words)

Section B: Response Validity and Psychological Functioning in a UK NHS Acquired Brain Injury Sample

Word Count: 7960 (plus 397 additional words)

Overall Word Count: 15904 (plus 688 additional words)

A thesis submitted in partial fulfilment of the requirements of Canterbury Christ Church University for the degree of Doctor of Clinical Psychology

SEPTEMBER 2018

SALOMONS
CANTERBURY CHRIST CHURCH UNIVERSITY

Acknowledgements

I am firstly extremely grateful to the participants and staff within the NHS service involved, without whom this study would not be possible. Particular thanks also go to Anna Isherwood, Aline Hardwick, and Kelly Llanfear for their support in completing the archival database. The contributions of my supervisors, Dr Monika Hunter and Dr Ndidi Boakye, as well as Dr Sabina Hulbert and Dr Holly Milling, are also greatly appreciated. Finally, my thanks go to my partner, Luke, and my family and friends (especially my Comma-Tamer) for their support throughout all stages of this thesis; truly a team ‘effort’!

Summary of the MRP Portfolio

Section A

A systematic literature review that aimed to explore the practice and beliefs of clinicians in relation to symptom and performance validity testing, following its endorsement by international professional bodies. Fourteen survey studies indicated that validity issues were reported in a substantial minority of assessments across medico-legal, forensic and clinical settings. Validity test use appears to be increasing, although the majority of clinicians reported to rely upon clinical judgement in their assessments, despite established research indicating its limited utility in detecting response invalidity. Clinical and research implications are discussed, particularly in light of the literature being dominated by North America.

Section B

An empirical study exploring performance and symptom validity in an NHS outpatient neuropsychology population. Archival data (N = 127) revealed a base rate of performance validity test (PVT) failure of up to 18%. A significant relationship was found between performance and symptom validity, as measured by the Personality Assessment Inventory (PAI). Furthermore, elevations in reported psychopathology were found in the PVT fail group compared to those who passed. Group differences in terms of demographic variables are explored. Findings are discussed in the context of existing literature and recommendations are made for future validity testing research, as well as clinical practice.

Section C

Appendices of supporting material.

Table of Contents

List of Tables and Figures

List of Abbreviations

PVT	Performance validity test
SVT	Symptom validity test
SPVT	Symptom or performance validity test
TOMM	Test of Memory Malingering
DS-SS	Digit Span age-corrected scaled score
ABI	Acquired brain injury
TBI	Traumatic brain injury
WAIS	Wechsler Adult Intelligence Scale
PAI	Personality Assessment Inventory
FSIQ	Full Scale IQ
PMIQ	Premorbid IQ

Section A

Assessment of Response Validity: A Systematic Review of Clinician Beliefs and Practices

7944 words (plus 291 additional words)

Abstract

Background: Research has shown that assessment data can be greatly affected by non-neurological dimensions of response, such as whether an individual is trying their best. Professional bodies in the USA and UK have endorsed the routine use of symptom and performance validity tests (SPVTs) to assess for this significant source of test variance. However, the impact of these recommendations on clinical practice has not yet been synthesised.

Method: Five electronic databases were systematically searched to identify studies exploring the practices and beliefs of clinicians in settings where there is opportunity for validity testing. Main findings are summarised and critically appraised.

Results: A total of 14 survey studies were included. Samples were international, although the majority were from North America. Validity issues were identified in a substantial minority of forensic and medico-legal cases, and a smaller minority of clinical assessments. The rate of SPVT use appears to be increasing, at least in secondary gain settings. However, clinical judgement continues to be relied upon by the majority, despite established research indicating its limited utility in detecting response invalidity.

Conclusions: There was variability in practitioners' adherence to professional recommendations regarding performance and symptom invalidity, although it appears that North America has progressed furthest in the field. Clinical and research implications are discussed.

Key words: Performance validity, symptom validity, clinician, practice, adherence

Assessment of Response Validity: A Systematic Review of Clinician Beliefs and Practices

Psychologists have historically used standardised instruments with the intention of measuring brain function and symptoms; however, test data is only valid if the examinee exerts adequate effort. There now exists an established evidence base indicating the presence of significant test data not fully explained by the brain condition itself. Green, Rohling, Lees-Haley and Allen (2001) showed that in the presence of compensation incentives, more than half of the statistical variance in neuropsychological test scores was explained by examinee 'effort', in contrast to just 11% explained by education, and 4% by age. Furthermore, global neuropsychological functioning was found to be suppressed 4.5 times more by suboptimal effort than severity of the brain injury. The authors highlight the vast implications of overlooking this source of variance. These include inappropriate diagnoses, treatment and social support, whilst also potentially unjustly impacting those with genuine impairments and service resources.

The ability of clinicians to detect suboptimal effort using clinical judgement has come under scrutiny. In a classic paper, Faust, Hart, Guilmette and Arkes (1988) found that 0% of their neuropsychologist sample were able to identify the profiles of simulators among those with genuine brain injuries. Even when informed that the base rate of malingering in the data was 50%, identification accuracy remained at chance level. Furthermore, the vast majority of the neuropsychologists indicated that they were highly or very highly confident in their judgement.

Research in the field has historically focused on forensic or litigating samples with incentive to 'fake bad', simply conceptualising examinee effort as malingering for financial gain (McMillan et al., 2009). Performance invalidity, symptom invalidity, suboptimal effort, response bias, dissimulation and malingering are terms used interchangeably in the literature (Bigler, 2012). Definitions of these terms can be found in Appendix A. The current thesis will

report research findings using the terms chosen by the authors to ensure the meaning is not altered; otherwise, symptom or performance validity will be used as these are considered the most accurate and least stigmatising descriptions (Larrabee, 2012).

Symptom and Performance Validity Tests

Validity tests attempt to offer a more objective method of assessing the validity of assessment data. Performance validity tests (PVTs) refer to the assessment of validity of performance on cognitive tasks, and symptom validity tests (SVTs) refer to the assessment of the validity of self-reported symptoms. SPVT will refer to both symptom and performance validity tests.

Many PVTs employ a very easy forced choice verbal recognition memory task that appears to the examinee to be more difficult, for example, the Test of Memory Malingering (TOMM, Tombaugh, 1996). Scores below chance are thought to be suggestive of malingering due to the likelihood of purposeful selection of incorrect items. Scores below a cut-off based upon normative data of known clinical groups are suggested to indicate invalid performance, but not intent to feign (Slick, Sherman & Iverson, 1999). PVTs should not be sensitive to general intellectual functioning, age, education, or brain condition (Green & Merten, 2013).

SVTs are concerned with the degree to which symptomatic complaint on self-report measures is reflective of ‘true’ symptoms, and are usually in the form of mood or personality inventories (such as the Minnesota Multiphasic Personality Inventory-II; Butcher, Dahlstrom, Graham, Tellegen & Kreammer, 1989).

As well as ‘stand-alone’ tests that have been developed to evaluate performance or symptom validity, the use of ‘embedded’ measures within existing tests can be used to identify invalid responding without increasing testing time.

SPVTs, like all neuropsychological instruments, possess imperfect psychometric qualities which vary across tests. It has been argued that these measures should prioritise specificity over sensitivity, since false positive errors may have less devastating consequences than false negatives (Greve & Bianchini, 2004). Vickery and colleagues (2001) found the average level of specificity across five PVTs to be 96%, but the average sensitivity was just 56%. This may be a factor in clinicians' use of SPVTs in clinical practice. The multivariate failure model (Larrabee, 2003) was proposed to address poor sensitivity rates, whereby failure on two or more PVTs can be understood as indicating invalid responding. This model has been found to produce good sensitivity and specificity in discriminating credible performance (Victor, Boone, Serpa, Buehler, & Ziegler, 2009). However, the true prevalence in a population, or base rate, is required for more accurate interpretation (McMillan et al., 2009).

Professional Recommendations

Position papers released in North America by the National Academy of Neuropsychology (Bush et al., 2005) and the American Academy of Clinical Neuropsychology (AACN; Heilbronner et al., 2009) have suggested that response invalidity is present in a sizeable minority of neuropsychological examinees, with higher base rates in secondary gain contexts (such as forensic settings). Both guidelines consider the inclusion of SPVTs to be a 'medical necessity', and to be the most valid approach in detecting response invalidity. Validation studies have established that more recently developed stand-alone SPVTs such as the TOMM possess sound psychometric properties in comparison to embedded tests (Tombaugh, 1996; Rees, Tombaugh, Gansler & Moczynski, 1998). Consequently, professional bodies have recommended a multi-method, multi-test approach, utilising both stand-alone and embedded measures. The use of SPVTs should also depend

upon the context of the assessment and patient factors (although if not employed clinicians should provide rationale as to why they were not utilised). In addition, clinicians should encourage examinees to give their best effort, and performance on SPVTs should be documented within reports.

Guidance in the UK (McMillan et al., 2009) has suggested that, in line with the USA, SPVTs should be routinely included in neuropsychological assessments (in both forensic and clinical settings). The recommendations offer some limited definitive advice to clinicians, such as the utility of employing both stand-alone and embedded measures, to advise examinees to try their best and that this will be assessed, and to report carefully on SPVT results (e.g. 'effort testing failed to indicate non-credible performance'). However, emphasis is made on the need for further research in the UK. More recently, further guidance from the AACN was released concerning the use of SPVTs in disability evaluations (Chafetz et al., 2015), which recommended their use in assessing pain complaints. To date, there have been no systematic reviews synthesising findings on validity testing practices, despite clear endorsement by international professional bodies. Without an understanding of clinical practice, the impact of guidance, as well as outstanding needs within the profession, remain unknown.

Review Aims

This review aimed to investigate the beliefs and practices of psychologists in relation to symptom and performance validity testing 12 years on from the influential US position paper (Bush et al., 2005), and eight years following the release of guidance in the UK (McMillan et al., 2009). The review will focus on a target population of clinicians who work in settings wherein there is opportunity for validity testing.

Main findings in light of methodological issues will be presented and synthesised, producing implications for future research and clinical practice.

Method

Literature Search

A total of five electronic databases were searched: PsycINFO, Medline, ASSIA, ERIC (EBSCO) and Web of Science.

Table 1

Electronic Search Strategy and Key Search Terms

Category 1: Clinicians		Category 2: Beliefs and practices		Category 3: Validity
psychologist* or clinical psychologist* or neuropsychologist* or expert* or clinician*	**AND**	attitude* or belief* or practice*	**AND**	effort test* or malinger* or symptom validity or performance validity or validity test* or response bias

*denotes truncation, looks for variants of words such as malinger, malingerer, malingering.

Study Selection

A flow diagram illustrating retrieved papers following application of search limitations (English language, peer-reviewed) is presented in Figure 1. These limits were applied to improve the quality of studies and to produce an appropriate amount of data for the current review. Titles and then abstracts were screened for eligibility. Several relevant journals (The Clinical Neuropsychologist, Archives of Clinical Neuropsychology, and Clinical Psychology Review) and reference lists of identified studies were searched for additional papers. Experts in the field were consulted regarding any outstanding papers and, finally, Google Scholar was used to hand-search for remaining literature.

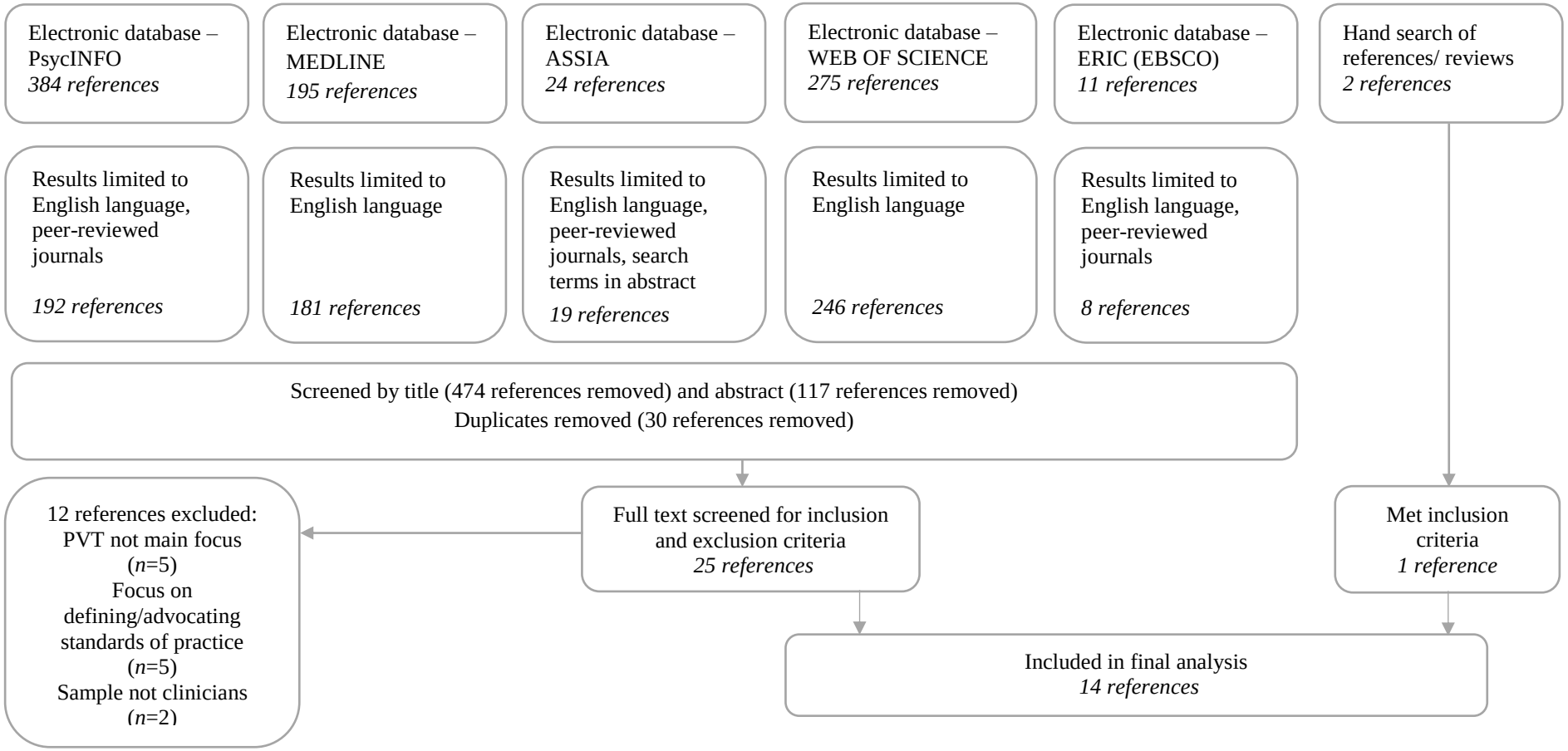

Figure 1. Flow Diagram Illustrating Literature Search

As demonstrated in Table 2, eligibility criteria were kept broad so as to include a diverse range of studies and produce more generalizable findings.

Table 2

Eligibility Criteria

Inclusion criteria	Exclusion criteria
Peer-reviewed journal paper	Not written in English language
Original empirical paper	Focus on defining or advocating validity testing standards of practice
Primarily focused on the assessment of response validity	Sample not clinicians

Literature Review

Table 3 introduces the 14 papers included in the final analysis. A more detailed summary table of findings is provided in Appendix B.

Table 3

Summary of Included Studies

Study	Participants	Design	Setting
Hirst et al. (2017)	*N* = 654 Licensed clinical psychologists with ≥100 post-graduate neuropsychological assessment hours; 21% were board certified in neuropsychology Mean years of practice = 16	Online survey emailed to National Academy of Neuropsychology (NAN) and International Neuropsychological Society (INS) members 9% response rate	90% USA/Canada, 10% international countries Approximately half sample conducted assessments in secondary gain contexts such as forensic work, disability claimants, and VA hospitals
Brooks, Ploetz, & Kirkwood (2016)	*N* = 282 Neuropsychologists working with children/adolescents Mean years of practice = 12	Online survey emailed via professional listservs, open for eight weeks	USA/Canada Majority of sample conducted clinical assessments but a third also conducted forensic assessments Mean number of assessments performed annually = 102
Schroeder, Martin, & Odland (2016)	*N* = 24 Neuropsychologists/experts (defined as being first author on four recent papers regarding validity testing or participation in the AACN response validity conference) Mean years of practice = 20	Online survey emailed to identified experts 50% response rate	USA 92% conducted clinical assessments and 91% forensic assessments 87% primarily evaluated adults and 12% worked across the lifespan
Young, Roper, & Arentsen (2016)	*N* = 172 Psychologists working with the Veterans Affairs healthcare system and likely practising neuropsychology	Email survey, open for one month 44% response rate	USA - Veterans Affairs healthcare system 43% conducted forensic assessments 16% board certified Mean number of assessments yearly = 155
Barker-Collo & Fernando (2015)	*N* = 73 Registered psychologists 89% of sample self-identified as clinical or educational	Advertised through the New Zealand Psychological Society and the New Zealand College of Clinical	New Zealand Most respondents were clinicians working for the Accident Compensation Corporation (ACC) or

	psychologists and others were 'generalists' Mean years of practice = 14 years	Psychology and provided weblink for online survey; open for four months	privately; minority also conducted medico-legal assessments
Martin, Schroeder, & Odland (2015)	*N* = 316 Licensed neuropsychologists who primarily assess adults Mean years of practice = 12	Online survey based on previous surveys, sent via professional neuropsychology email listservs and open for approximately three weeks	USA Majority (33%) worked in private practice, and 73% of sample did at least some forensic assessments
Allcott et al. (2014)	*N* = 73 Multi-disciplinary experts at consultant level (psychologists, psychiatrists, orthopaedic specialists, neurologists, & occupational therapists)	Emailed to members of the Directory of Expert Witnesses as well as other known experts; open for six months 25% response rate	UK Medico-legal settings
Dandachi-Fitzgerald, Ponds, & Merten (2013)	*N* = 515 96% psychologists, 3% physicians Mean years of practice = 10	Email survey sent to chairs of each of the European Societies of Neuropsychology to forward to respective members; six of 12 societies agreed to participate Survey open for 18 months Range of 6-25% response rates	Surveyed 6 European countries (Germany, Italy, Denmark, Finland, Norway, Netherlands) 95% conducted clinical assessments, 55% undertook forensic work Median assessments conducted in previous year = 70
McCarter, Walton, Brooks, & Powell (2009)	*N* = 130 Psychologists and neuropsychologists	Survey emailed to members of the BPS Division of Neuropsychology 22% response rate	UK 70% of sample conducted both clinical and forensic assessments, 29% solely clinical and 1% solely forensic
Sharland & Gfeller (2007)	*N* = 188 Clinical neuropsychologists (30% board certified in neuropsychology) Mean years of practice = 17	Paper surveys mailed to a random sample of approximately one third of NAN professional members 26% response rate	USA Unknown practice settings or proportion of clinical/forensic assessments conducted

Boccaccini, Boothby, & Overduin (2006)	*N* = 116 Pain specialists and clinical-forensic psychologists 34% had specialised training in forensic and pain assessment	Mailed questionnaire (including a vignette describing an attorney-referred case) to members of the American Pain Society, and relevant divisions of the American Psychological Association Asked respondents whether they would attempt to assess response validity, and methods they would use 18% response rate	USA Personal injury and medico-legal settings
Sullivan, Lange, & Dawes (2006)	*N* = 17 Members of the Australian Psychological Society, College of Clinical Neuropsychology and delegates from two Australian neuropsychology conferences Mean years of practice = 13 years	Emailed an online survey	Australia 64% worked in private practice settings 60% of respondents' work was clinical assessment and the remainder forensic work
Slick, Tan, Strauss, & Hultsch (2004)	*N* = 24 Neuropsychologists who were identified as experts in the area of civil litigation through their publication history 55% board certified in neuropsychology Mean years of practice = 15	Survey completed via email or over the telephone across a three month period 61% response rate	USA Clinical and medico-legal settings Majority (71%) had undertaken >20 assessments in the previous year
Mittenberg, Patton, Canyock, & Condit (2002)	*N* = 144 Members of the American board of neuropsychologists who were listed as actively practising as neuropsychologists Mean years of practice = 18	Paper surveys were mailed 37% response rate	USA/Canada Respondents engaged in both clinical and medico-legal/forensic work Mean number of assessments undertaken yearly = 252

Design of Studies

All studies employed a survey design to investigate beliefs and practices regarding SPVTs. Most developed an idiosyncratic questionnaire using software such as SurveyMonkey™ (SurveyMonkey Inc., 2015) based upon previous survey research in validity test use, which was adapted to suit the target audience. Conversely, Allcott et al. (2014) and McCarter, Walton, Brooks, and Powell (2009) designed a novel questionnaire based on the authors' clinical experience. Questionnaires were reported to examine participants' demographics, training and clinical practice, use of SPVTs and rationale, and practices regarding suspected invalid performance. Hirst et al. (2017) focussed specifically on adherence to validity testing recommendations, and Boccaccini, Boothby, and Overduin (2006) used a clinical vignette to investigate participants' hypothetical SPVT approaches. Five out of the 14 studies provided access to the questionnaire used.

Participants

The majority of samples included clinical psychologists, neuropsychologists, and experts with a doctoral degree, practicing within the field of neuropsychology in at least a part-time capacity.

Schroeder, Martin, and Odland (2016), as well as Slick, Tan, Strauss, and Hultsch (2004), investigated experts in the field of neuropsychological validity testing, which was defined as identification as first author on two recent papers regarding validity testing. Young, Roper, and Arentsen (2016) sampled psychologists employed within the Veterans Affairs (VA) healthcare system; a large, integrated system in the USA involving both healthcare provision and disability assessment.

Allcott et al. (2014) presented the only investigation of multi-disciplinary professionals at consultant level within personal injury settings, including psychologists,

psychiatrists, neurologists, and occupational therapists. Brooks, Ploetz, & Kirkwood (2016) sampled professionals who conducted assessments of children and adolescents, and Boccaccini et al. (2006) involved pain specialists and forensic psychologists.

Studies were predominantly based in North America and Canada. However, two were based in the UK (Allcott et al., 2014; McCarter et al., 2009), one was in Australia (Sullivan, Lange, & Dawes, 2006), and one was in New Zealand (Barker-Collo & Fernando, 2015). Additionally, Hirst et al. (2017) surveyed international neuropsychologists (although largely in the USA), and another study surveyed neuropsychologists across six European countries; Germany, Italy, Denmark, Finland, Norway and the Netherlands (Dandachi-Fitzgerald, Ponds, & Merten 2013).

The work settings of the participants varied; half of the surveys noted that respondents completed more clinical assessments than forensic, although the majority also completed some medico-legal work. The samples used by Allcott et al. (2014), Boccaccini et al. (2006), and Sullivan et al. (2006) completed only forensic, legal or disability claim cases. Barker-Collo and Fernando (2015) reported that most of their sample were employed in treatment settings in Accident Compensation Corporation (ACC)-funded and private practice in New Zealand. Similarly, the majority of the American neuropsychologists sampled by Hirst et al. (2017) practiced in settings where secondary gain was likely. In a UK study (McCarter et al., 2009), the majority (70%) of neuropsychologists reported conducting both clinical and forensic assessments, with a third completing solely clinical work and 1% solely forensic work.

All but one study investigated professionals working with adults, Brooks et al. (2016) being the only researchers exploring the use of SPVTs with children and adolescents. The vast majority of studies investigated professionals in the field of neuropsychology; however, Boccaccini et al. (2006) focussed on pain complaints in personal injury claims. Two studies

investigated the experiences of neuropsychologists working with clients presenting with a range of diagnoses such as head injury, PTSD, fibromyalgia, and mood disorders (Young et al., 2016; Mittenberg, Patton, Canyock, & Condit, 2002).

The clinical experience of the sample varied across studies, ranging from a mean of 10 years (Dandachi-Fitzgerald et al. 2013) to 20 years in the study by Slick et al. (2004), which sampled experts. Four studies did not provide data on years of experience.

Where reported, the mean number of assessments performed annually ranged from 30 (Dandachi-Fitzgerald et al., 2013) to 155 per respondent (Young et al., 2016).

Procedures

The majority of studies emailed participants a link to an online survey via broad list servers in the field, such as AACN, NPSYCH, and the British Psychological Society (Division of Neuropsychology). Dandachi-Fitzgerald et al. (2013) contacted the chairs of the European Societies of Neuropsychology and requested that they forward an email link to members of their respective societies, translated into their native languages if requested. Allcott et al. (2014) distributed their survey via email to experts identified from the Directory of Expert Witnesses, and, likewise, Slick et al. (2004) identified experts using PsycINFO searches. Sullivan et al. (2006) additionally invited conference attendees to take part and Barker-Collo and Fernando (2015) recruited their participants in professional society newsletters.

Main Findings

Base rates. Professionals' estimates of base rates of invalid performance were explored in the majority of studies. Dandachi-Fitzgerald et al. (2013) found that the base rate of 'insufficient effort' was estimated to be 10% in clinical assessments and 15% in forensic

assessments in their sample of neuropsychologists from six European countries. 'Malingering' was thought to occur in 4% of their clinical cases and 10% of forensic cases. Interestingly, a discrepancy was found between general estimated base rates of malingering and respondents' estimates of base rates in their own practice (general estimates were rated as 10% in clinical and 20% in forensic assessments). This suggests that neuropsychologists in the study believed that they personally encountered less malingering clients than their colleagues.

A quarter of Barker-Collo and Fernando's (2015) sample of psychologists predominantly working in treatment settings in ACC-funded and private practice indicated that performance invalidity issues occurred in 20-50% of cases. Hirst et al. (2017) also surveyed clinicians mostly practicing in secondary gain contexts, and found that respondents who followed validity testing recommendations reported significantly higher base rates of 'poor effort' and 'malingering' than those who did not follow all practice recommendations.

Approximately half of the experts surveyed by Slick et al. (2004) considered base rates of 'possible malingering' to be at least 10%, and a third considered rates to be at least 20%. Furthermore, two-thirds considered there to be the presence of 'definite malingering' in at least 10% of cases. This is in line with Mittenberg et al. (2002), who found base rates of 'probable malingering' and 'symptom exaggeration' were estimated to be approximately a third of disability evaluations and personal injury cases, and 8% of medical cases in their sample of North American neuropsychologists. Estimated base rates did not vary greatly across practice settings or geographic regions, or the number of assessments conducted annually.

Young et al. (2016) found that the base rate of SPVT failure as reported by neuropsychologists conducting routine outpatient clinical evaluations within the VA healthcare system was approximately 23% (three times that found by Mittenberg, et al. 2002).

The highest rate of SPVT failure was found in mild traumatic brain injury (mTBI) and PTSD disability evaluations, consistent with Mittenberg et al. (2002). Higher rates of SPVT failure were associated with clinician factors, such as practice being more focused on neuropsychological assessment, and greater frequency of SPVT usage; the less clinicians used SPVTs, the fewer failures they found (Young et al., 2016). There was no relationship between base rate of failure estimates and board certification in neuropsychology status, but there was a positive correlation between number of SPVTs used and professional organisation memberships ($p<.02$).

Allcott et al. (2014) found that 70% of their sample of UK multi-disciplinary consultants within personal injury settings indicated that three quarters of cases were 'genuine'; however, 25% considered half of their cases to be 'disingenuous'. Base rates of 'symptom exaggeration' in personal injury cases in Australia was 13%. Criminal cases received the highest estimate (17%) and medical or psychiatric the lowest (3%; Sullivan et al., 2006).

Using a relatively large and broad sample of US neuropsychologists, estimations of base rates of 'deliberate exaggeration' in medico-legal assessments was on average 20%, and 5% in cases with no obvious secondary gain (Sharland & Gfeller, 2007), similar to the findings of Mittenberg, et al. (2002) and Slick et al. (2004).

Overall, findings consistently suggested that a substantial minority of forensic and medico-legal cases and a smaller minority of clinical assessments were considered by professionals to produce invalid performances across a range of geographical locations. Estimated base rates in clinical evaluations were between 3 and 10%, in forensic cases estimates were 17-25%, and were variable in litigation settings (between 8-30%, the lowest estimates being found in New Zealand and the highest in North America).

The literature suggests that professionals' views of base rates have remained relatively stable over time, with medico-legal case estimates around 20-30% in 2002, to approximately 20% in 2007, 25% in 2004, and between 18-25% in 2017. Base rates of clinical assessments were estimated around 8% in 2002, 3% in 2005, 5% in 2007, and 10% in 2013.

Methods. Although base rates were relatively comparable, methods employed to assess response validity varied throughout the papers. The respondents investigated by Brooks et al. (2016) considered the assessment of validity to be 'multi-factorial', but mostly relied upon behavioural observations and discrepancies between self-report and records in their evaluations of children and adolescents. Furthermore, despite scores below cut-offs on stand-alone SPVTs receiving the most empirical support, these were only the seventh most popular method. This may indicate that the respondents based their decisions on factors other than independent research.

Similarly, Barker-Collo and Fernando (2015) found that their sample of registered psychologists in New Zealand was most likely to use clinical judgement to assess for performance validity (47%), with only 38% reporting use of embedded SPVTs.

Martin et al. (2015) found that when there was a discrepancy between SPVTs and qualitative measures of validity, the majority would have more confidence in SPVT results but a significant minority (13%) would give more weight to clinical judgement. The greatest proportion of respondents (35%) in this study indicated that they considered two or more 'failures' on PVTs to indicate questionable validity.

Both Sullivan et al. (2006) and Mittenberg et al. (2002) found that approximately two-thirds of their samples endorsed qualitative methods of assessing validity, such as inconsistencies in pattern of performance, severity of cognitive impairment, self-report and documented condition, whilst around half relied upon scores below cut-offs on SPVTs. Schroeder et al. (2016) similarly found that their sample used a broad range of methods to

assess validity, integrating both quantitative and qualitative methods. The majority of experts have also been found to rely upon discrepancies between self-reports and medical history, and complaints inconsistent with the severity of the condition (Allcott et al., 2014).

In a sample of specialists conducting pain assessments, the most popular method for assessing the validity of pain symptoms was to review collateral information in relation to pain symptoms, such as medical records, observations of pain-related behaviours and discrepancies between pain complaints and test data (Boccaccini et al., 2006).

Overall, the majority of studies suggested that the most commonly relied upon methods of detecting invalid responding were qualitative, and included clinical judgement in relation to inconsistencies between pattern of performance and condition, implausible self-reported symptoms, and inconsistencies between severity of symptoms and condition.

Frequency and manner of SPVT use. The vast majority of the studies reviewed reported on the frequency with which SPVTs were used by their respective samples of practitioners.

Around half of a sample of North American neuropsychologists reported they often or always included an SPVT in their assessments (Sharland & Gfeller, 2007); however, this had increased to 92% in a similar sample eight years later (Martin et al., 2015). The latter study also found that the use of embedded measures was more than 14 times greater than in Sharland and Gfeller's (2007) sample previously. Furthermore, the likelihood of using stand-alone measures was more than six times greater. However, the authors found no change in the popularity of using qualitative methods to assess invalidity which remained high. Similarly, approximately two-thirds of US respondents in 2016 indicated that they used SPVTs always or frequently across clinical and forensic contexts (Young et al., 2016), including both stand-alone and embedded measures.

According to the most recent survey in 2017, those who reportedly adhered to validity testing recommendations included a greater number of SPVTs in assessments on a routine basis in comparison to those who did not meet recommendations (an average of 10 measures compared to 5.8, $p<.001$; Hirst et al., 2017). Young et al. (2016) found that when respondents used SPVTs, the majority employed at least two stand-alone or embedded measures, but there was no consensus regarding the use of one or two SPVT failures to indicate invalid performance.

Consistent with Sharland and Gfeller (2007), Dandachi-Fitzgerald et al. (2013) found that European respondents indicated a greater occurrence of invalid responding than their use of SPVTs. Despite acknowledging the prevalence of invalid performance, 69% of respondents reported they often or always based their judgements on qualitative methods such as discrepancies between self-reports, records, and condition severity. Only 11% indicated systematically using SPVTs in clinical assessments, and just 44% in forensic assessments across the whole sample. Respondents in Norway were most likely to use SPVTs (86% in the majority of the forensic assessments and 54% in the majority of clinical assessments). Respondents in Italy reported the lowest rate of SPVT use (13% and 10% in the majority of forensic and clinical assessments respectively).

In a UK sample of neuropsychologists, more than 95% of those working within medico-legal settings indicated that they always commented on the examinee's approach to testing and level of co-operation, as well as 76% of those working in clinical settings (McCarter et al., 2009). However, validity testing was reported by only 59% to be incorporated into their medico-legal examinations. In addition, only 11% reported using SPVTs most of the time, and the majority indicated that they employed SPVTs rarely.

The rate of experts' SPVT use was found to be higher than non-experts; Schroeder et al. (2016) found that more than 90% of experts used both stand-alone and embedded SPVTs

in their assessments, and the majority reported that they gave more weight to SPVT results that their initial clinical judgement. The majority (79%) of experts in the study by Slick et al. (2004) used at least one SPVT per assessment, and all respondents who employed a fixed battery approach reported they included an SPVT routinely. In pain settings, 71% indicated that they assessed the validity of pain complaints, and this was comparable across pain and forensic specialists (Boccaccini et al., 2006). However, Allcott et al. (2014) found that 40% of their multi-disciplinary expert sample in UK personal injury settings indicated they did not express opinion on the validity of performance as a matter of course, and 11% had never considered performance validity. Unsurprisingly, 44% of respondents reported they did not routinely administer SPVTs.

In addition, Brooks et al. (2016) found that participants reported frequently using SPVTs in their assessments with children and adolescents; 92% reported they used at least one stand-alone or embedded validity test per assessment and an average assessment would include one stand-alone PVT, one-to-two embedded PVTs, and one-to-two embedded SVTs. Interestingly, this far exceeds that reported in several other studies using adult samples. The number of validity tests used per assessment with children and adolescents was not affected by the clinicians' level of training.

Lastly, clinicians conducting forensic evaluations were more likely than those solely working clinically to employ stand-alone SPVTs, both with adults (Slick et al., 2004) and children (Brooks et al., 2016).

In summary, the frequency of SPVT use was variable across evaluation setting, geographical location, and client characteristics. It appears that over time, the rate of SPVT use has increased, however, and validity testing recommendations are being more closely adhered to (Hirst et al., 2017).

In terms of the tests used to validate performance, a broad range of stand-alone and embedded measures were listed by study respondents; more detail can be found in Appendix B. By far the most commonly utilised stand-alone SPVT across the studies was the Test of Memory Malingering (TOMM; Tombaugh, 1996), in all but one study. Sharland and Gfeller (2007) found that as well as being the most frequently utilised test, professionals also had the greatest confidence in the TOMM; classification accuracy was rated at 7.5/10. In addition, the five highest ratings for classification accuracy were given to stand-alone tests; however, only the TOMM was among the 10 most frequently utilised PVTs.

The most popular SVT across studies was the MMPI-II, which was more widely used in the USA and New Zealand (Sharland & Gfeller, 2007; Barker-Collo & Fernando, 2015) than in the UK (McCarter et al., 2009).

In terms of validity measures embedded within existing tests, several of the studies found Reliable Digit Span (a calculation derived from the Digit Span subtest of the Wechsler Adult Intelligence Scale - Revised, Wechsler, 1981; Greiffenstein, Baker, & Gola, 1994) to be the most utilised (Brooks et al., 2016; Young et al., 2016; Martin et al., 2015; Sharland & Gfeller, 2007), along with the California Verbal Learning Test-II and Children's Version (CVLT-II and CVLT-C; Delis, Kaplan, Kramer, & Ober, 1994, 2000), a test of semantic verbal list learning.

Interestingly, none of the pain specialists surveyed by Boccaccini et al. (2006) endorsed any measure specifically intended to assess symptom validity. Qualitative comments suggested that respondents relied upon general pain and coping measures, despite none of the measures cited incorporating validity scales (for example, the McGill Pain Questionnaire; Melzack,1975).

Beliefs regarding validity testing. Martin et al. (2015) found that a vast majority (98%) of US neuropsychologists surveyed considered validity testing to be mandatory within

forensic assessments. Just 55% of the sample believed SPVTs to be mandatory in clinical settings, however, with 38% believing tests to be desirable. This is in contrast to McCarter et al. (2009) who found fewer still considered validity testing to be mandatory in UK clinical contexts (5-7%); 16% felt that validity testing was unnecessary in clinical assessments.

Approximately 70% of neuropsychologists surveyed by Hirst et al. (2017) believed that an SPVT should be included in every assessment. Of those who indicated they followed recommendations, a significantly greater number practised in adult settings compared to paediatric or geriatric settings (89% and 10% respectively), than those not following validity testing guidelines (64% and 35% respectively, $p<.001$).

Slick et al. (2004) explored the confidence of North American experts in their own abilities to detect 'exaggerated or faked deficits'. The average rating provided was 7.75/10, and ratings were weakly correlated with reported base rates of 'definite malingering', but were strongly correlated with estimates of 'possible malingering' ($r=-.13$, $p=.44$ and $r=-.79$, $p<.01$ respectively). This suggests that lower confidence in ability to detect malingering was reported by those who estimated a higher prevalence of malingering.

In terms of the presentations most likely be subject to validity concerns, Allcott et al. (2014) found that respondents provided the highest ratings for pain (headache; 50%) and cognitive complaints (35%). However, pain specialists who had not undertaken forensic training made several qualitative comments appearing to dispute the necessity of validating pain symptoms in the study by Boccaccini et al. (2006), including: "Pain is a subjective experience. Experts in pain are taught to believe the patient's reports. Diagnostic tests are not as useful for pain conditions as other medical problems" (p. 59).

The experts sampled by Schroeder et al. (2016) largely agreed with a general neuropsychologist sample regarding validity testing and were similarly knowledgeable about current recommendations (Martin et al., 2015). The latter study found a significant

relationship between validity testing beliefs and the number of articles read. The authors compared high and low readership groups and found that 62% respondents in the high readership group considered validity testing to be mandatory versus 40% in the low readership group (p<.001). The high readership group was also significantly more likely to strongly agree with the statement that "validity testing is more accurate than clinical impressions in determining patient credibility" than the low readership group (58% versus 33%, p<.001).

Likewise, Allcott et al. (2014) found that 55% of their UK sample were not able to list any peer-reviewed literature on the subject of performance validity, and half of respondents who indicated they routinely used SPVTs could not name any peer-reviewed research.

Justifications for use. The psychologists surveyed by Barker-Collo and Fernando (2015) reported they would decide whether to utilise SPVTs based upon various client characteristics, such as the presence of secondary gain, or unusual symptoms or history. Respondents reported using SPVTs due to endorsement by professional boards, awareness of support in the literature, to safeguard the validity of conclusions drawn, and to improve client care. Reasons provided for not using SPVTs included practical challenges such as time constraints, limited access to tests, and lack of training or experience. Comments also indicated that clinicians' reservations were concerned with the notion that validity testing does not reveal underlying motivations, and disapproval of using deception with clients.

Dandachi-Fitzgerald et al. (2013) found that SPVTs were not utilised in the presence of severe cognitive impairment. Furthermore, 23% of respondents indicated that they believed clinical cases to 'rarely malinger or exaggerate', and 23% felt that symptom invalidity was obvious from the examinees' presentation or from performance in other tests. Reasons provided for using SPVTs were related to awareness of the literature, SPVTs being necessary to validate other findings, and in line with recommendations from professional

bodies. A significantly greater number of respondents reported they utilised SPVTs “to cover my back” (p. 780) in comparison to the UK psychologists sampled by McCarter et al. (2009; 31% and 18% respectively, $p<.01$).

The most frequently reported justification for SPVT use in the UK study by McCarter et al. (2009) was the endorsement by the scientific and professional literature, as well as the need to validate the assessment findings overall. The most commonly endorsed reason for not including SPVTs was related to the belief that invalid responding was obvious from observations or other test results (29%). Respondents also reported that time constraints and a perception of low base rates of ‘malingering’ in clinical cases were justifications for the exclusion of SPVTs in assessments (27% and 26% respectively).

Allcott et al. (2014) also discovered scepticism in relation to validity testing in experts, who commented that “history and examination are best indicators”, “validity of such instruments remains questionable”, and “I am unaware of any reliable tests or procedures that are of help” (p. 72).

Providing warning. Where reported, all studies indicated that respondents encouraged the majority of examinees to do their best when beginning an assessment. Findings were more mixed on providing explicit warning that examinees would be completing tests sensitive to invalid performance, particularly in forensic assessments (Dandachi-Fitzgerald et al., 2013). Furthermore, Hirst el al. (2017) found that US neuropsychologists were significantly less likely than non-US respondents to provide explicit warning (22% and 32% respectively, $p=.003$). Clinicians working with children and adolescents were even less likely to disclose use of SPVTs (8% explicitly stated use; Brooks et al., 2016).

Interpretation. Professionals also had differing views on how to interpret SPVT failure. Schroeder et al. (2016) found that experts considered malingering to be the most

likely cause of SPVT failure in forensic settings, but a very infrequent cause of failure in clinical settings. Experts considered other factors, such as somatoform or conversion disorder, psychiatric issues or attitude towards testing (oppositional, non-compliant or indifferent behaviour), to be underlying SPVT failure in clinical contexts, although there was no consensus as to common underlying mechanisms.

The respondents surveyed by Martin et al. (2015) reported that the most likely cause of test invalidity in clinical cases was psychiatric issues (not including somatoform or conversion disorder). However, the most likely cause of SPVT failure in forensic settings was reported to be malingering. Malingering was listed to be the sixth most common reason for invalid responding in clinical settings. The least common underlying causes in both clinical and forensic settings were genuine cognitive impairment, and diagnosis threat.

A vast majority of the studies found that respondents preferred to report that test results were ‘inconsistent with severity of injury’, and that ‘no firm conclusions can be drawn’ (Dandachi-Fitzgerald et al., 2013; Sharland & Gfeller, 2007; Mittenberg, et al., 2002; Martin et al., 2015). Experts were more likely than general neuropsychologists to report that test data was invalid when SPVTs had been failed, as well as those practising in the USA compared to non-US clinicians (Slick et al., 2004; Sharland & Gfeller, 2007). More pejorative terms such as ‘malingering’ were not favoured by respondents across the studies. Martin et al. (2015) found that only 11% would use the term malingering, which is half that found eight years earlier (Sharland & Gfeller, 2007). However, experts were more likely than general clinicians to use this term (Schroeder et al., 2016). Findings also showed that most respondents preferred the professional term ‘PVT’ (74%), and just 14% preferred using ‘effort measure’. Interestingly, 23% of those surveyed by Young et al. (2016) reported using the terms somatoform (excessive somatic complaints) and cogniform (excessive cognitive complaints) labels when reporting on invalid test results, despite there currently being no

empirical support for SPVT failure being explained by these psychiatric conditions to the author's knowledge.

Of a multi-disciplinary sample of experts, 46% indicated they felt it to be desirable to know the amount of compensation being claimed when forming an opinion on response validity (Allcott et al., 2014).

Feedback and management. In terms of managing suspected invalid performance, a diverse range of responses was reported. The majority of an expert sample indicated they rarely or never confronted the examinee (Schroeder et al., 2016). This is significantly less likely than was found in a prior expert survey (0% versus 25%, $p<.01$; Slick et al., 2004) as well as in general neuropsychologists (4% versus 23%, $p<.01$; Martin et al., 2015).

Participants across the studies were split on ways of responding when suspecting invalid performance; the majority indicated they would administer additional SPVTs, some would continue as normal, and a smaller minority would discontinue (Martin et al., 2015; Dandachi-Fitzgerald et al., 2013; Slick et al., 2004; Hirst et al., 2017; Brooks et al., 2016). Interestingly, in forensic cases, neuropsychologists were more likely to continue as planned (75%) than to terminate the assessment (20%; Martin et al., 2015).

Concerningly, Hirst et al. (2017) reported that approximately one third of respondents indicated they mostly or always continued to interpret the assessment as usual even when SPVTs had been failed.

Critical Review

Papers were critiqued using the Center for Evidence Based Management Quality Appraisal Tool for surveys (CEBMa; 2014); Appendix C. This tool was chosen as it specifically critiqued surveys and therefore allowed for a more thorough comparison of studies that were highly homogenous in design. For example, the CEBMa Tool includes items

relating to survey selection bias and response rate, factors that may have been overlooked using a more general appraisal tool.

Table 4 illustrates the ratings for each study. The main methodological issues will be discussed in turn, followed by synthesised findings in light of the limitations discussed, with implications for further research and clinical practice.

All studies scored relatively comparably on the CEBMa appraisal tool, achieving between 36-55% of checklist items. Studies differed on their scores for response rate, statistical analysis and potential for confounding variables.

Table 4.
Results of Quality Appraisal (Center for Evidence Based Management, 2014)

	Hirst et al. (2017)	Brooks et al. (2016)	Schroeder et al. (2016)	Young et al. (2016)	Barker-Collo & Fernando (2015)	Martin et al. (2015)	Allcott et al. (2014)	Dandachi-Fitzgerald et al. (2013)	McCarter et al. (2009)	Sharland & Gfeller (2007)	Boccaccini et al. (2006)	Sullivan et al. (2006)	Slick et al. (2004)	Mittenberg et al. (2002)
1. Clearly focused question?	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
2. Study design appropriate?	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
3. Clearly described method of subject selection?	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
4. Possible selection bias?[a]	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
5. Sample representative?	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
6. Sample size based on power calculations?	N	N	N	N	N	N	N	N	N	N	N	N	N	N
7. Satisfactory response rate?	N	U	Y	Y	U	U	Y	N	Y	Y	N	U	Y	Y
8. Valid and reliable questionnaire?	N	N	N	N	N	N	N	N	N	N	N	N	N	N
9. Statistical significance assessed?	Y	Y	Y	Y	N	Y	N	Y	N	N	N	N	N	N
10. Confidence intervals given?	N	N	N	N	N	N	N	N	N	N	N	Y	N	N
11. Evidence of potential confounding variables?[a]	N	N	Y	Y	Y	Y	Y	N	Y	Y	Y	N	N	N
%	55	55	55	55	36	45	45	55	45	45	36	55	55	55

Note: Y = yes, N = no, U = unclear; Shaded areas identify failed items; Criteria 12 excluded as not relevant to current review; [a] Reverse scoring

Sample Representativeness

The majority of studies included samples that well represented their target population based on their focused question or issue. Half of the studies reviewed aimed to find out the general practices and beliefs of neuropsychologists and therefore sampled broadly without *a priori* hypotheses. Sharland and Gfeller (2007) surveyed a random sample of approximately one-third of members of a professional body, which may have decreased non-response bias and potentially increased the representativeness of the sample.

Four studies intended to explore the practices of experts and specialists within litigation settings and sampled using a variety of methods (Schroeder et al., 2016; Allcott et al., 2014; Boccaccini et al., 2006; Slick et al., 2004). Schroeder et al. (2016) and Slick et al. (2004) sampled more broadly and applied inclusion and exclusion criteria to define expert status. However, the former applied more stringent criteria making the sample more likely to represent experts in the field. Allcott et al. (2014) contacted all experts registered on the Directory of Expert Witnesses which is likely to have produced a sample representative of the target population. Although it was not possible to verify credentials of the VA healthcare clinicians surveyed by Young et al. (2016), it appears that due to the clinical activities reported by the sample, the population was relatively well represented.

Unfortunately, the generalisability of some studies was limited by poor response rates. Despite initially aiming to investigate international adherence to guidelines, only a small minority of international responses were received by Hirst et al. (2017), and the survey was only disseminated in English. Similarly, Brooks et al. (2016) received insufficient responses from practitioners working with the youngest children (below five years), limiting the generalisability of their findings to paediatric neuropsychologists. Furthermore, only six

countries agreed to participate in the study by Dandachi-Fitzgerald et al. (2013) which was intended to represent European practitioners.

Although all samples were scored as acceptable in terms of representativeness on the CEBMa checklist, some studies increased accessibility and generalisability of findings through appropriate and thorough dissemination to their target populations.

Selection Bias

All studies scored poorly on the CEBMa checklist due to using self-selecting or convenience sampling methods to greater or lesser extents, which inevitably introduces some level of bias. Nevertheless, performance validity is a trending topic in neuropsychology currently so can be considered relevant to most in the profession (Bigler, 2014).

A high level of selection bias is likely in the study by Barker-Collo and Fernando (2015) due to the sampling method of advertising the study in professional society newsletter; it is likely that those who responded were highly motivated to take part and may have had particular views on the topic.

The broad sampling method used by Dandachi-Fitzgerald et al. (2013) increased the risk of non-response bias, as not all of the participants approached would have conducted neuropsychological assessments. There was no way of discerning to what extent non-response bias affected the findings; however, the authors attempted to counter this by comparing findings to similar surveys in the USA and UK.

Sampling bias may also have been problematic in the survey by Brooks et al. (2016) due to the method of using an open online survey. However, in order to counter this the authors invited views of both practitioners regularly using SPVTs and those who did not.

Sample Size

Sample sizes varied throughout the studies reviewed. The mean sample size across all samples was 194 participants, with a range of 17 - 654. Power analysis was not conducted by any of the papers, meaning all studies scored poorly on this item on the CEBMa checklist.

Even without power calculations, it was clear that some studies suffered from small samples that ultimately impacted the robustness of the conclusions drawn. For example, the samples of five of the studies were under 100 (including *N*=17 in Sullivan et al., 2006, and *N*=24 in Schroeder et al., 2016). The majority of studies included sample sizes between 100 and 300, although it is noted that the sample of McCarter et al. (2009) was comparable to US studies with a much larger professional base of practitioners than the UK. Three studies benefitted from larger sample sizes between 300 and over 600 participants, which likely increased the generalisability of findings (Martin et al., 2015; Dandachi-Fitzgerald et al., 2013; Hirst et al., 2017).

Response Rate

Response rates were also variable throughout the studies, and it was not possible to report on response rates in some due to the sampling method utilised. Where reported, the mean response rate was 25%, with a range of 6 – 61%.

Of the studies achieving only a small response rate, Dandachi-Fitzgerald et al. (2013) received the lowest with 6% from their survey distributed in Denmark. Additionally, Hirst et al. (2017) gained a 9% response rate. Studies with relatively larger response rates (between 40-60%), included Young et al. (2016), Schroeder et al. (2016) and Slick et al. (2004), earning these studies higher scores on the CEBMa checklist. However, the majority of studies either did not report rates or had low response rates, and may reflect the opinions of only those motivated by the topic.

Measures

All studies utilised self-report measures, which are at risk of bias due to social desirability. Furthermore, it is not clear from the majority of studies whether surveys were anonymous.

None of the studies used standardised questionnaires with established validity and reliability due to the nature of the research being conducted. Eight papers based their questionnaires on other surveys that have yielded useful results which probably helped to increase the validity of their measures. However, the lack of standardised measures increased the risk of systematic bias. Brooks et al. (2016) suggested that unclear survey wording on one question had produced anomalous results. Similarly, the survey used by Barker-Collo and Fernando (2015) enabled respondents to select more than one response which made the findings unclear. For instance, it was not possible to discern whether respondents only used clinical judgement in assessment of effort or whether this was in conjunction with SPVTs.

In addition, none of the studies investigated actual prevalence of SPVT use from reports or databases by employing a retrospective cohort design, and instead relied upon estimates.

The majority of the surveys used only closed questions which limited the richness of findings compared to more open questions. For example, Young et al. (2016) neglected to explore practitioners' reasons for using specific terminology over others, or other perceived reasons for SPVT failure, which would have been a valuable addition to the literature.

Analyses

The majority of studies employed only descriptive statistics to analyse their data. However, a substantial minority of more recent studies utilised inferential statistics to compare findings to that of previous studies (Hirst et al., 2017; Brooks et al., 2016; Schroeder et al., 2016; Young et al., 2016; Martin et al., 2015; Dandachi-Fitzgerald et al., 2013). Only

one study included confidence intervals (Mittenberg et al., 2002), increasing the likelihood of obtaining a 'true' value.

Confounding Variables

There appear to be some common potentially confounding variables throughout the studies. The first is related to the terminology used to describe performance and symptom validity. There was significant variability in terms used, and therefore in interpretations of meaning (unsurprising given this issue continues to be debated in the wider profession; Bigler, 2012). Whilst some studies explored the preferred terms used by practitioners, none investigated the respondents' definitions of each term in their responses. Young et al. (2016) used 'SVT' to describe both symptom and performance validity tests. Furthermore, comparison of findings may have been impacted by the lack of consistency in the questions used across surveys, as terminology was not always identical.

Some studies pooled data which made conclusions less clear. Young et al. (2016) combined stand-alone and embedded tests and Sharland and Gfeller (2007) pooled the base rates of invalid performance across practice settings, which may have led to an overall over-estimation. Similarly, McCarter et al. (2009) failed to explore the proportion of litigation cases and forensic cases conducted by their sample, instead combining these practice settings. This is likely to have impacted findings due to the greater base rate of performance invalidity generally found in forensic settings.

In addition, there were inconsistencies in reporting of the average number of assessments completed in the last year by respondents (reported in only half of the studies). Therefore, participants may have been responding based on limited or no contemporary experience conducting assessments.

There were also inconsistencies in reporting of the average number of years of practice by respondents; although the majority of studies did report this. Barker-Collo and

Fernando (2015) noted that their sample of registered psychologists in general practice was likely to represent a highly diverse range of skills and experience. Varying levels of experience of practitioners was also reported by Martin et al. (2015), with half of their sample practicing in the field of neuropsychology for 10 years or less. Convenience sampling also meant that the credentials of the sample could not be verified in most cases. These factors may have decreased the likelihood of reaching the target population.

Synthesis and Conclusions

Taken together, findings were considered most convincing when there were adequate sample sizes and response rates, attempts to overcome selection bias and sophistication of analysis (as found in Hirst et al., 2017, and Dandachi-Fitzgerald et al., 2013).

Generally, the literature lacked diversity in methods of investigating the topic, and this produced data limited by the quantitative designs used. There is currently no qualitative research into the beliefs of professionals regarding performance invalidity to the author's knowledge. However, the samples employed in the studies under current review were generally representative of the target population, and findings were felt to be relevant and useful to practitioners in the field, holding limitations in mind.

Concerning base rates of response invalidity, a substantial minority of forensic and medico-legal cases and a smaller minority of clinical assessments were considered by professionals to produce invalid performances across a range of geographical locations. Estimated base rates in clinical evaluations were between 3 and 10%, in forensic cases estimates were 17-25%, and were variable in litigation settings (between 8-30%).

Concerningly, clinical judgement was relied upon by a vast proportion of respondents in the studies, despite established research indicating its limited reliability in detecting invalid performance (Faust et al., 1988).

Frequency of SPVT use was variable across evaluation setting, geographical location, and client characteristics. However, it appeared that over time the rate of SPVT use had increased, and validity testing recommendations had been more closely adhered to (Hirst et al. (2017).

In terms of the most frequently used stand-alone PVT, the TOMM was the most prevalent across service settings. Embedded measures were also frequently used according to the studies in this review, particularly within the CVLT and Digit Span tests.

The majority of professionals responding to the studies felt that SPVTs were mandatory in forensic settings, but not in clinical contexts. Justifications for excluding SPVTs were mostly related to belief that clinical cases rarely exaggerate and that symptom invalidity is obvious from other indicators, as well as time constraints. However, base rates reported by the studies reviewed also challenge the belief that SPVT failure is rare in clinical contexts. Reasons provided for using SPVTs were related to awareness of the literature and SPVTs being necessary to validate other findings.

The majority of the samples encouraged examinees to do their best, but did not provide explicit warning, a practice more commonly found in forensic settings. Views were mixed on how to manage invalid performance, but most would administer additional SPVTs. When reporting on performance or symptom invalidity, the majority stated that the test results were invalid, inconsistent with the severity of the injury, and that no firm conclusions could be drawn. Very few participants across the studies used pejorative terms such as ‘malingering’.

Overall, there was variability in practitioners’ adherence to professional recommendations regarding performance and symptom invalidity, which is unsurprising given the relative recency of these guidelines and the supporting evidence base; although it appeared that the USA had progressed furthest in the field. This review would support the call

for further research into response validity outside of the USA (McMillan et al., 2009). Nevertheless, the surveys examined suggested a general trend towards consideration of this substantial and complex source of test data variance, both in research and clinical practice.

Limitations and Future Directions

In order to make efforts to access all relevant papers for inclusion in this review, various terminology and definitions of performance invalidity have been included and findings grouped, which may pose a threat to the validity of conclusions. However, this is reflective of the interchangeable use of terms in the literature and emphasises the need for clearer definitions of this concept in future research. It is also acknowledged that the use of a survey-specific quality appraisal tool may have acted to focus the critique on survey design rather than other issues such as the quality and interpretation of the results.

In terms of clinical implications, the presence of out-dated and inaccurate beliefs and practices regarding validity testing by professionals suggests a need for more training as well as clearer and more consistent guidance from international professional bodies. It is also clinically implicated for professionals to keep more abreast of the literature to inform their practice. The current review suggests there is a need for clinicians to understand the contributing factors and mechanisms underlying SPVT failure and to consider this as part of a comprehensive biopsychosocial formulation. Broader research and clinical implications of the review are outlined in Appendix D.

Despite considerable research, there remain unanswered questions within the field, particularly the mechanisms underlying invalid performance. Moreover, further research is required into the base rates of SPVT failure across treatment settings and geographical locations, as the current literature is predominantly from North America using mostly litigating and forensic populations.

The majority of the studies reviewed suffered from low response rates. It may be beneficial for future studies to offer incentives to respondents, or to utilise an alternative study design such as focus groups or exit polls. Qualitative research would also provide richer information on the more complex and subtle factors associated with performance validity and clinicians' beliefs in relation to these.

Future research would benefit from being clearer in the definitions of performance invalidity utilised. In particular, an exploration into UK PVT and SVT failure rates would be fruitful, as well as the factors that influence these.

References

*Allcott, D., Anderson, S., Friedland, D., Leng, N., Gross, M., Skelton-Robinson, M., & Weller, M. (2014). How do experts reporting for the legal process validate symptoms? The results of a survey. *Medicine, Science and the Law*, *54*(2), 68-73. https://doi.org/10.1177/0025802413491247

*Barker-Collo, S. L., & Fernando, K. (2015). A survey of New Zealand psychologists' practices with respect to the assessment of performance validity. *New Zealand Journal of Psychology*, *44*(2), 35-42. Retrieved from http://www.psychology.org.nz/publications-media/new-zealand-journal-of-psychology/?#.WsXrS4jwbIU

Bennett-Levy, J., Klein-Boonschate, M. A., Batchelor, J., McCarter, R., & Walton, N. (1994). Encounters with Anna Thompson: The consumer's experience of neuropsychological assessment. *The Clinical Neuropsychologist*, *8*(2), 219-238. https://doi.org/10.1080/13854049408401559

Bigler, E. D. (2012). Symptom validity testing, effort, and neuropsychological assessment. *Journal of the International Neuropsychological Society*, *18*(4), 632-640. https://doi.org/10.1017/S1355617712000252

Bigler, E. D. (2014). Effort, symptom validity testing, performance validity testing and traumatic brain injury. *Brain Injury, 28*, 1623-1638. https://doi.org/10.3109/02699052.2014.947627

*Boccaccini, M. T., Boothby, J. L., & Overduin, L. Y. (2006). Evaluating the validity of pain complaints in personal injury cases: Assessment approaches of forensic and pain specialists. *Journal of Forensic Psychology Practice*, *6*(3), 51-62. https://doi.org/10.1300/J158v06n03_03

Bush, S. S., Ruff, R. M., Troster, A., Barth, J., Koffler, S. P., Pliskin, N. H., & Silver, C. H. (2005). NAN position paper: Symptom validity assessment: Practice issues and medical necessity. *Archives of Clinical Neuropsychology*, *20*(4), 419-426. https://doi.org/10.1016/j.acn.2005.02.002

Butcher, J. N., Dahlstrom, W. G., Graham, J. R, Tellegen, A. M., & Kreammer, B. (1989). *The Minnesota Multiphasic Personality Inventory-2 (MMPI-2) Manual for Administration and Scoring*. Minneapolis, MN: University of Minneapolis Press.

*Brooks, B. L., Ploetz, D. M., & Kirkwood, M. W. (2016). A survey of neuropsychologists' use of validity tests with children and adolescents. *Child Neuropsychology*, *22*(8), 1001-1020. https://doi.org/10.1080/09297049.2015.1075491

Center for Evidence Based Management (2014). *Critical appraisal of a cross-sectional study (survey).* Retrieved from https://www.cebma.org/wp-content/uploads/Critical-Appraisal-Questions-for-a-Cross-Sectional-Study-july-2014.pdf

Chafetz, M. D., Williams, M. A., Ben-Porath, Y. S., Bianchini, K. J., Boone, K. B., Kirkwood, M. W., ... & Ord, J. S. (2015). Official position of the American Academy of Clinical Neuropsychology Social Security Administration policy on validity testing: Guidance and recommendations for change. *The Clinical Neuropsychologist*, *29*(6), 723-740. https://doi.org/10.1080/13854046.2015.1099738

*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. https://doi.org/10.1093/arclin/act073

Delis, D. C., Kaplan, E., Kramer, J., & Ober, B. (1994). *California Verbal Learning Test—Children's Version.* San Antonio, TX: The Psychological Corporation.

Delis, D. C., Kaplan, E., Kramer, J., & Ober, B. (2000). *California Verbal Learning Test-II.* San Antonio, TX: The Psychological Corporation.

Faust, D., Hart, K. J., Guilmette, T.J., & Arkes, H. R. (1988). Neuropsychologists' capacity to detect adolescent malingerers. *Professional Psychology: Research and Practice, 19*, 508-51. Retrieved from http://www.apa.org/pubs/journals/pro/

Ferlie, E. B. & Shortell, S. M. (2001). Improving the quality of health care in the United Kingdom and the United States: a framework for change. *The Milbank Quarterly, 79,* 281-315. https://doi.org/10.1111/1468-0009.00206

Francke, A. L., Smit, M. C., de Veer, A. J., & Mistiaen, P. (2008). Factors influencing the implementation of clinical guidelines for health care professionals: A systematic meta-review. *BMC Medical Informatics and Decision Making*, *8*(1), 38-49. https://doi.org/10.1186/1472-6947-8-38

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. https://doi.org/10.1080/02699050110088254

Green, P. & Merten, T. (2013). Noncredible explanations of noncredible performance on symptom validity tests. In Carone, D. A. & Bush, S. S. (Eds.), *Mild traumatic brain injury: Symptom validity assessment and malingering* (pp.73-96). New York, NY: Springer.

Greiffenstein, M. F., Baker, W. J., & Gola, T. (1994). Validation of malingered amnesia measures with a large clinical sample. *Psychological Assessment, 6*(3), 218-224. http://dx.doi.org/10.1037/1040-3590.6.3.218

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. https://doi.org/10.1080/13854040903155063

*Hirst, R. B., Han, C. S., Teague, A. M., Rosen, A. S., Gretler, J., & Quittner, Z. (2017). Adherence to validity testing recommendations in neuropsychological assessment: A survey of INS and NAN members. *Archives of Clinical Neuropsychology*, *32*(4), 456-471. https://doi.org/10.1093/arclin/acx009

Larrabee, G. J. (2012). Performance validity and symptom validity in neuropsychological assessment. *Journal of the International Neuropsychological Society, 18*(04), 625-630. https://doi.org/10.1017/S1355617712000240

*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. https://doi.org/10.1080/13854046.2015.1087597

*McCarter, R. J., Walton, N. H., Brooks, D. N., & Powell, G. E. (2009). Effort testing in contemporary UK neuropsychological practice. *The Clinical Neuropsychologist*, *23*(6), 1050-1066. https://doi.org/10.1080/13854040802665790

McMillan, T. M., Anderson, S., Baker, G., Berger, M., Powell, G. E., & Knight, R. (2009). *Assessment of effort in clinical testing of cognitive functioning for adults*. Leicester, UK: The British Psychological Society.

Melzack, R. (1975). The McGill Pain Questionnaire: Major properties and scoring methods. *Pain*, *1*(3), 277-299. https://doi.org/10.1016/0304-3959(75)90044-5

*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. doi/abs/10.1076/jcen.24.8.1094.8379

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), 10-20. DOI: 10.1037/1040-3590.10.1.10

*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. https://doi.org/10.1080/13854046.2016.1177118

*Sharland, M. J., & Gfeller, J. D. (2007). A survey of neuropsychologists' beliefs and practices with respect to the assessment of effort. *Archives of Clinical Neuropsychology*, *22*(2), 213-223. https://doi.org/10.1016/j.acn.2006.12.004

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. https://doi.org/10.1076/1385-4046(199911)13:04;1-Y;FT545

*Slick, D. J., Tan, J. E., Strauss, E. H., & Hultsch, D. F. (2004). Detecting malingering: A survey of experts' practices. *Archives of Clinical Neuropsychology*, *19*(4), 465-473. https://doi.org/10.1016/j.acn.2003.04.001

*Sullivan, K., Lange, R. T., & Dawes, S. (2006). Methods of detecting malingering and estimated symptom exaggeration base rates in Australia. *Journal of Forensic Neuropsychology*, *4*(4), 49-70. https://doi.org/10.1300/J151v04n04_04

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

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

*Young, J. C., Roper, B. L., & Arentsen, T. J. (2016). Validity testing and neuropsychology practice in the VA healthcare system: Results from recent practitioner survey. *The Clinical Neuropsychologist*, *30*(4), 497-514. https://doi.org/10.1080/13854046.2016.1159730

Section B

Response Validity and Psychological Functioning in a UK NHS Acquired Brain Injury Sample

7960 words (plus 397 additional words)

Abstract

Objectives: Performance validity tests (PVTs) and symptom validity tests (SVTs) have been recommended by the British Psychological Society to assist clinicians in validating assessment data. The current study aimed to explore the base rate of PVT failure in an NHS neuropsychology service, a setting relatively unexplored. A secondary aim was to investigate the relationship between PVT and SVT performance. Lastly, group differences in those passing and failing PVTs were explored in terms of demographics, and psychological functioning.

Method: Archival test data (N = 127) was drawn from an NHS outpatient neuropsychology service. Participants completed one stand-alone PVT (the Test of Memory Malingering [TOMM]), one embedded PVT (Digit Span age-corrected scaled score [DS-SS]), and one SVT (the Personality Assessment Inventory [PAI]).

Results: The base rate of failure on any one PVT was 18%. The rate of TOMM failure was 12% and 4% additionally failed an embedded PVT. A significant relationship was found between PVT and SVT performance. Significantly elevated Paranoia, Anxiety-Related Disorders, and Schizophrenia PAI scales, as well as lower Full Scale IQ scores, were found in those who failed PVTs compared to those who passed. No other group differences on demographics were found, including reported financial incentive.

Conclusions: Findings suggest that PVT failure occurs in a sizable minority of NHS outpatients with acquired brain injuries (ABI), which is unlikely to be simply explained by malingering for financial gain. Elevations in reported psychopathological symptoms may be related to emotional and cognitive sequalae resulting from the ABI itself. Careful interpretation of neuropsychological test data is endorsed.

Key words: Performance validity, symptom validity, Test of Memory Malingering, Personality Assessment Inventory, neuropsychological assessment

Response Validity and Psychological Functioning in a UK NHS Acquired Brain Injury Sample

The practice of neuropsychology rests upon the assumption that brain functioning can be inferred from neuropsychological test performance (Schoenberg & Scott, 2011). The reliability and validity of conclusions drawn is dependent on a number of factors. These include examinee characteristics and influences, such as whether they try their best. The assessment of examinee 'effort' is concerned with capturing non-neurological dimensions of performance (Bigler, 2012) and has gained increased attention in the field since the turn of the century (Carone & Bush, 2013). Clinical judgement has been shown to be an unreliable method of identifying validity issues (Faust, Hart, Guilmette & Arkes, 1988). Research has therefore focused on the empirical testing of this construct.

The terminology used in the literature has varied widely (Larrabee, 2012); see Appendix A. In the current thesis, performance validity testing (PVT) will refer to the assessment of validity of performance on cognitive tasks, and symptom validity testing (SVT) will refer to the assessment of the validity of self-reported symptoms, consistent with Larrabee's (2012) recommendations. PVTs are usually within the format of a forced choice memory paradigm that appears difficult to examinees but actually involves very easy recognition memory tasks. They should be affected very little by brain trauma, age, overall intellectual functioning, and education (Carone & Bush, 2013). Below-chance performance is considered to indicate malingering. Scores below cut-off based upon normative data are suggestive of invalid responding, without implying intent to feign (Slick, Sherman & Iverson, 1999). SVTs are concerned with the degree to which symptomatic complaint on self-report measures is reflective of true symptoms, and are usually in the form of mood or personality inventories such as the Personality Assessment Inventory (PAI; Morey, 1991, 2007). As well as 'stand-alone' tests that have been specifically designed to evaluate performance or

symptom validity, the use of 'embedded' measures within existing tests can be used to identify invalid responding without increasing testing time. Several position papers have been released by professional bodies in both the USA and UK endorsing the routine use of validity tests in both clinical and forensic settings (Bush et al., 2005; McMillan et al., 2009).

Knowledge of the prevalence or 'base rate' of invalid responding in the population of interest is necessary for meaningful interpretation of performance (Crawford, 2003). Historically, invalid performance in neuropsychological evaluation has been assumed to be rare outside of medico-legal contexts (McCarter et al., 2009). However, research is increasingly suggestive of a far greater prevalence of underperformance across clinical as well as forensic settings (Bush et al., 2005). Hampson, Kemp, Coughlan, Moulin and Bhakta (2014) found that 27% of a National Health Service (NHS) sample of acquired brain injury (ABI) patients failed one PVT. Similarly, Bunnage, Eichinger, Pearce, Duckworth, and Newson (2008) found a base rate of PVT failure of 26% in a non-litigating NHS sample. This points to the potential oversight of a substantial source of variance, contributing to inaccurate conclusions regarding neuropsychological functioning. This has been highlighted by Fox (2011) who found that failure of just one PVT eliminated the correlation between neuropsychological test performance and brain injury severity.

Likely due to the lack of base rate data in UK clinical settings, clinicians continue to rely on clinical judgement in their interpretation of performance validity. McCarter, Walton, Brooks, and Powell (2009) found that only 5% of their sample of UK psychologists considered PVTs to be mandatory in clinical settings. Almost one third of the sample believed that invalid responding is obvious from observations or other test results.

Theories of Invalid Performance

The malingering hypothesis. Validity test failure has been extensively shown to be predicted by the provision of financial gain by means of litigation (Binder & Rohling, 1996). A meta-analysis by Iverson (2005) found that the overall effect of malingering on neuropsychological test scores was considerably higher than the effect of brain injury, depression, and benzodiazepine withdrawal. The literature is less clear, however, on explanations of malingering in the absence of financial incentive, where it might be assumed that there would no motivation to underperform. Suhr, Tranel, Wefel, and Barrash (1997) found that validity test failure was not found to be predicted by litigation status in a mixed sample with diagnoses of ABI, depression, or somatization disorders.

The cognitive impairment hypothesis. PVTs are designed to be insensitive to cognitive impairment and intelligence level, except in the presence of intellectual disabilities or significant neurodegenerative illness such as dementia (Demakis, Gervais, & Rohling, 2008; Tombaugh, 1996). However, the literature continues to link significantly lower Full Scale IQ (FSIQ) with PVT failure. Hampson et al. (2014) found a greater base rate of PVT failure to be associated with greater injury severity in various NHS ABI populations suggesting that PVTs may be measuring genuine impairment. However, some authors have attributed this effect to the presence of malingered neuropsychological impairment (Constantinou, Bauer, Ashendorf, Fisher, & McCaffrey, 2005). The risk of type I error, whereby invalid responding is falsely identified, and type II error, whereby invalid results are taken as valid, greatly depends upon the sensitivity and specificity of the PVTs employed. The multivariate failure model (Larrabee, 2003) was proposed to address poor sensitivity rates, whereby failure on two or more PVTs can be understood as indicating invalid responding. This has received support in the literature (Victor et al., 2009).

Psychogenic hypotheses. PVTs are designed to be insensitive to mood disturbance (Carone & Bush, 2005). A meta-analysis by Veiel (1997) reported to find major cognitive impairment in depression; however, none of the studies utilised PVTs. When these findings were re-analysed, there was no difference in performance across depressed or non-depressed groups when performance validity was accounted for (Rohling, Green, Allen & Iverson, 2002). Nevertheless, a relationship has been found between depressive symptomatology and levels of negative self-representation on SVTs (Morey, 2007).

Whilst there lacked consistent support for a linear relationship between affective distress and PVT failure in the literature (Ashendorf, Constantinou & McCaffrey, 2004), there appeared to be an interaction between elevated psychological symptom reporting and PVT underperformance (Dandachi-Fitzgerald et al., 2011). Sumanti, Boone, Savodnik and Gorsuch (2005) found elevations on Personality Assessment Inventory (PAI; Morey, 1991, 2007) scales pertaining to Somatic Complaints, Depression, Anxiety, Anxiety-Related Disorders and Schizophrenia in those failing PVTs compared to those who passed. Likewise, in their mixed clinical and litigating US sample, Whiteside et al. (2010) found significant associations between PVT failure and elevations on the PAI Somatic Complaints scale (specifically the Conversion subscale), with Schizophrenia, Anxiety, and Depression scales trending toward significance. Bigler (2012) has argued that unconscious processes such as a 'cry for help', diagnosis threat, or distorted expectations (for example, the impact of disability status or other labels on identity) may offer useful information in understanding the mechanisms underlying invalid performance. Research into diagnosis threat has demonstrated that cognitive test performance and perceived influence of symptoms on performance are influenced by performance expectations (Suhr & Gunstad, 2005). Bigler (2012) suggests that PVTs are no more immune to these effects than other cognitive tests, and that perception of 'illness' and related psychological state may explain 'near-pass', or above chance-level, PVT

performance. Indeed, the British Psychological Society (BPS) has warned clinicians to take care not to reinforce iatrogenic symptoms that may have developed through exposure to the disabled role or ill-health beliefs during the process of pursuing litigation or seeking treatment (McMillan et al., 2009).

It has also been argued that elevated psychological symptom reporting and PVT failure simply represent consistent exaggeration across assessment modalities (Haggerty, Frazier, Busch, & Naugle, 2007). There lacks consensus on whether PVTs and SVTs measure similar or different constructs. Van Dyke, Millis, Axelrod, and Hanks (2013) found that the domains were not consistently invalidated, and therefore endorsed the separate assessment of performance and symptom validity. However, Whiteside, Dunbar-Mayer, and Waters (2009) found correlations between PVT failure and SVT failure using the Personality Assessment Inventory validity scales (PAI; Morey, 1991, 2007). The authors further demonstrated that SVT performance could significantly predict PVT performance and argued for the presence of a 'defensive' response style (Gaasedelen, Whiteside & Basso, 2017).

Study Aims

The current study aimed to explore the base rate of PVT failure in a sample of NHS outpatients with acquired brain injuries. Larrabee's (2003, 2014) two-or-more-fails criterion will be applied to reduce the risk of type I error, which would add a novel element to the UK literature on performance validity. Furthermore, the BPS has highlighted the need for better understanding of the meaningfulness of PVT failure (McMillan et al., 2009). Without this, service-users could be subject to false positive diagnoses of suboptimal effort and associated invalid recommendations, and even incorrect social entitlements or legal verdicts (Mossman, Wygant & Gervais, 2012).

Secondly, this study aimed to investigate whether PVTs tend to be failed when SVT indicators are elevated, or whether they measure different domains of response. Exploration

into the relationships between cognitive and psychological functioning has been identified as requiring continued research to inform clinical practice (Whiteside et al., 2010). A third aim was to explore differences between individuals who pass and fail PVTs in terms of psychological functioning and personality traits as measured by the PAI. This poses a significant addition to the literature, since very few studies have investigated response validity using clinical samples, and to the researcher's knowledge none have explored the relationship with the PAI in the UK. Furthermore, group differences will be explored in relation to demographic variables, including potential financial incentive.

Hypotheses

It was hypothesised that base rates of failure on a single PVT would be approximately 10-15% based on previous research using mixed clinical samples (Whiteside et al., 2010). The base rate of multiple PVT failure was anticipated to be around 5%.

It was also hypothesised that a greater level of PVT failure would be found in participants with elevated PAI validity scales (Negative Impression Management and Infrequency scales), based on the North American literature (Whiteside et al., 2009; Gaasedelen et al., 2017).

Due to previous findings endorsing a relationship between PVT failure and elevations on measures of emotional and personality functioning (Sumanti et al., 2005; Whiteside et al., 2010), it was hypothesised that there would be significant positive relationships between PVT failure and the PAI scales, specifically Somatic Complaints, Depression, Anxiety, Anxiety-Related Disorders, and Schizophrenia. A second analysis would be performed on the subscales of any PAI clinical scales found to be significantly related to PVT performance.

Significant group differences were not anticipated for the demographic variables (age, diagnosis category, employment status, gender, and pre-morbid IQ [PMIQ]). However, it was anticipated that PVT failure would be associated with lower Full Scale IQ (FSIQ) and also

greater identification of potential financial incentive, based on previous findings (Hampson et al., 2014; Bianchini, Curtis, & Greve, 2006).

As well as PVT pass and fail groups, TOMM pass or failure will be separately analysed in order to provide comparison to previous research findings utilising this measure (Whiteside et al., 2009, 2010).

Method

Participants

Participants were drawn from consecutive referrals presenting to an NHS neuropsychology service in an urban setting in the South of England between February 2009 and March 2014. The service supported people with acquired neurological conditions referred from a number of regional sources. Participants attended an outpatient programme of assessment, treatment, or both. Referral criteria required that all service-users had capacity to consent to the assessment, which was assessed by the treating clinician.

Inclusion criteria were kept purposefully broad in line with the naturalistic design of the research, which aimed to recruit a sample representative of adults accessing NHS neuropsychology services in the UK. All participants were adults (aged 18 and over). The upper bound was set at 89 years since this is the lowest upper age limit of the measures included in the analysis.

Exclusion criteria for the current study were *a prior* diagnosis of intellectual disability, and progressive neurological disorder, such as multiple sclerosis or dementia. This was due to literature suggesting these populations are more likely to score below cut-off on PVTs (Boone & Lu, 1999) (excluded n=21). Participants were also excluded if more than 50% of their test data was missing (n=14).

Table 1

Participant Characteristics

		Descriptives	
Demographic	*n*	*M*	*SD*
Age	127	43.32	14.37
Time since injury (months)	124	56.30	94.47
PMIQ	119	104.52[a]	11.37
FSIQ	117	99.14[a]	17.26
	n	*%*	
Gender	127	100	
Male	84	66	
Female	43	34	
Diagnosis	127	100	
CVA/Stroke/AVM	37	29	
Tumour/cancer related	22	17	
mTBI	17	13	
modTBI	30	24	
sevTBI	5	4	
TBI severity unknown	2	2	
Hypoxia	5	4	
Encephalitis	4	3	
Infection/viral	2	2	
Epilepsy related	2	2	
Cyst	1	1	
Identified financial incentive	122	96	
Yes	27	21	
No	95	75	
Employment status	127	100	
Employed	37	29	
Unemployed	90	71	

Note. PMIQ = pre-morbid IQ; FSIQ = Full Scale IQ; CVA = cerebral vascular accident; AVM = arteriovenous malformation; mTBI = mild traumatic brain injury; modTBI = moderate traumatic brain injury; sevTBI = severe traumatic brain injury

[a] PMIQ and FSIQ values represent mean rank scores and not IQ scores

The final sample (*N*=127) consisted of male (*n*=43) and female (*n*=84) participants ranging in age at assessment from 18 to 74 years (*M*=43.32, *SD*=14.37). Participants presented with a range of acquired brain injuries (see Table 1).

The time since injury ranged from 1 to 545 months (*M*=56.30, *SD*=94.46). The majority of the sample indicated they were not in employment at the time of assessment (*n*=90). Furthermore, the presence of potential financial incentive was identified in 21% of the sample at the time of assessment (*n*=27). Potential financial incentive was routinely explored in the service during the clinical interview and included factors such as pursuing a compensation claim related to their ABI, or pursuing benefits such as disability living allowance, Personal Independence Payment, Employment and Support Allowance, Criminal Injuries Compensation Authority, or early retirement. Participants were not considered to be incentivised if their compensation claim had settled previously. It is acknowledged that this data may reflect self-report bias, and there could be many other types of incentive in this sample. For example, psychosocial incentives could include care elicited from others, or access to services.

Since participants' years of education was not available, a measure of PMIQ was utilised to indicate participants' long-standing intellectual functioning. An updated PMIQ functioning measure became available to the department in 2011. Therefore, 43% (*n*=55) of participants completed the Wechsler Test of Adult Reading (Wechsler, 2001) and 56% (*n*=72) completed the Test of Premorbid Functioning (Wechsler, 2011). Both of these measures involve an oral reading task suggested to remain relatively unaffected by brain injury (Brooks, Holdnack, & Iverson, 2011). Both have also been extensively validated for use with ABI populations (Green et al., 2008; Franzen, Burgess, & Smith-Seemiller, 1997). Analysis was completed to explore whether the PMIQ test used had any impact on performance validity; Chi-square tests for independence (with Yates Continuity Correction) indicated no

significant relationships. PMIQ scores were therefore combined across the sample and ranged from 62 to 129 (*M*=104.52, *SD*=11.37).

It was unfortunately not possible to gather data on participants' ethnicities. However, according to the equality information pertaining to outpatient activity published by the Trust (2014), key ethnic groups included White British (32%), White Other (12%), Black (10%), Asian (10%), Other (4%), Mixed (1%), and 'no data' (28%).

Power analyses were informed by previous research by Whiteside et al. (2009) and Whiteside et al. (2010) comparing PAI and TOMM performance in a US sample. Effect sizes ranged from *rs* = -.15 to .32 (small to medium effect; Cohen, 1992). Using the "G*Power 3" programme (Faul, Erdfelder, Lang, & Buchner, 2007), an allocation ratio of 0.18 was set to account for the estimated base rate of PVT failure (specifying alpha at 5% and desired power at 80%). The required total sample size to detect significant group differences on at least one PAI scale was estimated at 68.

Measures

The Personality Assessment Inventory (Morey, 1991, 2007). The PAI is a self-reported inventory designed to assess various domains of adult personality and psychopathology, comprised of 344 items which load onto 22 non-overlapping scales. These include four validity indices (Positive Impression Management, Negative Impression Management, Inconsistency, and Infrequency), and 11 clinical scales (Somatic Complaints, Anxiety, Anxiety Related Disorders, Depression, Mania, Paranoia, Schizophrenia, Borderline Features, Antisocial Features, Alcohol Problems, and Drug Problems), each with three to four subscales. Additionally, there are five treatment consideration scales (Aggression, Suicidal Ideation, Non-support, Stress and Treatment Rejection), as well as two interpersonal scales (Dominance and Warmth). Respondents are required to indicate the extent to which an item

applies to them using a four-point scale ranging from ‘false’ to ‘very true’. Further details can be found in Appendix F.

The PAI has been found to possess sound psychometric properties. Good test retest reliability has been demonstrated (Boyle & Lennon, 1994; Rogers, Flores, Ustad, & Sewell, 1995), as well as adequate internal consistency and reliability (Morey, 1991). The PAI has been validated for use with ABI populations (Demakis et al., 2007).

Cognitive performance validity tests. The current study will utilise two PVTs; one stand-alone measure (the TOMM; Tombaugh, 1996) and one embedded measure (Digit Span age-corrected scaled score [DS-SS], from the Wechsler Adult Intelligence Scale-III/IV; Wechsler, 1997, 2010). These two PVTs are endorsed by the BPS (McMillan et al., 2009) and are among the most commonly utilised in UK practice (McCarter et al., 2009). Furthermore, these measures operate across a variety of cognitive modalities (visual and auditory memory), in line with BPS recommendations (McMillan, 2009).

The Test of Memory Malingering (Tombaugh, 1996). The TOMM is a 50-item visual memory test designed to discriminate between genuine memory impairment and ‘malingered’ memory deficits. Individuals complete two learning trials and a supplementary retention trial. Tombaugh (1996) suggests a cut-off of 45 out of a possible 50 on Trial 2 to indicate suboptimal performance. The TOMM has demonstrated good specificity and sensitivity (Tombaugh, 1996, 1997; Haber & Fichtenberg, 2006), as well as good internal consistency, reliability, and convergent validity (Moore & Donders, 2004). Furthermore, the TOMM has been found to be relatively insensitive to affective distress (Boone, 2007).

Digit Span age-corrected scaled score (DS-SS). The DS-SS is an embedded PVT within the Working Memory Index of the Wechsler Adult Intelligence Scale-III/IV (Wechsler, 1997, 2010), whereby individuals are required to repeat increasing strings of numbers in the same order, reverse order, and in sequence. Axelrod, Fichtenberg, Millis, and

Wertheimer (2006) found improved specificity and sensitivity when using a scaled score Digit Span cut-off of five or less in their sample referred for neuropsychological assessment, in comparison to utilising the historically more popular Reliable Digit Span (Greiffenstein, Baker, & Gola, 1994). They noted that this cut-off minimizes false positive errors and achieves a "73% probability in support of a diagnosis of response bias" (p. 521).

An updated version of this measure was utilised by the service during the period sampled (the Digit Span subtest from the WAIS-III and WAIS-IV; Wechsler, 1997, 2011). It was decided to pool data using both versions based upon previous research suggesting that the Digit Span subtest in both versions is highly correlated (Robbins, 2014). Furthermore, analyses were conducted to explore group differences relating to the test version used. No significant associations between PVT performance and the version used were found.

Design and Procedure

With permission of the host Trust, two research assistants were briefed on the project and collated raw archival neuropsychological test data from patient archives. Archival files were available from 2009 until the clinic was discontinued in 2014. This data was anonymised at the point of entry onto a password-protected database through the use of participant numbers and stored securely on an encrypted USB. Data was cleaned and quantitatively analysed by the researcher, and kept in a secure location. A between-subjects design was used to investigate group differences; no variables were manipulated. Data will be retained securely for ten years in line with University regulations.

The service employed a comprehensive neuropsychological assessment using a fixed-battery approach, administered in a fixed order for all participants over two days across two consecutive weeks. Although the dataset was relatively complete due to the use of a fixed battery approach, the number of participants included in analyses addressing each research question varied somewhat due to missing data points for some cases. Pairwise deletion was

employed in correlational analyses.

Ethical Considerations

Ethical approval was granted by the NHS REC Proportionate Review Service for the use of anonymous archival data, which was made available by the Trust and nevertheless stored securely in line with university regulations. It was not possible for any individual participant to be identified according to their test data. No risks for participants were identified. The archival database was also partly accessed as part of a separate thesis project; ethics applications, analyses and write-up were completed independently (Appendix I provides further information). Presentation of findings to the neuropsychology department within the NHS Trust involved has been planned following completion of the project.

Results

Analysis was run to assess distribution of data using IBM's Statistical Package for the Social Sciences (SPSS), version 24. Since none of the performance validity variables were determined to be normally distributed and unequal group sizes were expected, non-parametric equivalents were utilised throughout.

The skewness and kurtosis of performance validity variables were examined in order to identify outliers in the data which were then verified to identify any error in data entry.

Base Rates of PVT failure

Analysis was conducted with Trial 2 of the TOMM using a cut-off of 45 based on the manual recommendations (Tombaugh, 1996). An age-corrected scaled score of five or below on Digit Span was used as a cut-off based on recommendations in the literature for achieving optimal sensitivity and specificity (Axelrod et al., 2006).

Table 2

PVT Failure Base Rates

	n	Number of fails	Base rate (%)
Failure of TOMM	127	15	12
Failure of DS-SS	91	12	13
Failure of ≥1 PVT*	127	23	18
Failure of 2 PVTs	91	4	4

Note. PVT = Performance validity test; TOMM = Test of Memory Malingering; DS-SS = Digit Span age-corrected scaled sore

* Failure of TOMM and DS-SS includes participants failing both PVTs, i.e. 11 failed TOMM only, 8 failed DS-SS only, and 4 failed both, therefore 23 failed ≥1 PVT.

Table 2 presents the PVT failure rate according to cut-offs and group comparisons of interest. As hypothesised, a small minority failed two PVTs (TOMM and DS-SS; 4%). The rate of TOMM failure (12%) was consistent with the hypothesis and previous findings (Whiteside et al., 2010), but the base rate of failure on any one PVT was somewhat greater than expected (TOMM or DS-SS; 18%).

Since there were only four participants in the 'two PVT fails' group, the analysis was conducted using a 'one or more PVT fails' group, or essentially PVT pass versus failure. Group differences were analysed in order to ensure that the one or more PVT fails group was not significantly skewed by the inclusion of the two PVT fails group.

Bonferroni adjustments were utilised throughout all analyses. Although when applied strictly the significance should be smaller than the critical *p* value, the result was considered significant if it was equal to or smaller than the critical *p* value. This was decided in an effort to reduce the likelihood of type II errors, since Bonferroni adjustments are considered a highly conservative method when applied to a high number of comparisons (Napierala, 2012).

Demographics Effects

Initial exploratory analyses were employed to examine the demographic variables for significant associations across the groups (PVT pass or fail, and TOMM pass or fail) and the PAI variables using Chi-square tests for independence[1], or Mann Whitney *U* tests to explore differences in group means. A Bonferroni correction was applied; the new familywise error rate to detect statistical significance was $p \leq .025$.

Table 3 shows that there were no significant differences in mean age across the PVT pass or fail groups, or the TOMM pass or fail groups. Chi-square tests for independence and Mann-Whitney *U* tests revealed no significant relationships between the PVT or TOMM pass or fail groups in relation to diagnosis category, time since injury, presence of identified financial incentive, or employment status. Furthermore, there were no significant relationships found across the TOMM groups in relation to gender. A significant relationship was found between the PVT pass and fail groups, and gender (χ^2 (1, N = 127) = 5.19, p=.023, phi = -.22). However, examination of crosstabulation indicated there was no meaningful gender difference found in the group of interest (PVT fail group).

Mann-Whitney *U* tests showed no significant differences across the PVT and TOMM pass and fail groups in terms of PMIQ. Finally, differences in current FSIQ across groups were explored using Mann-Whitney *U* tests. A significant difference was found in FSIQ scores between the PVT pass and fail groups. A significantly lower IQ score was found in the PVT fails group (*Mdn* = 89, *n* =22) in comparison to the pass group (*Mdn* = 100, *n* = 95), *U* = 637, *z* = -4.04, p<.000, *r* = -.37 (medium effect). FSIQ was also significantly lower in the TOMM fail group (*Mdn* = 89, *n* = 14) compared to the TOMM pass group (*Mdn* = 100, *n* = 103), *U* = 356, *z* = -3.07, p=.002, *r* = -.28 (medium effect).

[1] With Yates Continuity Correction

Table 4 presents a small number of statistically significant relationships between the demographic variables and the PAI variables after Bonferroni corrections (the new familywise error rate for the validity scales was $p \leq .013$ and $p \leq .003$ for the clinical scales). Gender was found to be significantly related to Antisocial score, with males scoring higher (M = 54.29, SD = 11.23, n =83) than females (M = 47.93, SD = 6.88, n = 40), r = .28, n = 123, p=.002. Age was significantly inversely correlated with Borderline score (r = -.29, p=.001) and Aggression score (r = -.31, p=.001). Time since injury was positively correlated with Negative Impression Management score (r = .26, p=.005).

Interestingly, there were a number of significant inverse relationships found between PMIQ score and the PAI variables, namely the Inconsistency scale (r = -.34, p<.000), Somatic Complaints (r = -.30, p = .001), Anxiety-Related Disorders (r = -.36, p<.000), Paranoia (r = -.28, p=.002), Borderline (r = -.31, p=.001), Antisocial (r = -.33, p<.000), Drug Problems (r = -.39, p<.000) and Aggression scores (r = -.30, p=.001). Furthermore, a number of significant inverse correlations were found between current FSIQ score and the PAI variables; including Inconsistency (r = -.34, p<.000), Somatic Complaints (r = -.32, p=.001), Anxiety-Related Disorders (r = -.32, p=.001), Drug Problems (r = -.37, p<.000) and Suicidality scales (r = -.30, p=.001).

In summary, initial analysis indicated no significant relationships between the demographic variables and PVT or TOMM groups, with the exception of current FSIQ. There were a small number of statistically significant correlations with the PAI variables, and FSIQ was significantly related to Inconsistency, Somatic Complaints, Anxiety-Related Disorders, Drug Problems and Suicidality scores. Given the overall lack of significant associations between the demographic variables and the performance validity variables, overall analysis utilising the entire sample was deemed appropriate.

Table 3
Demographic Effects for the PVT Variables

Demographic variable		*p*	
		PVT pass or fail	TOMM pass or fail
A	Gender	.023*	.160
B	Age	.385	.609
C	Diagnosis	.808	.235
D	Time since injury	.084	.048
E	Financial incentive	.836	.509
F	Employment status	.401	1.000
G	PMIQ	.037	.144
H	FSIQ	.000*	.002*

*Significant at the $p \leq .025$ level after Bonferroni corrections

Table 4
Demographic Effects for the PAI Variables

	p																					
	PAI validity scales (p<.013)											*PAI clinical scales (p<.003)*										
	INC	INF	NIM	PIM	SOM	ANX	ARD	DEP	MAN	PAR	SCZ	BOR	ANT	ALC	DRG	AGG	SUI	STR	NON	RXR	DOM	WAR
A	.319	.987	.696	.837	.018	.125	.124	.041	.570	.744	.770	.259	**.002***	.012	.967	.157	.107	.041	.126	.018	.545	.084
B	.826	.771	.027	.081	.823	.198	.067	.121	.011	.003	.243	**.001***	.009	.981	.833	**.001***	.188	.084	.171	.077	.686	.250
C	.254	.314	.342	.637	.542	.378	.483	.062	.706	.539	.301	.532	.181	.607	.172	.254	.183	.563	.292	.132	.169	.574
D	.481	.283	**.005***	.531	.049	.899	.576	.418	.521	.125	.104	.632	.984	.758	.331	.543	.205	.257	.512	.059	.043	.273
E	.795	.865	.043	.119	.065	.753	.900	.096	.233	.526	.414	.247	.032	.497	.146	.320	.749	.318	.907	.694	.378	.706
F	.012	.693	.386	.717	.240	.514	.909	.646	.613	.561	.747	.996	.947	.712	.251	.492	.117	.184	.503	.793	.667	.590
G	**.000***	.261	.014	.233	**.001***	.003	**.000***	.015	.056	**.002***	.055	**.001***	**.000***	.631	**.000***	**.001***	.046	.062	.445	.057	.365	.226
H	**.000***	.058	.004	.949	**.001***	.018	**.001***	.134	.171	.020	.090	.027	.084	.940	**.000***	.049	**.001***	.109	.220	.215	.958	.989

Note. A = gender; B = age; C = diagnosis category; D = time since injury; E = financial incentive; F = employment status; G = pre-morbid IQ; H = Full Scale IQ
PAI validity scales: INC = Inconsistency; INF = Infrequency; NIM = Negative Impression Management; PIM = Positive Impression Management
PAI clinical scales: SOM = Somatic Complaints; ANX = Anxiety; ARD = Anxiety-Related Disorders; DEP = Depression; MAN = Mania; PAR = Paranoia; SCZ = Schizophrenia; BOR = Borderline; ANT = Antisocial; ALC = Alcohol Problems; DRG = Drug Problems; AGG = Aggression; SUI = Suicide; STR = Stress; NON = Non-support; RXR = Treatment Rejection; DOM = Dominance; WAR = Warmth
*Validity scales significant at the $p \leq .013$ and clinical scales significant at the $p \leq .003$ level after Bonferroni corrections

Performance Validity and Symptom Validity

It was hypothesised that significant associations would be found between elevations on certain PAI validity scales (Negative Impression Management and Infrequency scales) and PVT failure (PVT pass and fail, and TOMM pass and fail groups). Spearman's correlation coefficients were calculated in order to identify significant relationships between the variables (Table 5).

After Bonferroni corrections (the new error rate was $p \leq .012$), results showed a medium positive correlation between PVT performance and the Negative Impression Management scale ($rs = .34$, $n = 123$, $p < .000$), with high scores associated with PVT fails. There was a significant positive relationship between Infrequency and PVT performance before Bonferroni corrections but not after, therefore this scale may be considered to be trending towards significance.

Table 5

Correlations between PVT Performance (Pass and One or More Fails) and SVT Performance

	PAI validity scales			
	INC	INF	NIM	PIM
Correlation coefficient	.10	.19	.34*	-.17
Sig. (2-tailed)	.294	.039	.000	.063
n	121	123	123	123

Note: INC = Inconsistency; INF = Infrequency; NIM = Negative Impression Management; PIM = Positive Impression Management

*significant at the $p \leq .012$ level after Bonferroni corrections

To investigate group differences in SVT performance, Mann-Whitney tests were utilised with PVT pass or fail as the grouping variable (see Table 6). A Bonferroni correction was applied; the new rate for significance was $p \leq .012$. A significant difference was found in

Negative Impression Management scores between the PVT pass and fail groups. A Mann-Whitney test revealed that scores were significantly higher in the PVT fail group (*Mdn* = 66, *n* = 20) than the pass group (*Mdn* = 55, *n* = 103), *U* = 754.0, *z* = -3.80, *p*<.000, *r* = .34 (medium effect; Cohen, 1992). Before Bonferroni corrections, Infrequency scores were significantly higher in the PVT fails group compared to the pass group. However, this result did not hold when adjusting for multiple comparisons.

The hypothesis was partially supported; Negative Impression Management scores were significantly higher in the one or more PVT fails group compared to PVT pass group, but there were no significant differences held in Infrequency scores after Bonferroni adjustments.

To demonstrate that this result was not being driven by the multiple PVT fails cases, the analysis was re-run exploring the PVT pass and one PVT fail groups. The same effect was found; Negative Impression Management scores were significantly higher in the one fail group (*Mdn* = 66, *n* = 17) than the pass group (*Mdn* = 55, *n* = 103); *U* = 716.5, *z* = -3.19, *p*=.001, *r* = -.29.

Since the bivariate statistical analyses (Spearman's correlation) and test of difference (Mann Whitney *U*) generated highly similar results, the test of difference analysis will be reported only for the remaining analysis.

Mann-Whitney tests were utilised with TOMM performance (pass or fail) as the grouping variable (see Table 6). Consistent with the PVT group, a significant difference was found in Negative Impression Management scores between TOMM pass and fail groups after Bonferroni corrections (the new familywise error rate was *p*≤.012). Negative Impression Management scores were significantly higher in the TOMM fail group (*Mdn* = 66, *n* = 13) than the TOMM pass group (*Mdn* = 55, *n* = 110), *U* = 368.0, *z* = -2.87, *p*=.004, *r* = .26 (small

to medium effect). There were no other significant group differences on any of the other PAI validity scales.

The hypothesis that Negative Impression Management would be significantly higher in the TOMM fail group compared to TOMM pass was supported. However, there were no significant differences in Infrequency scores across the groups.

Table 6

Group Comparisons for the SVT Variables (PAI Validity Scales)

	PAI validity scales			
	INC	INF	NIM	PIM
PVT pass/fail groups				
Mann-Whitney *U*	1190.0	1047.5	754.0	1080.0
Z	-1.05	-2.06	-3.80	-1.86
Sig. (2-tailed)	.292	.039	.000*	.063
TOMM pass/fail groups				
Mann-Whitney *U*	678.5	613.5	368	607
Z	-.20	-.84	-2.87	-.89
Sig. (2-tailed)	.844	.400	.004*	.373

Note: INC = Inconsistency; INF = Infrequency; NIM = Negative Impression Management; PIM = Positive Impression Management
*significant at the $p \leq .012$ level after Bonferroni corrections

Performance Validity, Personality and Psychological Functioning

It was hypothesised that significant associations would be found between PVT failure and elevations on Somatic Complaints, Depression, Anxiety, Anxiety-Related Disorders, and Schizophrenia PAI scales.

Mann-Whitney *U* tests were used with PVT pass or fail as the grouping variable (see Table 7). After Bonferroni corrections (the new error rate was $p \leq .003$), significant differences

were found in Schizophrenia, Anxiety-Related Disorders, and Paranoia scores between PVT pass and fail groups. Schizophrenia scores were significantly higher in the PVT fail group (*Mdn* = 63.5, *n* = 20) than the pass group (*Mdn* = 55, *n* = 104), *U* = 804.0, *z* = -3.54, *p*<*.000*, *r* = -.32 (medium effect). Similarly, Anxiety-Related Disorders scores were significantly higher in the PVT fail group (*Mdn* = 59.5, *n* = 20) than the pass group (*Mdn* = 53.5, *n* = 104), *U* = 887.5, *z* = -3.05, *p*=.002, *r* = -.27 (small to medium effect). Paranoia scores were also higher in the fails group (*Mdn* = 56.5, *n* = 20) than the pass group (*Mdn* = 50, *n* = 104), *U* = 908.5, *z* = -2.93, *p*=.003, *r* = -.26 (small to medium effect). Before Bonferroni corrections, Somatic Complaints, Anxiety, Depression, Borderline, Suicidality, Non-Support, Treatment Rejection and Warmth scores were significantly higher in the PVT fails group compared to the pass group.

Again, the analysis was re-run exploring the pass and one PVT fail group to explore the effect of the two PVT fails cases. Mann-Whitney *U* tests showed that scores were significantly higher in the one fail group compared to the pass group for Schizophrenia (*U* = 759.0, *z* = -2.95, *p*=.003, *r* = -.27), Paranoia (*U* = 824.0, *z* = -2.54, *p*=.011, *r* = -.23), and Anxiety-Related Disorders (*U* = 832.5, *z* = -2.48, *p*=.013, *r* = -.23). However, only Schizophrenia scores remained significant following Bonferroni corrections, suggesting the two fails cases had some impact on the Paranoia and Anxiety-Related Disorders scores in the analysis.

Exploratory *post-hoc* Mann-Whitney *U* tests utilizing the subscales from the PAI scales found to be significantly related to PVT failure (Anxiety-Related Disorders, Paranoia and Schizophrenia) were then run. The Anxiety-Related Disorders scale is comprised of Obsessive Compulsive Disorder, Phobias, and Traumatic Stress subscales, the Paranoia scale includes Hypervigilance, Persecution, and Resentment subscales, and finally the Schizophrenia scale contains Paranoia, Social Detachment, and Thought Disorder subscales.

Utilising the new familywise error rate of $p \leq .006$, only Paranoia-Hypervigilance scores were significantly higher in the one or more PVT fails group (*Mdn* = 57, *n* = 11) than the pass group (*Mdn* = 48, *n* = 57); *U* = 261.0, *z* = -2.81, *p*=.005, *r* = -.34 (medium effect).

The hypothesis was partially supported; Anxiety-Related Disorders and Schizophrenia scores were significantly higher in the one or more PVT fails group compared to the pass group, but there were no significant differences held in Somatic Complaints, Depression or Anxiety scores after Bonferroni adjustments. In addition, significantly higher Paranoia scores were found in the PVT fails group compared to the pass group. Furthermore, it was found that the Paranoia-Hypervigilance subscale specifically was greater in the PVT fail group than the pass. However, there were no other significant subscale group differences after adjustments for multiple comparisons.

To explore TOMM performance and the PAI clinical scales, Mann-Whitney tests were utilised, with TOMM pass or fail as the grouping variable (see Table 7). In contrast to the analysis of PVT performance, no significant differences were found in the PAI clinical scale scores between TOMM pass and fail groups after Bonferroni corrections ($p \leq .003$). Before adjustments for multiple comparisons, significantly higher Anxiety, Anxiety-Related Disorders, Depression, Schizophrenia, and Suicidality scores were found in the TOMM fail group in comparison to the pass group. Therefore, there appeared to be a trend in the data consistent with previous research (Whiteside et al., 2010). Trending subscales were analysed exploratively to see whether there were any group differences. Mann-Whitney *U* tests showed that no scores were significantly higher in the one or more PVT fails group in comparison to the pass group.

Table 7

Group Comparisons for the SVT Variables (PAI Clinical, Treatment Consideration and Interpersonal Scales)

	PAI scales																	
Group	SOM	ANX	ARD	DEP	MAN	PAR	SCZ	BOR	ANT	ALC	DRG	AGG	SUI	STR	NON	RXR	DOM	WAR
PVT pass or fail groups																		
Mann-Whitney U	1026.0	940.5	887.5	918.5	1398.0	908.5	804.0	944.5	1317.0	1193.5	1281.5	1200.0	992.5	1097.5	968.5	1041.5	1295.5	1042.0
Z	-2.24	-2.74	-3.05	-2.87	-0.07	-2.93	-3.54	-2.72	-0.46	-1.19	-0.68	-1.15	-2.38	-1.61	-2.52	-2.09	-0.59	-2.08
Sig. (2-tailed)	.025	.006	**.002***	.004	.944	**.003***	**.000***	.007	.646	.232	.498	.250	.017	.108	.012	.037	.557	.037
TOMM pass or fail groups																		
Mann-Whitney U	532.0	362.0	380.0	386.0	624.0	472.0	395.0	490.5	502.5	464.0	666.5	622.0	413.0	663.0	586.0	517.5	659.0	539.5
Z	-1.38	-2.81	-2.67	-2.62	-0.60	-1.89	-2.54	-1.73	-1.59	-1.93	-0.19	-0.57	-2.36	-0.17	-0.88	-1.46	-0.26	-1.28
Sig. (2-tailed)	.167	.005	.008	.009	.546	.059	.011	.083	.112	.054	.846	.568	.018	.867	.378	.143	.798	.202

Note. SOM = Somatic Complaints; ANX = Anxiety; ARD = Anxiety-Related Disorders; DEP = Depression; MAN = Mania; PAR = Paranoia; SCZ = Schizophrenia; BOR = Borderline; ANT = Antisocial; ALC = Alcohol Problems; DRG = Drug Problems; AGG = Aggression; SUI = Suicide; STR = Stress; NON = Non-support; RXR = Treatment Rejection; DOM = Dominance; WAR = Warmth
*significant at the adjusted $p \leq .003$ level

Finally, a hierarchical regression was utilised to explore whether PVT performance could be predicted by SVT performance (Negative Impression Management), or elevated psychopathological scales (Schizophrenia, Anxiety-Related Disorders and Paranoia). Preliminary analyses were conducted to check for violations of the assumptions of linearity, homoscedasticity, and unrestricted range. The assumption of normality of residuals may have been violated. It was decided to proceed on balance that findings are not usually vulnerable to effects of small deviations from normality (Tabachnick & Fidell, 2007); however, results should be interpreted with caution.

Table 8

Hierarchical Regression Model Predicting PVT Performance

	PVT performance				
Predictor	R^2	ΔR^2	β	*F*	*p*
Step 1	.095	.095		12.22	**.001****
NIM			.308		**.001****
Step 2	.122	.027		3.96	**.005***
SCZ			.126		.305
ARD			.003		.983
PAR			.158		.262

Note. NIM = Negative Impression Management; SCZ = Schizophrenia; ARD = Anxiety-Related Disorders; PAR = Paranoia

*$p < .01$, **$p < .001$

Negative Impression Management was entered at Step 1 explaining 9.5% of the variance in PVT performance (F (1, 117) = 12.22, p<.001). After entry of the Schizophrenia, Anxiety-Related Disorders, and Paranoia scales at Step 2 the total variance explained by the model as a whole was 12.2%, F (3, 114) = 3.96, p<.005. Only Negative Impression

Management made a unique significant contribution (β = .308, p<.001), with Schizophrenia, Anxiety-Related Disorders and Paranoia explaining an additional 2.7% of variance. This was a non-significant contribution, R^2 change = .027, F change (3, 114) = 1.184, p=.319.

Discussion

Base Rates of PVT Failure

The base rates of PVT failure found were in support of the hypotheses; TOMM failure was found in 12% of the sample, and 4% failed both PVTs. However, the rate of one or more failures on any PVT (TOMM or DS-SS) exceeded expectations and was in fact 18%. Bunnage et al. (2008) and Hampson et al. (2014) found base rates of PVT failure as high as 26% and 27% respectively in their NHS ABI samples using the Word Memory Test (Green, 2003). This is also an interesting finding considering a survey of UK neuropsychologists found that just 16% utilised PVTs in their clinical practice, believing base rates of PVT failure in clinical cases to be low (McCarter et al., 2009).

Demographics Effects

There were few relationships or group differences found in terms of PVT and PAI performance on the demographic variables. This is in support of the hypothesis and in line with previous research (Armistead-Jehle, 2010).

There were no significant relationships found between the demographic variables and performance validity, with the exception of current FSIQ; PVT failure was associated with significantly lower current FSIQ. This could be due to the PVTs used being sensitive to cognitive impairment, as suggested by Hampson et al. (2014). Alternatively, the result could be understood as consistent underperforming on both PVTs and other cognitive tests, including the measure of FSIQ.

There were a small number of statistically significant correlations with the PAI variables; males scored higher on the Antisocial scale than females, and younger participants scored higher on the Borderline and Aggression scales than older. These findings are unsurprising given young males are generally found to exhibit more disinhibited and aggressive behaviours (Dumais et al., 2005). In addition, participants with lower PMIQs scored higher on a number of PAI scales in comparison to higher PMIQ. Lower current FSIQ scores were similarly related to higher psychopathology scores. This may indicate a greater vulnerability to mental distress in those with lower cognitive functioning, which is supported by the intellectual disabilities literature (Smiley, 2005). Finally, a positive relationship was found between Negative Impression Management and the time since injury, which could be suggestive of a 'cry for help' related to chronicity of problematic brain injury sequalae.

The hypothesis that the PVT failure rate would be increased where financial incentive had been identified was not supported by the data; no group differences were found, consistent with Suhr et al. (1997). This is a highly interesting finding as previous research has focused on malingering as an explanation for PVT failure (Bianchini et al., 2006). This hypothesis cannot be ruled out due to the possibility of other psychosocial incentives operating, for example, time off work, or access to services. However, the inclusion of this variable in the current study nevertheless presents a novel addition to the literature and provides some information regarding certain types of external incentives.

Performance Validity and Symptom Validity

Both PVT and TOMM failure groups were found to be significantly associated with higher scores on the Negative Impression Management scale of the PAI. This finding supports the view that elevations on this scale can be expected in individuals performing below threshold on PVTs, and that PVTs and SVTs are related, consistent with Whiteside et al. (2009) and Haggerty et al. (2007). However, the effect size for group differences between

TOMM performance and Negative Impression Management in the current study was slightly smaller than Whiteside and colleagues (2009).

Although the PAI Infrequency scale has previously been found to be related to PVT failure (Whiteside et al., 2009), this was not supported in the current study and could be attributable to differences in sample characteristics. There were no significant associations found between the other PAI validity scales (Positive Impression Management and Inconsistency) and PVT failure. This may suggest exaggeration of difficulties in the PVT fail group, since there were no indications of validity threats otherwise.

These findings endorse elevated SVT performance as a useful indicator for risk of cognitive PVT failure. Likewise, PVT failure may indicate exaggerated responding on SVTs. However, it is suggested that neither PVTs nor SVTs can provide comprehensive information pertaining to the intentions and motivations underlying test-taking behaviour, and that evaluation of both domains remains valuable in contributing to a comprehensive biopsychosocial formulation (McMillan et al., 2009).

Performance Validity, Personality and Psychological Functioning

For the PVT failure group, the hypothesis was partially supported; Anxiety-Related Disorders and Schizophrenia scores were significantly higher in participants who failed any one PVT in comparison to the pass group. Furthermore, Paranoia scores were found to be higher in the PVT fail group compared to pass, which had not been expected based on previous research. Analysis of subscales revealed that Paranoia-Hypervigilance scores were driving this group difference. This is partially consistent with the findings of Sumanti et al. (2005) who showed that PVT failures were related to elevated scores on PAI Somatic Complaints, Depression, Anxiety, Anxiety-Related Disorders, and Schizophrenia scales in a psychiatric sample.

It was further found that only Negative Impression Management was able to significantly predict PVT performance; none of the PAI clinical scales made significant contributions to the predictive model. However, findings should be interpreted with caution due to potential violation of the assumption of normality of residuals.

In contrast, no significant differences were found in the expected PAI clinical scales between TOMM pass and fail groups. Nevertheless, before adjustments for multiple comparisons, the Depression, Anxiety, Anxiety-Related Disorders, Schizophrenia and Suicidality scales appeared to be trending towards significance. This is comparable to the findings of Whiteside et al. (2010) in their US study utilising the TOMM. Furthermore, it was found that there were no significant differences in terms of subscale scores between the groups. This suggested that the significant group differences on Anxiety-Related Disorders, Schizophrenia, and Paranoia (Hypervigilance) in the PVT analysis were driven by Digit Span performance rather than TOMM performance. It may be that Digit Span suffers from weaker sensitivity and specificity; however, the elevated PAI scales generally concur with other findings in the literature, which decreases the likelihood that significant findings are the result of type I error.

Although significant relationships were found between PVT failure and self-reported psychological symptoms, it is not possible to infer the causality of PVT failure. However, since Negative Impression Management was the only validity scale significantly related to PVT performance, and was the only significant unique predictor of PVT performance, on balance it seems likely that scale elevations were subject to at least some level of symptom exaggeration. It is unclear why these scales would be subject to a greater level of exaggeration than others; Appendix F can be referred to for the PAI items comprising each scale and subscale. Negative impression management could also be understood in the context of the experience of stigma and shame following brain injury (Hagger & Riley, 2017). Nochi

(1998) explored the impact of undergoing neuropsychological assessment on the self-image of TBI survivors and argued that ongoing complications, such as litigation, can result in adjustment difficulties and feelings of helplessness and persecution. The author suggested that the assessment process can provide opportunity for individuals to communicate and legitimise their struggles. It may be that the less visible, non-physical consequences of ABI, which may have been tapped by the PAI, create a need for individuals to communicate a more negative impression to professionals in order to get their needs met. In addition, it could be hypothesised that impression management requires a level of performance monitoring that may be impaired following ABI, and particularly in frontal lobe injuries (Rabinowitz & Levin, 2014).

It could be argued that some of the items included on the Anxiety-Related Disorders, Schizophrenia and Paranoia scales are related to brain injury sequalae or cognitive impairment. For example, within the Anxiety-Related Disorders scale, 'I have impulses that I fight to keep under control' could be understood as relating to problems with disinhibition rather than OCD. The items within the Traumatic Stress subscale may relate to sustaining the brain injury itself and ongoing difficulties in this population, for example, 'I can't seem to get over some things from my past'. Elevations on the Phobia subscale may have been due to indirect consequences of the ABI, for example, 'I don't mind driving on freeways'. Furthermore, the Schizophrenia scale could be considered to tap into cognitive or social difficulties arising from brain trauma, for example 'My thinking has become confused', and 'I just don't seem to relate to people very well'. Items on the Paranoia scale are themed around Hypervigilance, Persecution and Resentment. Social and interpersonal difficulties such as irritability, and poor social communication and social problem-solving skills are common after brain injury (Schoenberg & Scott, 2011). It could be that the group who performed more poorly on validity tests were experiencing a greater level of these difficulties

or adjustment difficulties in comparison to those who performed well. Till, Christensen and Green (2009) explored the use of the PAI with ABI populations and similarly found a number of transdiagnostic items on the Schizophrenia, Depression, and Somatic Complaints scales, noting items related to "anti-social behaviours, history of substance abuse and psychiatric problems of an anxiety-related and paranoid nature" (p. 663). The authors concluded that high levels of psychopathology and personality disturbance are often found on measures used with individuals with ABIs, but that this may be attributable to the cognitive and physical sequelae of the injury rather than representing psychiatric disorder. It is also acknowledged that there exists a high prevalence of comorbidity in ABI populations (Rogers & Read, 2007). The current study supports the use of caution when interpreting elevations on the Schizophrenia scale when using the PAI with individuals with acquired brain injuries (Morey, 2003), and additionally endorses the careful interpretation of elevations on the Anxiety-Related Disorders and Paranoia scales.

This explanation may be corroborated by the finding that the median FSIQ score was 11 points lower in the PVT failure group compared to the pass group (taking the fail group into the 'low average' IQ category from the 'average' category). Since PMIQ scores were comparable across the PVT pass and fail groups, it seemed more likely that the result was due to cognitive impairment arising from the brain injury, or secondary to higher levels of exaggeration.

The lack of relationship between PVT performance and somatic preoccupation in the current study is puzzling since somatization has historically been found to be the most consistently elevated scale in those performing poorly on PVTs (Sumanti et al., 2005; Boone & Lu, 1999; Whiteside et al., 2010). This finding refutes the idea that motivation during neuropsychological assessment is mediated by understanding and response to physiological symptoms (Whiteside et al., 2010; Boone & Lu, 2010). It may be that PVTs in the current

study tapped a construct other than motivation, such as complex cognitive or neuropsychological sequalae resulting from brain injury. In addition, no significant elevations were found in depression and anxiety scores in this sample, suggesting that PVT failure is unlikely to be attributable to affective distress. This was consistent with previous findings (Ashendorf et al, 2004). Research suggesting that cognitive performance is dependent on the interaction between PVT failure and psychological symptomatology (Green, Rohling, Allen, & Iverson, 2001) may be applicable. The current findings may reflect an interaction of factors that underlie both neuropsychological test performance, psychological symptom reporting and 'effort' rather than linear, causal relationships.

The findings also relate to the broader literature on experiences of neuropsychological assessments. Keady and Gilliard (2002) explored service-users' experiences of dementia assessments and identified a high prevalence of anxiety and uncertainty. The authors argued that feelings of perceived threat, particularly in the context of poor rapport with the examiner, can lead to the adoption of coping strategies to create distance, such as defensiveness, confrontation, resistance, and passivity. This was found to be exacerbated by cognitive fatigue. Those struggling more with the cognitive, emotional and behavioural sequalae of brain injury, and particularly when insight into difficulties is high, may experience the assessment as particularly distressing (Paterson & Scott-Findlay, 2002). This may further contextualise performance.

Knowledge of the base rate of PVT failure and relationship with psychological functioning in UK neuropsychology settings can offer clinicians a potentially useful tool in assessing the extent to which test performance can be confidently attributed to brain injury. Although PVT failure in itself cannot definitively identify invalid performance due to PVTs being imperfect measures, it could stimulate further and more nuanced exploration of an individual's needs.

Limitations and Future Directions

There are a number of limitations to be borne in mind when evaluating the conclusions of the present study. Firstly, the archival nature of the research limits experimental manipulation of variables and, therefore, no causal statements can be made. However, the benefits of using naturalistic clinical data lie in its ecological validity, and consequent generalisability to NHS neuropsychology practice.

A further limitation concerns the lack of data regarding participants' ethnicities, years of education, and English not being the first language; risk factors for PVT failure (Victor et al., 2009). Although the tests of pre-morbid functioning utilised have been validated for use with brain injury populations (Green et al., 2008; Franzen et al., 1997) these could also have been subject to biased responding. Future studies would likely benefit from gathering such demographic data. The pooling of PMIQ and FSIQ data based upon different tests and versions also presents an important limitation. However, analysis on the impact of the test or version suggested no significant effect on study variables.

Despite the sample size being comparable to, and often exceeding, published literature in the field (Locke et al, 2008, Van Dyke et al., 2013), groups were unequal due to the nature of PVT failure. Fidelity to Larrabee's criterion for detecting invalid responding was intended; however, the classification was relaxed from two or more PVT fails to any one PVT fail due to small numbers in the comparison group. This increased the likelihood of type I error. Potential misclassifications represent a pervasive challenge for all performance validity research (Hawes & Boccaccini, 2009). Furthermore, the current study was not concerned with malingering diagnosis per se, but rather with the meaning of PVT failure. It has been shown that when even one PVT is failed, the correlation between cognitive test performance and the documented brain injury is lost (Fox, 2011).

The use of cut-off scores may pose a methodological problem in the current research; non-neurological test-taking behaviour is increasingly being considered to be on a continuum rather than a binary taxonomy (Bigler, 2012). Above-chance but below cut-off performance is potentially an important future direction for validity research, particularly in clinical settings. Additionally, more recently developed supplementary PAI scales such as the Malingering Index and Defensiveness Index (Morey, 2007) were not available in the current study but would pose valuable lines of enquiry in future research.

It is hoped the current findings will inform UK psychologists in their clinical practice, and their decision to use PVTs, interpretation of test data, and wider biopsychosocial formulation. Research on this topic, as well as clinical practice, would benefit from continuing to explore the non-neurological factors influencing performance, rather than focussing on malingering in isolation. As Iverson and Binder (2000) propose, "the well-informed clinician will seek to identify all variables that may affect symptom reporting or neuropsychological test performance and be careful not to over- or under-interpret evidence of negative response bias" (p. 853). It is also suggested that more patient-centred, qualitative lines of enquiry may be particularly informative in our understanding of patients' needs and this complex construct.

Conclusions

This thesis endorses the view that PVT failure occurs in a sizable minority of NHS ABI patients, which is unlikely to be simply explained by malingering for financial gain. It is suggested that further exploration of interactions between psychogenic factors and validity test performance could reduce false positive diagnoses and associated invalid recommendations. Although the study is limited by methodological issues related to naturalistic design, within this design also lies its strengths. It is hoped that the findings will

be directly applicable to current NHS clinical neuropsychology practice, and contribute to the provision of comprehensive and valid assessments of those who use these services.

References

Armistead-Jehle, P. (2010). Symptom validity test performance in US veterans referred for evaluation of mild TBI. *Applied Neuropsychology*, *17*(1), 52-59. https://doi.org/10.1080/09084280903526182

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. https://doi.org/10.1016/S0887-6177(02)00218-4

Axelrod, B. N., Fichtenberg, N. L., Millis, S. R., & Wertheimer, J. C. (2006). Detecting incomplete effort with Digit Span from the Wechsler Adult Intelligence Scale - Third Edition. *The Clinical Neuropsychologist*, *20*(3), 513-523. https://doi.org/10.1080/13854040590967117

Crawford, J. R. (2003). Psychometric foundations of neuropsychological assessment. In L. H. Goldstein & J. McNeil (Eds.), *Clinical neuropsychology: A practical guide to assessment and management for clinicians*. Chichester, UK: Wiley.

Carone, D. A. & Bush, S. S. (2013). *Mild traumatic brain injury: Symptom validity assessment and malingering*. New York, NY: Springer.

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. https://doi.org/10.1080/13854040600875203

Bigler, E. D. (2012). Symptom validity testing, effort, and neuropsychological assessment. *Journal of the International Neuropsychological Society*, *18*(4), 632-640. https://doi.org/10.1017/S1355617712000252

Binder, L. M., & Rohling, M. L. (1996). Money matters: A meta-analytic review of the effects of financial incentives on recovery after closed-head injury. *The American Journal of Psychiatry*, *153*(1), 7-10. Retrieved from https://ajp.psychiatryonline.org/

Boone, K. B. (Ed.). (2007). *Assessment of feigned cognitive impairment: A neuropsychological perspective*. New York, NY: Guilford Press.

Boone, K. B., & Lu, P. H. (1999). Impact of somatoform symptomatology on credibility of cognitive performance. *The Clinical Neuropsychologist*, *13*(4), 414-419. https://doi.org/10.1076/1385-4046(199911)13:04;1-Y;FT414

Boyle, G. J., & Lennon, T. J. (1994). Examination of the reliability and validity of the Personality Assessment Inventory. *Journal of Psychopathology and Behavioral Assessment*, *16*(3), 173-187. http://dx.doi.org/10.1007/BF02229206

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. https://doi.org/10.1177/1073191110385316

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*(1), 138-139. Retrieved from https://www.cambridge.org/core/journals/journal-of-the-international-neuropsychological-society

Bush, S. S., Ruff, R. M., Troster, A., Barth, J., Koffler, S. P., Pliskin, N. H., & Silver, C. H. (2005). NAN position paper: Symptom validity assessment: Practice issues and medical necessity. *Archives of Clinical Neuropsychology*, *20*(4), 419-426. https://doi.org/10.1016/j.acn.2005.02.002

Butcher, J. N., Graham, J. R., Ben-Porath, Y. S., Tellegen, A., & Dahlstrom, W. G. (2001). *MMPI-2: Minnesota Multiphasic Personality Inventory-2*. Minnesota, MN: University of Minnesota Press.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159. Retrieved from http://www.apa.org/pubs/journals/bul/

Constantinou, M., Bauer, L., Ashendorf, L., Fisher, J. M., & McCaffrey, R. J. (2005). Is poor performance on recognition memory effort measures indicative of generalized poor performance on neuropsychological tests?. *Archives of Clinical Neuropsychology*, *20*(2), 191-198. https://doi.org/10.1016/j.acn.2004.06.002

Crawford, J. R. (2013). Quantitative aspects of neuropsychological assessment. In Goldstein, L. H. & McNeil, J. (Eds.), *Clinical Neuropsychology: A Practical Guide to Assessment and Management for Clinicians* (pp. 129-159). Chichester, UK: Wiley.

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. https://doi.org/10.1093/arclin/act073

Demakis, G. J., Gervais, R. O., & Rohling, M. L. (2008). The effect of failure on cognitive and psychological symptom validity tests in litigants with symptoms of post-traumatic stress disorder. *The Clinical Neuropsychologist*, *22*(5), 879-895. https://doi.org/10.1080/13854040701564482

Demakis, G. J., Hammond, F., Knotts, A., Cooper, D. B., Clement, P., Kennedy, J., & Sawyer, T. (2007). The Personality Assessment Inventory in individuals with traumatic brain injury. *Archives of Clinical Neuropsychology*, *22*(1), 123-130. https://doi.org/10.1016/j.acn.2006.09.004

Dumais, A., Lesage, A. D., Alda, M., Rouleau, G., Dumont, M., Chawky, N., ... & Turecki, G. (2005). Risk factors for suicide completion in major depression: A case-control study of impulsive and aggressive behaviors in men. *American Journal of Psychiatry*, *162*(11), 2116-2124. https://doi.org/10.1176/appi.ajp.162.11.2116

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175-191. https://doi.org/10.3758/BF03193146

Faust, D., Hart, K. J., Guilmette, T.J., & Arkes, H. R. (1988). Neuropsychologists' capacity to detect adolescent malingerers. *Professional Psychology: Research and Practice, 19,* 508-51. Retrieved from http://www.apa.org/pubs/journals/pro/

Fox, D. D. (2011). Symptom validity test failure indicates invalidity of neuropsychological tests. *Clinical Neuropsychologist, 25*(3), 488-495. https://doi.org/10.1080/13854046.2011.554443

Franzen, M. D., Burgess, E. J., & Smith-Seemiller, L. (1997). Methods of estimating premorbid functioning. *Archives of Clinical Neuropsychology*, *12*(8), 711-738. https://doi.org/10.1016/S0887-6177(97)00046-2

Gaasedelen, O. J., Whiteside, D. M., & Basso, M. (2017). Exploring the sensitivity of the Personality Assessment Inventory symptom validity tests in detecting response bias in a mixed neuropsychological outpatient sample. *The Clinical Neuropsychologist*, *31*(5), 844-856. https://doi.org/10.1080/13854046.2017.1312700

Green, P. (2003). *Green's Word Memory Test*. Kelowna, BC, Canada: Green's Publishing.

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. https://doi.org/10.1080/02699050110088254

Green, R. E., Melo, B., Christensen, B., Ngo, L. A., Monette, G., & Bradbury, C. (2008). Measuring premorbid IQ in traumatic brain injury: An examination of the validity of the Wechsler Test of Adult Reading (WTAR). *Journal of Clinical and Experimental Neuropsychology*, *30*(2), 163-172. https://doi.org/10.1080/13803390701300524

Green, P. & Merten, T. (2013). Noncredible explanations of noncredible performance on symptom validity tests. In Carone, D. A. & Bush, S. S. (Ed.), *Mild traumatic brain injury: Symptom validity assessment and malingering* (pp.73-96). New York, NY: Springer.

Greiffenstein, M. F., Baker, W. J., & Gola, T. (1994). Validation of malingered amnesia measures with a large clinical sample. *Psychological Assessment, 6*(3), 218-224. http://dx.doi.org/10.1037/1040-3590.6.3.218

Haber, A. H., & Fichtenberg, N. L. (2006). Replication of the Test of Memory Malingering (TOMM) in a traumatic brain injury and head trauma sample. *The Clinical Neuropsychologist*, *20*(3), 524-532. https://doi.org/10.1080/13854040590967595

Hagger, B. F., & Riley, G. A. (2017). The social consequences of stigma-related self-concealment after acquired brain injury. *Neuropsychological Rehabilitation, 27,* 1-20. https://doi.org/10.1080/09602011.2017.1375416

Haggerty, K. A., Frazier, T. W., Busch, R. M., & Naugle, R. I. (2007). Relationships among Victoria Symptom Validity Test indices and Personality Assessment Inventory validity scales in a large clinical sample. *The Clinical Neuropsychologist*, *21*(6), 917-928. https://doi.org/10.1080/13854040600899724

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. https://doi.org/10.1080/09084282.2013.787425

Hawes, S. W., & Boccaccini, M. T. (2009). Detection of overreporting of psychopathology on the Personality Assessment Inventory: A meta-analytic review. *Psychological Assessment*, *21*(1), 112-124. http://dx.doi.org/10.1037/a0015036

IBM Corp. (2013). *IBM SPSS Statistics for Windows, Version 24.0*. Armonk, NY: IBM Corp.

Iverson, G. L. (2005). Outcome from mild traumatic brain injury. *Current Opinion in Psychiatry*, *18*(3), 301-317. doi:10.1097/01.yco.0000165601.29047.ae

Iverson, G. L., & Binder, L. M. (2000). Detecting exaggeration and malingering in neuropsychological assessment. *The Journal of Head Trauma Rehabilitation*, *15*(2), 829-858. http://dx.doi.org/10.1097/00001199-200004000-00006

Keady, J., & Gilliard, J. (2002). The experience of neuropsychological assessment for people with suspected Alzheimer's disease. In Harris, P. B. (Ed.), *The person with Alzheimer's disease: Pathways to understanding the experience* (pp. 3-28). Maryland, MD: John Hopkins University Press.

Larrabee, G. J. (2003). Detection of malingering using atypical performance patterns on standard neuropsychological tests. *Clinical Neuropsychologist, 17*(3), 410-425. https://doi.org/10.1076/clin.17.3.410.18089

Larrabee, G. J. (2012). Performance validity and symptom validity in neuropsychological assessment. *Journal of the International Neuropsychological Society, 18*(04), 625-630. https://doi.org/10.1017/S1355617712000240

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. doi:10.1093/arclin/acu019

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. Retrieved from https://www.ncbi.nlm.nih.gov/labs/journals/neurorehabilitation/

McCarter, R. J., Walton, N. H., Brooks, D. N., & Powell, G. E. (2009). Effort testing in contemporary UK neuropsychological practice. *The Clinical Neuropsychologist*, *23*(6), 1050-1066. https://doi.org/10.1080/13854040802665790

McMillan, T. M., Anderson, S., Baker, G., Berger, M., Powell, G. E., & Knight, R. (2009). *Assessment of effort in clinical testing of cognitive functioning for adults*. Leicester, UK: The British Psychological Society.

Moore, B. A., & Donders, J. (2004). Predictors of invalid neuropsychological test performance after traumatic brain injury. *Brain Injury*, *18*(10), 975-984. https://doi.org/10.1080/02699050410001672350

Morey, L. C. (1991). *Personality Assessment Inventory professional manual.* Odessa, FL: Psychological Assessment Resources.

Morey, L. C. (2003). *Essentials of PAI interpretation.* New York, NY: Wiley.

Morey, L. C. (2007). *Personality Assessment Inventory professional manual* (2nd ed.). Lutz, FL: Psychological Assessment Resources.

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), 815-22. doi:10.1037/a0028195

Napierala, M. A. (2012). What is the Bonferroni correction. *AAOS Now*, *6*(4), 40-41. Retrieved from https://www.aaos.org/aaosnow/?ssopc=1

Nochi, M. (1998). Struggling with the labelled self: People with traumatic brain injuries in social settings. *Qualitative Health Research*, *8*(5), 665-681. https://doi.org/10.1177/104973239800800507

Paterson, B., & Scott-Findlay, S. (2002). Critical issues in interviewing people with traumatic brain injury. *Qualitative Health Research*, *12*(3), 399-409. https://doi.org/10.1177/104973202129119973

Rabinowitz, A. R., & Levin, H. S. (2014). Cognitive sequelae of traumatic brain injury. *The Psychiatric Clinics of North America*, *37*(1), 1-11. doi:10.1016/j.psc.2013.11.004

Robbins, J. (2014). *The neuropsychological application of the WAIS-IV over the WAIS-III*. Retrieved from http://nsuworks.nova.edu/cps_stuetd/91

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. https://doi.org/10.1016/S0887-6177(01)00109-3

Rogers, J. M., & Read, C. A. (2007). Psychiatric comorbidity following traumatic brain injury. *Brain Injury*, *21*(13), 1321-1333. https://doi.org/10.1080/02699050701765700

Rogers, R., Flores, J., Ustad, K., & Sewell, K. W. (1995). Initial validation of the personality assessment inventory - Spanish version with clients from Mexican American communities. *Journal of Personality Assessment*, *64*(2), 340-348. https://doi.org/10.1207/s15327752jpa6402_12

Schoenberg, M. R. & Scott, J. G. (2011). *The little black book of neuropsychology: A syndrome-based approach.* New York, NY: Springer.

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. https://doi.org/10.1076/1385-4046(199911)13:04;1-Y;FT545

Smiley, E. (2005). Epidemiology of mental health problems in adults with learning disability: an update. *Advances in Psychiatric Treatment*, *11*(3), 214-222. https://doi.org/10.1192/apt.11.3.214

Suhr, J. A., & Gunstad, J. (2005). Further exploration of the effect of "diagnosis threat" on cognitive performance in individuals with mild head injury. *Journal of the International Neuropsychological Society*, *11*(1), 23-29. https://doi.org/10.1017/S1355617705050010

Suhr, J., Tranel, D., Wefel, J., & Barrash, J. (1997). Memory performance after head injury: Contributions of malingering, litigation status, psychological factors, and medication use. *Journal of Clinical and Experimental Neuropsychology*, *19*(4), 500-514. https://doi.org/10.1080/01688639708403740

Sumanti, M., Boone, K. B., Savodnik, I., & Gorsuch, R. (2006). Noncredible psychiatric and cognitive symptoms in a workers' compensation "stress" claim sample. *The Clinical Neuropsychologist*, *20*(4), 754-765. doi:10.1080/13854040500428467

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics (5th ed.)*. Boston, MA: Allyn & Bacon/Pearson Education.

Till, C., Christensen, B. K., & Green, R. E. (2009). Use of the Personality Assessment Inventory (PAI) in individuals with traumatic brain injury. *Brain Injury*, *23*(7), 655-665. doi:10.1080/02699050902970794

Tombaugh, T. (1996). *Test of Memory Malingering*. Toronto, 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. http://dx.doi.org/10.1037/1040-3590.9.3.260

Van Dyke, S. A., Millis, S. R., Axelrod, B. N., & Hanks, R. A. (2013). Assessing effort: Differentiating performance and symptom validity. *The Clinical Neuropsychologist*, *27*(8), 1234-1246. https://doi.org/10.1080/13854046.2013.835447

Veiel, H. O. (1997). A preliminary profile of neuropsychological deficits associated with major depression. *Journal of Clinical and Experimental Neuropsychology*, *19*(4), 587-603. https://doi.org/10.1080/01688639708403745

Victor, T. L., Boone, K. B., Serpa, J. G., Buehler, J., & Ziegler, E. A. (2009). Interpreting the meaning of multiple symptom validity test failure. *The Clinical Neuropsychologist*, *23*(2), 297-313. https://doi.org/10.1080/13854040802232682

Wechsler, D. (1997). *Wechsler Adult Intelligence Scale – Third Edition.* San Antonio, TX: Pearson Assessment.

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

Wechsler, D. (2010) *Wechsler Adult Intelligence Scale – Fourth Edition*. San Antonio, TX: Pearson Assessment.

Wechsler, D. (2011). *Test of Premorbid Functioning. UK version (TOPF UK).* London, UK: Pearson Assessment.

Whiteside, D. M., Clinton, C., Diamonti, C., Stroemel, J., White, C., Zimberoff, A., & Waters, D. (2010). Relationship between suboptimal cognitive effort and the clinical scales of the Personality Assessment Inventory. *The Clinical Neuropsychologist*, *24*(2), 315-325. doi:10.1080/13854040903482822

Whiteside, D. M., Dunbar-Mayer, P., & Waters, D. P. (2009). Relationship between TOMM performance and PAI validity scales in a mixed clinical sample. *The Clinical Neuropsychologist*, *23*(3), 523-533. doi:10.1080/13854040802389169

Section C

Appendices of Supporting Material

Appendix A

Response Validity Terminology

In the current thesis, the following terminology has been utilised, based upon definitions provided by Larrabee (2012), McMillan et al. (2009), Bush et al. (2005) and Bigler (2012);

- **Response validity** – an umbrella term concerning the validity of responses (both performance and symptom responses);
- **Performance validity** – the validity of performance on cognitive tasks;
- **Symptom validity** – the validity of self-reported symptoms;
- **Effort** – "Motivation to comply with implicit or explicit test instructions with regard to speed, accuracy or other performance requirement. Failure on a test of effort means that someone has performed poorly on the test (below a suitable cut-off or low absolute score), and where the test was appropriate for that person, that they performed below their capability as determined by other criteria" (BPS guidance; McMillan et al., 2009, p. 18). "'Failure' reflects non-neurological factors that reduce neuropsychological test scores and invalidates findings" (Bigler, 2012, p. 632);
- **Malingering** – "The intentional production of false or exaggerated symptoms, motivated by external incentives. Although symptom validity tests are commonly referred to as malingering tests, malingering is just one possible cause of invalid performance" (NAN Position Paper; Bush et al., 2005, p. 420);
- **Response bias** – "An attempt to mislead the examiner through inaccurate or incomplete responses or effort" (NAN Position Paper; Bush et al., 2005, p. 420);
- **Dissimulation** – "The falsification or misrepresentation of symptoms by over representation or under representation, with an intention to appear different from the 'true' state" (BPS guidance; McMillan et al., 2009, p. 18).

Appendix B

Summary Table of Included Studies (Section A)

Study	**Design/Sample**	**Key Findings**			**Limitations**
		Base Rates of SPVT Failure	*Clinical Use of SPVTs*	*Beliefs Regarding SPVT Use*	
Hirst et al. (2017)	*N* = 654 Licensed clinical psychologists with ≥100 post-grad clinical neuropsychological assessment hours; 21% were board certified in neuropsychology Online survey emailed to National Academy of Neuropsychology (NAN) and International Neuropsychological Society (INS) members 9% response rate Approximately half conducted assessments in secondary gain contexts such as forensic work, disability claimants, and VA hospitals	Respondents who followed recommendations reported significantly higher base rates of probable poor effort (22% versus 18%, *p*=.008), definite poor effort (15% versus 10%, *p*<.001), probable malingering (10% versus 7%, *p*=.002), & estimated base rates of definite malingering (7% versus 5%, *p*=.028) than respondents who did not follow practice recommendations	International psychologists adhere to NAN/AACN recommendations as well as US psychologists Clinicians working with paediatric and geriatric populations did not follow the guidance as closely as those working with adults More experienced neuropsychologists were less likely to adhere to guidance An average test battery included 6 embedded/stand-alone SPVTs. Those who adhered to recommendations typically employed a significantly greater number of validity measures than those who did not adhere (average of 10 measures compared to 5.8, *p*<.001) Majority of both US and international respondents indicated that they mostly or always encouraged examinees to give their best effort (91% and 88% respectively) US respondents were significantly less likely than non-US to provide explicit warning that effort tests would be used (22% and 32% respectively, *p*=.003) Sample was divided on whether to provide	70% believed that an SPVT should be included in every assessment A greater proportion of less experienced respondents believed that every test battery should contain validity testing than more experienced respondents (78% and 62% respectively, *p*<.000)	Unstandardized survey limited by self-selection bias and self-report Despite aiming to investigate international adherence to guidelines, only 17% of responses were international and survey was only disseminated in English Low response rate so may only reflect opinions of those motivated by the topic Failed to report average number of assessments conducted in the

	Mean years of practice = 16		immediate feedback, as well on whether to discontinue the assessment if detecting suspect effort. Approximately one third indicated they mostly/always continued to interpret the assessment even when SPVTs had been failed		last year
Brooks, Ploetz, & Kirkwood (2016)	*N* = 282 Neuropsychologists working with children/adolescents in North America/Canada Online survey emailed via professional listservs, open for 8 weeks Mean years of practice = 12 Majority conducted clinical assessments but a third also conducted forensic assessments Mean number of assessments performed yearly = 102	Not reported	Majority utilised clinical judgement methods; 92% endorsed behavioural observations of poor compliance, 90% endorsed discrepancies between records, self-report and observed behaviours 92% used at least one stand-alone or PVT and 88% used at least one SVT per assessment, 60% used embedded validity tests. An average assessment included 1 stand-alone PVT, 1-2 embedded PVTs, and 1-2 embedded SVTs Number of validity tests used was not affected by clinicians' level of training Those who conducted forensic assessments administered more SPVTs (Cohen's *d*=.57) Most utilised stand-alone PVTs were the Test of Memory Malingering (TOMM), Medical Symptom Validity Test (MSVT), and Word Memory Test (WMT). Most utilised embedded PVTs were the Reliable Digit Span (RDS) & California Verbal Learning Test (CVLT). Most utilised SVTs were the Behavior Rating Inventory of Executive Function (BRIEF) validity indicators & Behavior Assessment System for Children-2 (BASC-2) validity indicators 95% often/always encouraged examinees to give their best "effort"; 76% never/rarely warned examinees that tests are sensitive to exaggeration	71% believed validity testing to be mandatory in forensic assessments, 53% believed should be mandatory in psychiatric facilities, half believed desirable in schools SPVTs felt to be possible to administer and not unnecessary in any setting 76% utilised SPVTs due to research evidence, 68% in order to validate other test scores, 64% due to own experience supporting use, & 18% due to third party instruction Not utilising SPVTs was most commonly due to difficulty in interpretation for very young children and children with severe cognitive impairment	Unstandardized survey limited by self-selection bias and self-report Potential sampling bias due to methodology; however, authors addressed by inviting views of both practitioners regularly using SPVTs and those who did not Insufficient responses gained for practitioners working with the youngest children (below five years), limiting generalisability Not possible to report response rate due to design Unclear survey wording may have produced

			40% often/always reported that results were 'inconsistent with severity of condition', and 93% never/rarely use the term 'malingering'		anomalous results
Schroeder, Martin, & Odland (2016)	*N* = 24 Online survey emailed to North American neuropsychologists/experts (defined as being first author on four recent papers regarding validity testing or participation in the AACN response validity conference) 50% response rate 92% conducted clinical assessments and 91% forensic assessments 87% primarily evaluated adults and 12% worked across the lifespan Mean years of practice = 20	Not reported	Experts had similar beliefs and practices to non-expert neuropsychologists in a prior study (Martin, Schroeder, & Odland, 2015) Respondents used a broad range of methods to identify suspect performance More than 90% of experts often/almost always used both stand-alone and embedded SPVTs Majority reported that in 95% of cases they gave more weight to SPVT results that their initial clinical judgement 95% often/always encouraged examinees to try their best, and 25% often/always directly warn that tests are sensitive to poor effort If suspecting response invalidity, vast majority rarely/never directly confronted the examinee, and would not prematurely terminate the assessment 47% would state that responses suggested 'malingering'	100% of experts considered validity testing to be mandatory in forensic examinations and 65% felt SPVTs to be mandatory in clinical assessments Experts considered malingering to be the most likely cause of SPVT failure in forensic settings, but a very infrequent cause of failure in clinical settings Experts considered other factors, such as somatoform/conversion disorder, psychiatric issues or attitude towards testing (oppositional, non-compliant or indifferent behaviour), to be underlying SPVT failure in clinical contexts, although there was no consensus as to common underlying mechanisms	Unstandardized survey limited by self-selection bias and self-report Small sample Generalisability – only North American psychologists working with adult populations Failed to report average number of assessments conducted in the last year
Young, Roper, & Arentsen (2016)	*N* = 172 US psychologists working with the Veterans Affairs healthcare system and likely practising neuropsychology	Overall, 28% of examinees were estimated to fail 1 PVT, 18% fail 2, and 13% fail 3 or more Mean estimated failure rates varied across settings; 23% in clinical	Approximately two-thirds indicated that they used SPVTs always/frequently across clinical and forensic contexts When utilised, 89% employed 2 or more SPVTs There was no consensus regarding the use of 1 or 2	Factors considered to limit SPVT use were time constraints, and influence of supervisors/organisations	Unstandardized survey limited by self-selection bias and self-report Unclear survey wording meant PVT and SVT use

	Email survey, open for one month 44% response rate 43% conducted forensic assessments, & 16% board certified Mean number of assessments yearly = 155	outpatients, 12% in inpatients, and 39% in disability exams Greater rate of failure estimated for mTBI and PTSD populations The less likely respondents were to use SPVTs, the lower their estimated base rates of failure Board certification status not associated with estimated PVT failure rates Base rates in VA system were comparable to other US medico-legal settings	SPVT failures to indicate invalid performance (45% and 47% respectively) Respondents with a greater number of professional organisation memberships tended to employ more PVTs (p<.02); no correlation was found for board certification Stand-alone PVTs were used always/frequently 63% of the time, embedded were used in 73% of cases, and SVTs utilised in 43% of cases The most commonly employed stand-alone PVTs were the TOMM, Rey-15 Item, and WMT. The most commonly employed embedded PVTs were the CVLT-II Forced Choice, RDS, Wisconsin Card Sorting Test (Failure to Maintain Set), and Digit Span Age-Corrected Scaled Score. The MMPI-2 was the most utilised SVT Respondents were most likely to report on results as 'poor/suboptimal effort'. 'Malingering', 'feigning' and 'disability seeking' were the least popular descriptions		was combined Failed to report number of years of practice of respondents
Barker-Collo & Fernando (2015)	N = 73 Registered psychologists in New Zealand Advertised through NZ Psychological Society and NZ College of Clinical Psychology and provided weblink for online survey; open for 4 months 89% of sample self-	Majority of respondents (32%) reported 5-20% of cases presented with suspect effort 24% of sample reported 1-5% of their cases present with suspect effort 24% indicated suspect effort in 20-50% of examinees	Majority of respondents (56%) assessed response validity in <50% of examinees 75% reported using multiple methods to assess response validity Most utilised methods were clinical judgement (47%) and SVTs (38%), such as the MMPI and PAI When used, most popular stand-alone PVTs were the TOMM (39%), WMT (26%), and the Rey 15-Item (28%)	SPVTs were employed in secondary gain contexts or when clients presented with unusual symptoms/inconsistent history Respondents reported using SPVTs due to endorsement by professional boards, awareness of support in the literature, to safeguard the validity of conclusions drawn, and to improve client care	High level of selection bias likely due to the sampling method and likelihood that respondents were highly motivated to take part Relatively small sample Not possible to

	identified as clinical or educational psychologists and others were 'generalists' Mean years of practice = 14 years Most respondents were clinicians working for the Accident Compensation Corporation (ACC) or privately; minority also conducted medico-legal assessments			Reasons given for not assessing response validity included when the population/context was deemed inappropriate, as well as practical challenges such as time restraints, limited access to tests and lack of training/experience Respondents were also concerned with over-reliance on test scores, the notion that validity testing does not reveal underlying motivations, and disapproval of using deception with clients	report response rate due to design Unstandardized survey Did not utilise inferential statistics Failed to report average number of assessments conducted in the last year
Martin, Schroeder, & Odland (2015)	*N* = 316 Licensed North American neuropsychologists who primarily assess adults Online survey based on previous surveys, sent via professional neuropsychology email listservs and open for approximately 3 weeks Majority (33%) worked in private practice, and 73% of sample did at least some forensic assessments Mean years of practice = 12	Not reported	92% often or always use embedded and stand-alone measures to assess response validity 35% indicated that they most commonly use 2 or more PVT 'failures' to indicate cognitive invalidity 13% reported they relied on clinical judgement, but 89% agreed or strongly agreed that validity testing is more accurate than clinical judgement Respondents used mean of 1.6 stand-alone and 3.2 embedded measures in clinical assessments and 2.4 stand-alone and 3.9 embedded SPVTs in forensic evaluations Most commonly employed stand-alone SPVTs were the TOMM and WMT. RDS and the CVLT-2 were the most endorsed embedded SPVTs. Most utilised SVTs were the MMPI and PAI 97% often/always encouraged examinees to try their best, and 38% often/always explicitly warned	98% believed SPVTs to be mandatory in forensic settings and 55% mandatory in clinical settings SPVTs considered by majority to be more accurate than clinical judgement; forensic neuropsychologists significantly more likely to strongly hold this belief than clinical workers ($p < .001$) Most likely cause of test invalidity in clinical cases was believed to be psychiatric issues (not including somatoform or conversion disorder), and most likely cause in forensic settings was reported to be malingering Least common underlying	Unstandardized survey limited by self-selection bias and self-report North America only – limited generalisability Not possible to report response rate due to design Failed to report average number of assessments conducted in the last year Experience was variable; majority of sample had

			examinees that SPVTs would be utilised When suspecting invalid performance, majority would administer additional SPVTs and would not directly confront the examinee 91% often/always reported that 'test results are inconsistent with the severity of injury'. Majority (74%) used the term PVT, and just 11% preferred 'malingering'	causes in both clinical and forensic settings were considered to be genuine cognitive impairment, and stereotype/diagnosis threat Respondents who read more SPVT literature considered validity testing to be significantly more valuable than those who read less ($p<.001$), and were significantly more likely to strongly agree that PVTs are more accurate than clinical judgement ($p<.001$)	practiced in neuropsychology for 10 years or less
Allcott et al. (2014)	*N* = 73 UK multi-disciplinary experts at consultant level in medico-legal settings (psychologists, psychiatrists, orthopaedic specialists, neurologists, & occupational therapists) Emailed to members of the Directory of Expert Witnesses as well as other known experts; open for 6 months 25% response rate	70% of respondents indicated that three quarters of examinees were 'genuine', and 25% considered half of their cases to be 'disingenuous'	Majority (49%) relied upon discrepancies between self-reports and medical history to assess response validity 44% of respondents reported they did not routinely administer SPVTs 40% of respondents did not standardly express opinion on the validity of performance and 11% reported they had never considered performance validity	25% believed 'malingering' to be a medical diagnosis 55% of their UK sample were not able to list any peer-reviewed literature on the subject, and half of respondents who indicated they routinely use PVTs could not name any peer-reviewed research Noted scepticism in relation to validity testing in qualitative comments, e.g. "history and examination are the best indicators" 46% felt it was desirable to know the amount of compensation being claimed when forming an opinion.	Unstandardized survey limited by self-selection bias and self-report Relatively small sample Did not utilise inferential statistics Failed to report average number of assessments conducted in the last year Failed to report number of years of practice of respondents

Dandachi-Fitzgerald, Ponds, & Merten (2013)	*N* = 515 Surveyed 6 European countries (Germany, Italy, Denmark, Finland, Norway, Netherlands) Email survey sent to chairs of each of the European Societies of Neuropsychology to forward to respective members; six of 12 societies agreed to participate Survey open for 18 months Range of 6-25% response rates 96% psychologists, 3% physicians 95% conducted clinical assessments, 55% undertook forensic work Mean years of practice = 10 Median assessments conducted in previous year = 70	Base rates of insufficient effort estimated to be 10% in clinical assessments & 15% in forensic assessments Malingering was thought to occur in 4% of their clinical cases and 10% of forensic cases Discrepancy was found between general estimated base rates of malingering and respondents' estimates of base rates in their own practice (general estimates were rated as 10% in clinical and 20% in forensic assessments)	69% of respondents reported they often/always base their judgements on qualitative methods such as discrepancies between self-reports, records, and condition severity Only 11% indicated systematically using SPVTs in clinical assessments, and just 44% in forensic assessments across the whole sample Respondents in Norway were most likely to use SPVTs (86% in the majority of the forensic assessments and 54% in the majority of clinical assessments) Respondents in Italy reported the lowest rate of SPVT use (13% and 10% in the majority of forensic and clinical assessments respectively) When used, the most popular stand-alone PVTs were the Amsterdam Short-Term Memory Test, Rey 15-Item, and the TOMM. The most commonly utilised embedded PVTs was the Rey Auditory Verbal Learning Test (however, >50% of respondents indicated never utilising embedded tests) Respondents were divided on whether they warn examinees about SPVT use; however, most encouraged examinees to give their best effort If suspecting poor effort, majority would continue the assessment and encourage the examinee to give good effort, but were divided on administering additional SPVTs or directly confronting the examinee 66% indicated they would often/always state that test results are 'inconsistent with severity of injury',	Majority believed could rely on clinical judgement to assess response validity Reasons provided for not utilising SPVTs included presence of severe cognitive impairment (47%), poor effort being obvious in the pattern of other test scores (25%), and poor effort being rare in clinical settings and therefore validity testing is unnecessary (23%) Reasons provided for using SPVTs were related to having read the literature (63%), SPVTs being necessary to validate other findings (59%) and in line with recommendations from professional bodies (59%). 31% of respondents endorsed using SPVTs to "cover my back"	Unstandardized survey limited by self-selection bias and self-report Representativeness of Western European neuropsychologists limited due to only 6 of 12 countries responding Some low response rates e.g. 6% in Denmark Broad sampling method increased the risk of non-response bias, as not all of the participants approached would have conducted neuropsychological assessments. No method of discerning to what extent non-response bias affected findings

			& 60% would indicate that 'no firm conclusions can be drawn'		
McCarter, Walton, Brooks, & Powell (2009)	*N* = 130 UK survey emailed to members of the BPS Division of Neuropsychology 22% response rate 70% of sample conducted both clinical and forensic assessments, 29% solely clinical and 1% solely forensic	5-7% of clinical cases were estimated to require validity assessment, and 60% of medico-legal assessments were considered to require SPVT on a mandatory basis	16% of respondents in clinical settings used SPVTs the majority of the time, compared to 73% in medico-legal settings Clinicians working in medico-legal settings were more likely to always comment on test taking behaviour than those in clinical settings (95% and 76% respectively) However, SPVTs were employed standardly by 59% in medico-legal assessments and only 11% of clinical assessments; majority in clinical settings utilised SPVTs in fewer than 5% of cases The TOMM was most popular in both clinical (32%) and medico-legal work (58%).The WMT was utilised by 34% of medico-legal workers but none of the clinical workers. The Rey 15-Item was also used by a sizeable minority (15%) Most popular embedded measures were comparisons of recognition memory and free recall scores in clinical assessments (8%), and Raven's Progressive Matrices in medico-legal work (8%). However, 11% reported using their own idiosyncratic methods SVTs were employed by a minority of medico-legal workers but very rarely in clinical settings	60% of medicolegal workers considered SPVT use to be mandatory in legal cases, compared to 5% in clinical settings. 16% of clinical workers considered SPVTs to be 'unnecessary', and majority (55%) indicated PVTs were 'optional' Justifications given for validity testing included endorsement by scientific/professional literature, and need to validate the assessment findings overall Most commonly endorsed reason for not including PVTs was related to the belief that invalid responding is obvious from observations or other test results (29%) Respondents also reported time constraints and perception of low base rates of malingering in clinical cases as justifications for the exclusion of SPVTs in assessments	Unstandardized survey limited by self-selection bias and self-report Failed to report years of practice or number of assessments undertaken yearly Failed to explore proportion of clinical and forensic work by sample

Sharland & Gfeller (2007)	*N* = 188 Paper surveys mailed to a random sample of approximately one third of NAN professional members (clinical neuropsychologists) 26% response rate Mean years of practice = 17 30% board certified in neuropsychology Unknown practice settings or proportion of clinical/forensic assessments conducted	In their own practice, median base rate of probable insufficient effort was 10%, and definite insufficient effort was rated at 5% In general practice, median base rate of deliberate exaggeration in medico-legal assessments was 20%, and in cases with no obvious secondary gain estimates were 5% Ranges of base rates were 'considerable' (0-90%)	56% of respondents reported they often/always included an SPVT Most utilised method of assessing response validity was to compare severity of cognitive impairment with severity of the condition (88% often/always) 63% often/always relied upon stand-alone PVTs, 46% often/always used embedded measures, and 55% often/always utilised SVTs The TOMM and the Rey-15 were the most utilised stand-alone PVTs. The CVLT and RDS were the most frequently used embedded PVTs. The MMPI-2 was the most utilised SVT 89% often/always provided encouragement to examinees to try their best 22% often/always warned that tests are sensitive to effort, but 52% never/rarely provide warning Respondents most commonly reported that 'test results were inconsistent with the severity of the injury', and least likely to report that 'test results suggest or indicate malingering'	Participants had the greatest confidence in the TOMM; classification accuracy was rated at 7.5/10	Unstandardized survey limited by self-selection bias and self-report Failed to report years of practice or number of assessments conducted in past year Lower than desired response rate Failed to report practice settings or proportion of clinical/forensic assessments conducted
Boccaccini, Boothby, & Overduin (2006)	*N* = 116 Pain specialists and clinical-forensic psychologists working in personal injury medico-legal settings Mailed questionnaire (including a vignette describing an attorney-referred case) to members of	Not reported	71% indicated that they assess the validity of pain complaints, and this was comparable across pain specialists (68%), forensic specialists (74%), and forensic-pain specialists (76%) Most popular method was to review collateral information, such as reports of other professionals and medical records, observations of pain-related behaviours and discrepancies between pain complaints and test data	Qualitative comments by respondents without forensic training appeared to dispute the necessity of validating pain symptoms e.g. 'diagnostic tests are not as useful for pain conditions as other medical problems'	Relatively low response rate Unstandardized survey limited by self-selection bias and self-report Failed to report respondents' number of years of

	the American Pain Society, and relevant divisions of the American Psychological Association Asked respondents whether they would attempt to assess response validity, and methods they would use 18% response rate 34% had specialised training in forensic and pain assessment		29% reported they would use SVTs. The most commonly endorsed SVT was the MMPI-2 (56% of clinicians trained in forensic and pain assessment). The TOMM was endorsed by a minority of respondents and these tended to be forensic specialists No respondents endorsed any measure specifically intended to assess malingered pain. Qualitative comments suggested that respondents relied upon general pain and coping measures in their assessment of symptom validity, however, none of the measures cited incorporated validity scales		practice Failed to report average number of assessments conducted in the last year 80% coder agreement limits reliability
Sullivan, Lange, & Dawes (2006)	*N* = 17 Members of the Australian Psychological Society, College of Clinical Neuropsychology and delegates from two Australian neuropsychology conferences were emailed an online survey Mean years of practice = 13 years 64% worked in private practice settings 60% of respondents' work was clinical assessment and the remainder forensic work	Base rates of probable symptom exaggeration reported to be 17% of forensic cases, 13% of disability/personal injury assessments, and 4% of clinical cases (medical/psychiatric) The highest base rates were found in mild head injury cases and the lowest were associated with cases involving vascular dementia	84% indicated they routinely screen for response validity in litigation cases, and 38% routinely screen in clinical cases Respondents reported considering an average of 6.6/9 possible indicators when assessing response validity Methods most frequently endorsed by respondents were inconsistencies between severity of cognitive impairment and condition (68%), inconsistent pattern of performance and condition (66%), discrepancies between observations, self-reports and records (64%) and scores below cut-offs on SPVTs (59%) The most popular stand-alone PVTs were the Rey 15-Item and the TOMM. The RAVLT recognition score was the most frequently endorsed embedded measure	Not reported	Very small sample size Not possible to report response rate due to design Unstandardized survey limited by self-selection bias and self-report

Slick, Tan, Strauss, & Hultsch (2004)	*N* = 24 North American neuropsychologists who were identified as experts in the area of civil litigation through their publication history Survey completed via email or over the telephone across a 3 month period 61% response rate 55% board certified in neuropsychology Mean years of practice = 15 Majority (71%) had undertaken >20 assessments in the previous year Conducted both clinical and medico-legal evaluations	50% of respondents considered base rates of possible malingering to be at least 10%, and a third considered rates to be at least 20% of cases Two-thirds considered there to be the presence of definite malingering in at least 10% of cases	The majority (79%) used at least one PVT per assessment, and all respondents who employed a fixed battery approach reported they included a PVT routinely Multiple methods of assessing response validity were utilised; an average of 7.5/9 possible methods were considered by respondents when evaluating performance validity Most frequently used stand-alone PVTs were the TOMM and Rey15-Item 89% of respondents reported they encouraged clients to try their best Respondents were divided on whether they gave examinees warning that tests are sensitive to invalid responding When suspecting invalid responding, the majority reported they would administer additional SPVTs (73%), and a minority would discontinue the assessment (16%) Over 90% indicated they often/always reported that test results were invalid, or inconsistent with injury severity. The majority (54%) reported never/rarely using the term malingering	The average rating of confidence in respondents' own abilities to detect response invalidity was 7.75/10 Ratings were weakly correlated with reported base rates of definite malingering, but were strongly correlated with estimates of possible malingering (r=-.13, p=.44 and r=-.79, p<.01 respectively); a lower confidence in ability to detect malingering was reported by those who estimated higher prevalence of malingering	Unstandardized survey limited by self-selection bias and self-report Unclear whether base rates were estimates or accurate ratings Small sample size but good response rate
Mittenberg, Patton, Canyock, & Condit (2002)	*N* = 144 Paper surveys were mailed to members of the American board of neuropsychologists who were listed as actively practising as neuropsychologists in the	Base rates of probable malingering and symptom exaggeration were estimated to be 30% in disability evaluations, 29% in personal injury cases, 19% in criminal cases and 8% of medical cases; base	The most common method for assessing response validity was comparing severity of injury/condition with severity of cognitive impairment (65%), or pattern of cognitive impairment (64%) 57% relied upon scores below cut-offs on forced choice PVTs, and 38% utilised scores below cut-	Not reported	Unstandardized survey limited by self-selection bias and self-report Proportion of clinical, medico-legal and forensic

	USA/Canada 37% response rate Respondents engaged in both clinical and medico-legal/forensic work Mean years of practice = 18 Mean number of assessments undertaken yearly = 252	rates did not vary greatly across practice settings or geographic regions, or the number of assessments conducted annually	offs on SVTs Respondents endorsed an average of 7.5/9 possible validity indicators in forming a clinical opinion Presentations most likely to be associated with invalid performance in litigation settings were mild head injury (39%), fibromyalgia (35%), chronic pain (31%) and neurotoxic injuries (27%)	assessments not reported

Note. PVT = Performance validity Test; SVT = Symptom validity Test; SPVT = Symptom and performance validity test; NAN = National Academy of Neuropsychology; AACN = American Academy of Clinical *Neuropsychology*; INS = International *Neuropsychological* Society; TOMM = Test of Memory Malingering; MSVT = Medical Symptom Validity Test; WMT = Word Memory Test; RDS = Reliable Digit Span; CVLT-II = California Verbal Learning Test – Version 2; BRIEF = Behavior Rating Inventory of Executive Function; BASC-2 = Behavior Assessment System for Children - Version 2; mTBI = Mild traumatic brain injury; VA = Veterans Affairs; MMPI-II = Minnesota Multiphasic Personality Inventory – Version 2; ACC = Accident Compensation Corporation; PAI = Personality Assessment Inventory

Appendix C

Center for Evidence Based Management Quality Appraisal Tool (Section A)

CEBMa center for Evidence-Based Management

Critical Appraisal of a Survey

Appraisal questions	Yes	Can't tell	No
1. *Did the study address a clearly focused question / issue?*			
2. *Is the research method (study design) appropriate for answering the research question?*			
3. *Is the method of selection of the subjects (employees, teams, divisions, organizations) clearly described?*			
4. *Could the way the sample was obtained introduce (selection)bias?*			
5. *Was the sample of subjects representative with regard to the population to which the findings will be referred?*			
6. *Was the sample size based on pre-study considerations of statistical power?*			
7. *Was a satisfactory response rate achieved?*			
8. *Are the measurements (questionnaires) likely to be valid and reliable?*			
9. *Was the statistical significance assessed?*			
10. *Are confidence intervals given for the main results?*			
11. *Could there be confounding factors that haven't been accounted for?*			
12. *Can the results be applied to your organization?*			

Adapted from Crombie, *The Pocket Guide to Critical Appraisal*; the critical appraisal approach used by the Oxford Centre for Evidence Medicine, checklists of the Dutch Cochrane Centre, BMJ editor's checklists and the checklists of the EPPI Centre.

Appendix D

Broader Research and Clinical Implications of the Review (Section A)

Broader Literature	Research and Clinical Implications
Service users' experiences of testing	One of the few studies exploring service-users' experiences of neuropsychological assessments found half of the sample experienced significant fatigue and feelings of frustration, and a quarter indicated feeling anxious (Bennett-Levy, Klein-Boonschate, Batchelor, McCarter, & Walton, 1994). Further research on this topic is needed to contextualise test performance, and to broaden clinicians' beliefs about the meaning of SPVT failure.
Clinicians' experiences of talking with service users about response validity, for example, with regard to informed consent	The review findings are linked to wider professional issues around informed consent, and how to negotiate this clinical dilemma when using SPVTs. Although professional recommendations provide limited guidance on providing warning to service-users that SPVTs will be used, this is by no means definitive and the issue of informed consent is not elaborated on by either the US or UK recommendations (McMillan et al., 2009; Bush et al., 2005; Heilbronner et al., 2009).
The content of the professional recommendations themselves	Francke, Smit, de Veer and Mistiaen (2008) carried out a systematic meta-review of factors affecting healthcare guidance adherence and found higher rates of implementation when guidelines were simply explained and easy to understand. The authors also found that targeted implementation interventions were necessary following the release of new guidance. Ferlie and Shortell (2001) argue that practice change interventions need to operate across multiple levels; individual clinicians, teams, organisations, and wider systems such as professional bodies. It may be that SPVT guidance needs to be more clearly communicated and disseminated in more accessible ways.
Understanding differences in base rate in different settings	The studies reviewed were suggestive of variable base rates of SPVT failure across different clinical settings and populations. This has important clinical implications for services in developing an understanding of what SPVT failure means in practice. This may involve consideration of the psychometric properties of SPVTs, confounding variables such as mood or cultural factors, or the potential for malingering.

Appendix E

Approval Letter from Research Ethics Committee (Section B)

This has been removed from the electronic copy

APPENDICES OF SUPPORTING MATERIAL

This has been removed from the electronic copy

APPENDICES OF SUPPORTING MATERIAL

This has been removed from the electronic copy

APPENDICES OF SUPPORTING MATERIAL

This has been removed from the electronic copy

APPENDICES OF SUPPORTING MATERIAL

This has been removed from the electronic copy

APPENDICES OF SUPPORTING MATERIAL

This has been removed from the electronic copy

APPENDICES OF SUPPORTING MATERIAL

This has been removed from the electronic copy

Appendix F

Personality Assessment Inventory (PAI) Scale and Subscale Descriptions

The PAI contains four validity scales. The Positive Impression Management (PIM), and Negative Impression Management (NIM) scales are concerned with the extent to which respondents present themselves in a favourable or unfavourable manner, and includes exaggerated, bizarre and highly unlikely symptoms. The Inconsistency (INC) scale identifies individuals who are not responding consistently to similar items. Finally, the Infrequency (INF) scale reflects the level of random or careless responding.

The clinical scales of the PAI were developed based upon a construct validation framework in relation to psychiatric diagnostic categories relied upon at the time (Morey, 1991), and include Somatic Concerns (SOM), Anxiety (ANX), Anxiety Related Disorders (ARD), Depression (DEP), Mania (MAN), Paranoia (PAR), Schizophrenia (SCZ), Borderline features (BOR), Antisocial features (ANT), Alcohol Problems (ALC), and Drug Problems (DRG). The treatment consideration scales relate to factors that may influence engagement in treatment, and comprise Aggression (AGG), Suicidal Ideation (SUI), Nonsupport (NON), Stress (STR), and Treatment Rejection (RXR). Finally, the interpersonal scales explore Dominance (DOM) and Warmth (WAR).

Morey (1991, 2007) suggests that *t* scores of 70 or above on a scale are suggestive of significant symptoms.

APPENDICES OF SUPPORTING MATERIAL

This has been removed from the electronic copy

APPENDICES OF SUPPORTING MATERIAL

This has been removed from the electronic copy

APPENDICES OF SUPPORTING MATERIAL

This has been removed from the electronic copy

APPENDICES OF SUPPORTING MATERIAL

This has been removed from the electronic copy

APPENDICES OF SUPPORTING MATERIAL

This has been removed from the electronic copy

APPENDICES OF SUPPORTING MATERIAL

This has been removed from the electronic copy

APPENDICES OF SUPPORTING MATERIAL

This has been removed from the electronic copy

APPENDICES OF SUPPORTING MATERIAL

This has been removed from the electronic copy

APPENDICES OF SUPPORTING MATERIAL

This has been removed from the electronic copy

APPENDICES OF SUPPORTING MATERIAL

This has been removed from the electronic copy

Appendix G

End of Study Form to the Research Ethics Committee

DECLARATION OF THE END OF A STUDY

(For all studies except clinical trials of investigational medicinal products)

To be completed in typescript by the Chief Investigator and submitted to the Research Ethics Committee (REC) that gave a favourable opinion of the research within 90 days of the conclusion of the study or within 15 days of early termination.
For questions with Yes/No options please indicate answer in bold type.

1. Details of Chief Investigator

Name:	Jessica Hooker
Address:	Salomons Centre for Applied Psychology, Canterbury Christ Church University, 1 Meadow Rd, Tunbridge Wells TN1 2YG
Telephone:	0333 011 7101
Email:	
Fax:	NA

2. Details of study

Full title of study:	Effort test failure and psychological functioning in a UK NHS acquired brain injury population
Research sponsor:	Salomons Centre for Applied Psychology, Canterbury Christ Church University
Name of REC:	[removed to protect anonymity]
REC reference number:	16/LO/2092

3. Study duration

Date study commenced:	25 February 2017
Date study ended:	09 March 2018
Did this study terminate prematurely?	*Yes / **No*** *If yes, please complete sections 4, 5, 6, & 7.* *If no, please go direct to section 8.*

4. Recruitment

Number of participants recruited	
Proposed number of participants to be recruited at the start of the study	
If different, please state the reason or this	

5. Circumstances of early termination

What is the justification for this early termination?	

6. Temporary halt

Is this a temporary halt to the study?	*Yes / No*
If yes, what is the justification for temporarily halting the study? When do you expect the study to re-start?	*e.g. Safety, difficulties recruiting participants, trial has not commenced, other reasons.*

7. Potential implications for research participants

Are there any potential implications for research participants as a result of terminating/halting the study prematurely? Please describe the steps taken to address them.	

8. Final report on the research

Is a summary of the final report on the research enclosed with this form?	***Yes*** */ No* *If no, please forward within 12 months of the end of the study.*

9. Declaration

Signature of Chief Investigator:	
Print name:	Jessica Hooker
Date of submission:	11 April 2018

Appendix H

End of Study Report to the Research Ethics Committee and R&D Department

Study title: Effort test failure and psychological functioning in a UK NHS acquired brain injury population
REC reference: 16/LO/2092
IRAS project ID: 216551

Aims of the Study

Neuropsychological assessments are not valid if the examinee does not try hard (exerts maximum effort) on the tests. Little research exists looking at the issue of effort in NHS populations. Performance validity tests (PVTs) and symptom validity tests (SVTs) have been recommended by the British Psychological Society to assist clinicians in validating assessment data. This study aimed to explore the prevalence (base rate) of PVT failure in the outpatient neuropsychology service, _____. A secondary aim was to investigate the relationship between PVT and SVT performance. Lastly, group differences in those passing and failing PVTs were explored in terms of demographics, and psychological functioning as measured by the Personality Assessment Inventory (PAI). Participants completed one stand-alone PVT (the Test of Memory Malingering [TOMM]), one embedded PVT (Digit Span age-corrected scaled score [DS-SS]), and one SVT (PAI validity scales]).

Findings

Anonymised archival neuropsychological test data (N = 127) spanning 2009 to 2014 were quantitatively analysed. The base rate of failure on any one PVT was 18%. The rate of TOMM failure was 12% and 4% additionally failed an embedded PVT. A significant relationship was found between PVT and SVT performance; participants who failed PVTs reported higher Negative Impression Management scores on the PAI than those who passed (p < .000, r = .34; medium effect size). Significant elevations were also found on the Schizophrenia (p < *.000*, r = -.32; medium effect size), Anxiety-Related Disorders (p = .002, r = -.27; small to medium effect size), and Paranoia (p = .003, r = -.26; small to medium effect size) PAI scales in those who failed one or more PVT compared to those who passed. Additionally, the PVT fail group attained significantly lower Full Scale IQ scores compared to the pass group (p < .000, r = -.37; medium effect size), but pre-morbid IQ scores were comparable across the groups. No other group differences on demographics were found.

Conclusions and Implications

Findings suggest that PVT failure occurs in a sizable minority of NHS acquired brain injury outpatients, which is unlikely to be simply explained by malingering for financial gain. Elevations in reported psychopathological symptoms may be related to emotional and cognitive sequalae resulting from the ABI itself. For example, the Schizophrenia scale could be considered to tap into cognitive or social difficulties arising from brain trauma, for

example ‘My thinking has become confused’, and ‘I just don’t seem to relate to people very well’. Careful interpretation of neuropsychological test data is recommended. It is hoped the current findings will inform UK psychologists in their clinical practice, and contribute to the provision of comprehensive and valid assessments of those who use these services. Research on this topic, as well as clinical practice, would benefit from continuing to explore the non-neurological factors influencing performance, rather than focussing on malingering in isolation, in order to reduce false positive diagnoses and associated invalid recommendations. This study formed the major part of a doctoral thesis for a qualification in Clinical Psychology (DClinPsy) and will be examined by the Salomons Centre for Applied Psychology, Canterbury Christ Church University in April/May 2018. Presentation of findings to staff within the _____ neuropsychology department has been provisionally arranged for 04 September 2018. Upon finalisation the project will be submitted to a peer reviewed journal.

Please do not hesitate to contact me should you require any further information or have any questions.

Jessica Hooker
Trainee clinical psychologist
Salomon’s Centre for Applied Psychology
Canterbury Christ Church University
1 Meadow Road
Tunbridge Wells
TN1 2YG
Tel: 01227 92 7073
Email:

Appendix I

Description of Joint Work

Section B was conducted in part collaboration with a University College London DClinPsy student, Anna Isherwood, whose thesis was completed in early 2018. The current study focused on performance validity testing in terms of differences in symptom validity test performance and self-reported affective and personality variables. My colleague's thesis aimed to explore the presence of a general downgrading of abilities across multiple domains of cognitive functioning in PVT fail groups. Only a subset of the data was shared between the projects (57%) and my colleague additionally accessed separate data from a different research site. Completion of the archival database was done jointly. Ethical applications, analysis and write-up has been conducted separately.

Appendix J

Instructions for Submission to The Clinical Neuropsychologist

Instructions for authors

Thank you for choosing to submit your paper to us. These instructions will ensure we have everything required so your paper can move through peer review, production and publication smoothly. Please take the time to read and follow them as closely as possible, as doing so will ensure your paper matches the journal's requirements. For general guidance on the publication process at Taylor & Francis please visit our Author Services website.

AUTHORSERVICES
Supporting Taylor & Francis authors
SCHOLARONE MANUSCRIPTS™

This journal uses ScholarOne Manuscripts (previously Manuscript Central) to peer review manuscript submissions. Please read the guide for ScholarOne authors before making a submission. Complete guidelines for preparing and submitting your manuscript to this journal are provided below.

Contents

About the Journal

The Clinical Neuropsychologist is an international, peer-reviewed journal publishing high-quality, original research. Please see the journal's Aims & Scope for information about its focus and peer-review policy.

Please note that this journal only publishes manuscripts in English.

The Clinical Neuropsychologist accepts the following types of article: Original Articles, Review Articles, Grand Rounds Articles, Book Reviews.

Authors are strongly encouraged to consult the TCN reporting guidelines checklist when preparing or editing their manuscript. Gross disregard for the reporting guidelines could result in the manuscript being returned without a review.

Peer Review

Taylor & Francis is committed to peer-review integrity and upholding the highest standards of review. Once your paper has been assessed for suitability by the editor, it will then be single blind peer reviewed by independent, anonymous expert referees. Find out more about what to expect during peer review and read our guidance on publishing ethics.

Preparing Your Paper

Structure

Your paper should be compiled in the following order: title page; abstract; keywords; main text introduction, materials and methods, results, discussion; acknowledgments; declaration of interest statement; references; appendices (as appropriate); table(s) with caption(s) (on individual pages); figures; figure captions (as a list).

Word Limits

Please include a word count for your paper. There are no word limits for papers in this journal.

Style Guidelines

Please refer to these quick style guidelines when preparing your paper, rather than any published articles or a sample copy.

Please use American spelling style consistently throughout your manuscript.

Please use single quotation marks, except where ‘a quotation is “within” a quotation’. Please note that long quotations should be indented without quotation marks.

Headers should be in sentence case. p - lower case and italicised. P-values should not have a zero before the decimal point. n - lower case and italicised SD - upper case and italicised ns - lower case and italicised.

Formatting and Templates

Papers may be submitted in Word or LaTeX formats. Figures should be saved separately from the text. To assist you in preparing your paper, we provide formatting template(s).

Word templates are available for this journal. Please save the template to your hard drive, ready for use.

A LaTeX template is available for this journal. Please save the LaTeX template to your hard drive and open it, ready for use, by clicking on the icon in Windows Explorer.

If you are not able to use the template via the links (or if you have any other template queries) please contact authortemplate@tandf.co.uk.

References

Please use this reference guide when preparing your paper.

An EndNote output style is also available to assist you.

Checklist: What to Include

1. **Author details.** Please include all authors' full names, affiliations, postal addresses, telephone numbers and email addresses on the cover page. Where available, please also include ORCiDs and social media handles (Facebook, Twitter or LinkedIn). One author will need to be identified as the corresponding author, with their email address normally displayed in the article PDF (depending on the journal) and the online article. Authors' affiliations are the affiliations where the research was conducted. If any of the named co-authors moves affiliation during the peer-review process, the new affiliation can be given as a footnote. Please note that no changes to affiliation can be made after your paper is accepted. Read more on authorship.
2. Should contain a structured abstract of 250 words. A structured abstract should cover (in the following order): Objective: A brief statement of the purpose of the study. Method: A summary of the participants as well as descriptions of the study design, procedures, and specific key measures, to the extent that space allows. Results: A summary of the key findings. Conclusions: Clinical and theoretical implications of the findings. NOTE: If your manuscript is a critical review or a commentary, you can omit the Results portion of the abstract. However, retain that portion for systematic reviews and meta-analyses. Read tips on writing your abstract.
3. **Graphical abstract** (optional). This is an image to give readers a clear idea of the content of your article. It should be a maximum width of 525 pixels. If your image is narrower than 525 pixels, please place it on a white background 525 pixels wide to ensure the dimensions are maintained. Save the graphical abstract as a .jpg, .png, or .gif. Please do not embed it in the manuscript file but save it as a separate file, labelled GraphicalAbstract1.
4. You can opt to include a **video abstract** with your article. Find out how these can help your work reach a wider audience, and what to think about when filming.
5. Read making your article more discoverable, including information on choosing a title and search engine optimization.
6. **Funding details.** Please supply all details required by your funding and grant-awarding bodies as follows:
   For single agency grants

This work was supported by the [Funding Agency] under Grant [number xxxx].
For multiple agency grants
This work was supported by the [Funding Agency #1] under Grant [number xxxx]; [Funding Agency #2] under Grant [number xxxx]; and [Funding Agency #3] under Grant [number xxxx].

7. **Disclosure statement.** This is to acknowledge any financial interest or benefit that has arisen from the direct applications of your research. Further guidance on what is a conflict of interest and how to disclose it.
8. **Data availability statement.** If there is a data set associated with the paper, please provide information about where the data supporting the results or analyses presented in the paper can be found. Where applicable, this should include the hyperlink, DOI or other persistent identifier associated with the data set(s). Templates are also available to support authors.
9. **Data deposition.** If you choose to share or make the data underlying the study open, please deposit your data in a recognized data repository prior to or at the time of submission. You will be asked to provide the DOI, pre-reserved DOI, or other persistent identifier for the data set.
10. **Geolocation information.** Submitting a geolocation information section, as a separate paragraph before your acknowledgements, means we can index your paper's study area accurately in JournalMap's geographic literature database and make your article more discoverable to others. More information.
11. **Supplemental online material.** Supplemental material can be a video, dataset, fileset, sound file or anything which supports (and is pertinent to) your paper. We publish supplemental material online via Figshare. Find out more about supplemental material and how to submit it with your article.
12. **Figures.** Figures should be high quality (1200 dpi for line art, 600 dpi for grayscale and 300 dpi for colour, at the correct size). Figures should be supplied in one of our preferred file formats: EPS, PS, JPEG, GIF, or Microsoft Word (DOC or DOCX). For information relating to other file types, please consult our Submission of electronic artwork document.
13. **Tables.** Tables should present new information rather than duplicating what is in the text. Readers should be able to interpret the table without reference to the text. Please supply editable files.
14. **Equations.** If you are submitting your manuscript as a Word document, please ensure that equations are editable. More information about mathematical symbols and equations.
15. **Units.** Please use SI units (non-italicized).

Using Third-Party Material in your Paper

You must obtain the necessary permission to reuse third-party material in your article. The use of short extracts of text and some other types of material is usually permitted, on a limited basis, for the purposes of criticism and review without securing formal permission. If you wish to include any material in your paper for which you do not hold copyright, and which is not covered by this informal agreement, you will need to obtain written permission from the copyright owner prior to submission. More information on requesting permission to reproduce work(s) under copyright.

Submitting Your Paper

This journal uses ScholarOne Manuscripts to manage the peer-review process. If you haven't submitted a paper to this journal before, you will need to create an account in ScholarOne. Please read the guidelines above and then submit your paper in the relevant Author Centre, where you will find user guides and a helpdesk.

If you are submitting in LaTeX, please convert the files to PDF beforehand (you will also need to upload your LaTeX source files with the PDF).

Please note that *The Clinical Neuropsychologist* uses Crossref™ to screen papers for unoriginal material. By submitting your paper to *The Clinical Neuropsychologist* you are agreeing to originality checks during the peer-review and production processes.

On acceptance, we recommend that you keep a copy of your Accepted Manuscript. Find out more about sharing your work.

Data Sharing Policy

This journal applies the Taylor & Francis Basic Data Sharing Policy. Authors are encouraged to share or make open the data supporting the results or analyses presented in their paper where this does not violate the protection of human subjects or other valid privacy or security concerns.

Authors are encouraged to deposit the dataset(s) in a recognized data repository that can mint a persistent digital identifier, preferably a digital object identifier (DOI) and recognizes a long-term preservation plan. If you are uncertain about where to deposit your data, please see this information regarding repositories.

Authors are further encouraged to cite any data sets referenced in the article and provide a Data Availability Statement.

At the point of submission, you will be asked if there is a data set associated with the paper. If you reply yes, you will be asked to provide the DOI, pre-registered DOI, hyperlink, or other persistent identifier associated with the data set(s). If you have selected to provide a pre-registered DOI, please be prepared to share the reviewer URL associated with your data deposit, upon request by reviewers.

Where one or multiple data sets are associated with a manuscript, these are not formally peer reviewed as a part of the journal submission process. It is the author's responsibility to ensure the soundness of data. Any errors in the data rest solely with the producers of the data set(s).

Publication Charges

There are no submission fees or page charges for this journal.

Colour figures will be reproduced in colour in your online article free of charge. If it is necessary for the figures to be reproduced in colour in the print version, a charge will apply.

Charges for colour figures in print are £300 per figure ($400 US Dollars; $500 Australian Dollars; €350). For more than 4 colour figures, figures 5 and above will be charged at £50 per figure ($75 US Dollars; $100 Australian Dollars; €65). Depending on your location, these charges may be subject to local taxes.

Copyright Options

Copyright allows you to protect your original material, and stop others from using your work without your permission. Taylor & Francis offers a number of different license and reuse options, including Creative Commons licenses when publishing open access. Read more on publishing agreements.

Complying with Funding Agencies

We will deposit all National Institutes of Health or Wellcome Trust-funded papers into PubMedCentral on behalf of authors, meeting the requirements of their respective open access policies. If this applies to you, please tell our production team when you receive your article proofs, so we can do this for you. Check funders' open access policy mandates here. Find out more about sharing your work.

Open Access

This journal gives authors the option to publish open access via our Open Select publishing program, making it free to access online immediately on publication. Many funders mandate publishing your research open access; you can check open access funder policies and mandates here.

Taylor & Francis Open Select gives you, your institution or funder the option of paying an article publishing charge (APC) to make an article open access. Please contact openaccess@tandf.co.uk if you would like to find out more, or go to our Author Services website.

For more information on license options, embargo periods and APCs for this journal please go here.

My Authored Works

On publication, you will be able to view, download and check your article's metrics (downloads, citations and Altmetric data) via My Authored Works on Taylor & Francis Online. This is where you can access every article you have published with us, as well as your free eprints link, so you can quickly and easily share your work with friends and colleagues.

We are committed to promoting and increasing the visibility of your article. Here are some tips and ideas on how you can work with us to promote your research.

Article Reprints

You will be sent a link to order article reprints via your account in our production system. For enquiries about reprints, please contact the Taylor & Francis Author Services team at reprints@tandf.co.uk. You can also order print copies of the journal issue in which your article appears.

Queries

Should you have any queries, please visit our Author Services website or contact us at authorqueries@tandf.co.uk.

Updated 24-01-2018