Effort testing in dementia assessment: A systematic review

Claire McGuire, Stephanie Crawford and Jonathan J Evans

Abstract

Objective

Interpretation of neuropsychological test data is only valid when appropriate effort has been exerted. Research however suggests that neuropsychologists do not always formally test for effort and that this may especially be the case in the context of dementia assessment. This review systematically examined the literature that has investigated the use of both purpose-built and embedded effort-sensitive indices in dementia, mild cognitive impairment (MCI) and healthy control samples. The aim was to determine which tests of effort are most sensitive to suboptimal effort and least sensitive to the type of cognitive impairment seen in dementia.

Methods

A systematic search of databases was conducted to October 2017. There was no start date.

Results

Twenty five studies were included for review. The studies were divided into two categories according to methodology. One category of studies (n=5) was reviewed using a tailored methodological quality rating checklist whilst the remaining studies (n=20) were reviewed using the Crowe Critical Appraisal Tool (CCAT).

Conclusions

The results of this review suggest that PVTs which take a hierarchical approach to effort testing such as the WMT, MSVT and NV-MSVT are preferable for use with older adults who are under investigation for possible dementia. These tests go above and beyond the traditional pass/fail approach of more traditional tests of effort since they allow the examiner to analyse

the examinee's profile of scores. The methodological limitations and challenges involved in this field of research are discussed.

Introduction

Cognitive testing is used in many clinical settings, alongside information gathered from other sources, to develop a comprehensive understanding of a person's difficulties. Scores on cognitive tests are usually interpreted alongside published normative data which assume that the examinee has put forth good effort. The value and accuracy of an assessment therefore relies on the quality of the data to be interpreted and, as such, it is of great importance that the clinician has evaluated the examinee's level of effort and motivation during the assessment process. Effort testing is considered a crucial component of neuropsychological evaluation according to both the British Psychological Society (BPS; McMillan et al., 2009) and the American Academy of Clinical Neuropsychology (AACN; Heilbronner et al., 2009).

This has led to the creation of both purpose-built tests designed to detect non-credible effort such as the Test of Memory Malingering (TOMM; Tombaugh, 1996), the Word Memory Test (WMT; Green, 2003) and the Rey 15-Item Test (RFIT; Rey, 1964) and those which have been developed from existing neuropsychological test batteries such as the Effort Index (EI; Silverberg, Wertheimer, & Fichtenberg, 2007) and the Effort Scale (ES; Novitski, Steele, Karantzoulis & Randolph, 2012) both derived from the Repeatable Battery for the Assessment of Neuropsychological Status (RBANS; Randolph, Tierney, Mohr & Chase, 1998). Effort tests are also known as symptom validity tests (SVTs) or performance validity tests (PVTs) however Larrabee (2012) distinguishes between these two terms. SVTs are self-report measures which tell the examiner whether a person's symptomatic complaint is reflective of their true experience of symptoms whereas PVTs are cognitive tests which allow

the examiner to know whether the examinee's test performance is reflective of true cognitive ability. For the purposes of this review, the term performance validity test (PVT) will be used alongside test of effort.

There also exist published criteria which can aid clinicians in making a judgement about an examinee's level of effort or motivation. One such set of criteria was published by Slick, Sherman and Iverson (1999) in a landmark paper which encouraged clinicians to apply a 'discrepancy method' to their judgement of poor effort (e.g. attending to inconsistencies between test scores and observed behaviours from the same domain). Slick et al (1999) propose inspecting an examinee's pattern of performance (e.g. comparing scores on easy and difficult items to identify inconsistent patterns of scores) alongside evidence from other sources such as the presence or absence of a substantial, external incentive and significant inconsistencies or discrepancies in the individual's self-report.

Despite these recommendations, however, it appears that effort tests are not always routinely administered as part of cognitive assessment. One study which surveyed 130 UK neuropsychologists about their practices and beliefs regarding tests of effort, found that whilst 59% of respondents working in forensic settings said that they always used a test of effort, only 15% of respondents working in other clinical settings said the same. One third of the respondents in this study said that there was no need to use a dedicated test of effort because non-credible symptoms were evident from the results of other tests and from the client's general presentation during assessment (McCarter, Walton, Brooks & Powell, 2009). Research has found, however, that clinicians' subjective evaluation of test validity is often highly inaccurate (Faust, 1995; Faust, Hart & Guilmette, 1988; Heaton, Smith, Lehman & Vogt, 1978) and that analysing performance on traditional neuropsychological measures alone is an unreliable method of detecting invalid or malingered performance (Van Gorp et al., 1999). It should be noted however that there is a cultural divide regarding the use of effort

testing as research has shown that a majority of clinicians in the United States routinely test for effort as part of neuropsychological assessment (Schroeder, Martin & Odland, 2016).

It may be that some clinicians consider effort tests to only be of relevance when there is suspicion that an individual is deliberately feigning symptoms and indeed the majority of the literature on effort testing focuses on populations in which the deliberate feigning of symptoms is thought to be most prevalent, e.g. medico-legal settings and disability payment assessments. Nevertheless, literature does exist which examines the validity and reliability of effort testing in various clinical groups such as brain injury (Green, Rohling, Lees-Haley & Allen, 2001; Hampson, Kemp, Coughlan, Moulin & Bhakta, 2014), depression (Ashendorf, Constantinou & McCaffrey, 2004), chronic fatigue syndrome (van der Werf, Prins, Jongen, van der Meer & Bleijenberg, 2000) and conversion and somatoform disorders (Boone & Lu, 1999).

There is, however, a lack of information on how people with varying degrees of mild cognitive impairment (MCI) or dementia would be expected to perform on tests of effort as individuals with dementia are often excluded from samples used for effort test validation (Dean, Victor, Boone, Philpott & Hess, 2009).

It may be that tests of effort are not routinely validated in samples of older adults who are under investigation for memory problems because it is thought that they are unlikely to be feigning symptoms. Indeed, a study found that as few as 2% of litigants and those seeking other forms of compensation alleged dementia (Mittenberg, Patton, Canyock & Condit, 2002). It is important however to approach the issue of effort from a wider definition than that of deliberate malingering. There are many reasons unrelated to financial gain that can result in non-credible effort such as: depression, medication side effects, stress, lack of interest, fatigue, lack of comprehension of the utility of the tests or motivation to be in a 'sick role' (Barker, Horner & Bachman, 2010). When conducting cognitive assessment with older

people it is also important to consider medical and physical issues which could impact on an examinee's performance on a test of effort such as visual impairment/disturbance, language difficulties and chronic health conditions which can cause fatigue.

In order for clinicians to be able to adequately assess the reliability of data resulting from cognitive assessment in older adults presenting with memory problems, they must know which effort tests are the most suitable for use with this population.

To date there is no systematic review which examines the literature on the use of effort testing in dementia assessment.

Systematic review objectives

This review evaluates the literature on effort testing in dementia assessment with the following objectives: review which effort tests provide the lowest rate of false-positive error in people with MCI and dementia and to examine the relationship between dementia severity and false-positive rates.

Methods

Search Strategy

The following electronic bibliographic databases were searched: PsycINFO, Cinahl, EMBASE, Medline and Psychology and Behavioural Sciences Collection. The search did not have a start date limit. The end date was October 2017. The following search criteria were used in all databases: ([malingering OR "test* of effort" OR "symptom validity test*" OR "symptom validity" OR "effort test*" OR "validity test*" OR "performance validity" OR "non-credible effort" OR "suspect effort"] AND [dementia OR MCI OR "mild cognitive impairment" OR "geriatric*"]). Titles and abstracts of studies identified were examined to identify those pertaining to effort testing in dementia assessment. Reference lists of all

included papers were also examined to identify any further relevant studies. All the titles and abstracts of identified papers featuring the use of effort testing in dementia assessment were screened against the following criteria:

Inclusion criteria: Studies investigating the performance of dementia and/or MCI samples on tests of effort.

Exclusion Criteria: Studies which solely used a sample of participants asked to simulate MCI or dementia. Single case studies.

Methodological quality

In order to rate the methodological quality of the studies included in this review, the studies were separated into two different categories based on their methodology. The first category pertained to papers in which a reference standard is used to establish if a diagnosis is present or absent in the participants (in this case the diagnosis would be credible or non-credible effort) and then results on the index test (the effort test(s) of interest) are compared between the two groups. A reference standard is the best available method for establishing the presence or absence of a particular diagnosis. To rate the papers included in this first category (n=5), a checklist was developed based on the SIGN Methodology Checklist 5 for Studies of Diagnostic Accuracy (SIGN, 2007) and the Standards for the Reporting of Diagnostic accuracy studies statement (STARD; Bossuyt et al., 2015). The quality rating checklist had a maximum score of 28 points.

The second category covers the majority of the papers included in this review (n=20) in which the researchers have recruited a sample of participants whom they consider not to meet the diagnosis (of non-credible effort). In these studies the researchers have either excluded

participants who may have motivation to feign impairment (involvement in litigation/in receipt of disability payments) or they have assumed that their sample will exert credible effort by virtue of having a diagnosis of dementia/MCI. To rate the methodological quality of these papers, the Crowe Critical Appraisal Tool (CCAT; Crowe & Sheppard, 2011) was used. The CCAT contains 54 reporting items in 8 categories and has a maximum score of 40 points.

All papers were rated by the author. A second rater assessed 12/25 (48%) of the papers to examine the inter-rater reliability of the checklists. Across all the checklist items in the quality rating tools, there was 84% agreement between raters. Where discrepancies occurred, these were resolved through discussion.

Outcome of search process

A total of 25 papers met the inclusion criteria and are included in this review. Figure 1 is a flow diagram illustrating the systematic process of identifying the 25 papers included.

Results

In this review, sensitivity (also called the true positive rate) refers to the ability of the tests to identify non-credible effort when non-credible effort is present. Specificity (also called the true negative rate) refers to the ability of the tests to identify credible effort when credible effort is present. The majority of the studies included in this review involved participants who were deemed to be exerting credible effort due to not being involved in litigation or by virtue of having an established diagnosis of MCI or dementia and therefore having little to no reason to feign impairment (n=20). This means that the methodology involved administering tests of effort to participants who were already deemed to be exerting

credible effort. These studies cannot possibly investigate the sensitivity levels of the effort test(s) in question (there are no true positives present in their samples). They report specificity levels only.

The studies (n=5) which include both participants who are and are not exerting credible effort are able to report both sensitivity and specificity levels with the exception of Schroeder et al. (2012) who used the RBANS Effort Scale as a reference standard but who deemed all of their participants to be exerting credible effort therefore they report specificity levels only. See Tables 1 to 12 for data extraction tables which include demographic information and sensitivity and specificity levels where appropriate. This information is listed per PVT.

The majority of the studies included in this review therefore report specificity levels but not sensitivity levels.

The results of the studies included in this review will be reported by effort test and grouped by purpose-built vs embedded effort tests:

Purpose-built PVTs

1. Test of Memory Malingering (TOMM; Tombaugh, 1996).
2. Rey 15 Item Test (RFIT; Rey, 1964).
3. The Coin in the Hand (CIH; Kapur, 1994).
4. Word Memory Test (WMT; Green, 2003).
5. Medical Symptom Validity Test (MSVT; Green, 2004).
6. Non-Verbal Symptom Validity Test (NV-MSVT; Green, 2008).
7. Amsterdam Short Memory Test (ASTM; Schagen, Schmand, de Sterke &, Lindeboom, 1997).
8. Dot Counting Test (DCT; Rey, 1941).

9. Finger Tapping Test (FTT; Reitan & Wolfson, 1993).

Embedded PVTs

1. Repeatable Battery for the Assessment of Neurological Status (RBANS; Randolph, Tierney, Mohr & Chase, 1998)
    a. Effort Index (EI; Silverberg, Wertheimer & Fichtenberg, 2007).
    b. Effort Scale (ES; Novitski, Steele, Karantzoulis & Randolph, 2012).
    c. Two novel indices (PVI and CRIER) (Paulson, Horner & Bachman, 2015).
2. Reliable Digit Span (RDS; Greiffenstein, Baker & Gola, 1994).

Please note that the study by Dean, Victor, Boone, Philpott and Hess (2009) included in this review investigated a total of 18 stand-alone and embedded effort tests. It was out with the scope of this review to include all of these PVTs however those which have also been investigated by other studies have been included. These are: TOMM, RFIT, DCT and RDS.

Purpose-built PVTs

1. *Test of Memory Malingering (TOMM; Tombaugh, 1996).*

Seven of the 25 papers included in this review investigated the use of the TOMM, one of the most widely used PVTs. The results of the studies can be found in Table 1. The TOMM is a picture-recognition, forced-choice, purpose-designed effort test consisting of two learning trials and an optional retention trial. A cut-off of <45 for Trial 2 suggests that a score lower than this is indicative of poor effort.

Across the seven studies, pass rates for the dementia groups (Trial 2 <45) ranged from a low of 24% (Teichner & Wagner, 2004) to a high of 95% (Rudman, Oyebode, Jones & Bentham, 2011). Only two of the seven studies investigated the utility of the TOMM in MCI samples (Teichner & Wagner, 2004; Walter, Morris, SwierVosnos & Pliskin, 2014). Both studies found similar pass rates for these samples (91.7% and 90.3% respectively).

Drawing comparisons between the results of these studies is compromised to an extent, because they use different criteria for diagnosing dementia (DSM-III, DSM-IV and the ADRDA-NINDS) and they also assess cognitive function using different tools (five report MMSE scores, one uses the RBANS and one the CAMCOG). This is important because one reason for the discrepancy in results across studies might be that the samples include individuals with significantly different levels of cognitive function. The difference in results reported by Tombaugh (1997) and Teichner and Wagner (2004) may be explained by dementia severity. Tombaugh states that 4/37 dementia participants (i.e. 10% of their dementia sample) who scored below 40 on the TOMM, had MMSE scores of 7, 15, 16 and 19. The paper does not, however, give any detail about the MMSE scores of the rest of the sample (presumably the remaining participants all of whom scored > 45 in this sample had MMSE scores of >19). In Teichner and Wagner's (2004) sample however, 9/21 (i.e. 42.9% of

their dementia sample), had MMSE scores lower than 19. It may therefore be that Teichner and Wagner's (2004) sample was more cognitively impaired than that of Tombaugh's (1997). Rudman and colleagues (2011) found a specificity of 95% for the TOMM, however this was in a sample of mildly impaired dementia participants. Specificity dropped to 36% in their moderate to severely impaired sample.

Bortnik et al. (2013), Teichner and Wagner (2004) and Tombaugh (1997) all give data for alternative cut-offs. A cut-off of 40 produced a specificity of 89.2% in Tombaugh's sample however the same cut-off yielded only 48% specificity in Teichner and Wagner's dementia sample (compared to 97.3% of their MCI group). In the Bortnik et al. (2013) good effort dementia sample, the TOMM reached 95% specificity when the cut-off was lowered to 34

There are also some flaws in the reporting of results across studies. Tombaugh (1997) which received a score of 30/40 on the CCAT, states that 'a cutting score of 45 on Trial 2 produced a high level of specificity. It correctly classified 95% of all non-demented patients (91% of all patients) as non-malingering' (p.265). As previously mentioned however, 89.1% of Tombaugh's (1997) dementia sample passed the TOMM but only when a cut-off score of <40 was used. This drops to 72.9% with the recommended cut-off of <45, meaning that approximately one in four of their dementia patients were incorrectly classified as putting forth non-credible effort.

Additionally, with the exception of Bortnik, Horner and Bachmann (2013) sample sizes across studies were small.

Table 1. Data extraction table for the Test of Memory Malingering (TOMM; Tombaugh, 1996) – all results refer to a cut-off score of <45/40 on TOMM Trial 2 unless otherwise stated.

Study	Sample (n)	Mean Age (SD)	Mean Education (SD)	Impairment measure	Mean Score on Impairment Measure (SD)	±TOMM Sensitivity (%)	*TOMM Specificity (%)	Quality Rating score
Walter et al. (2014)	Dementia (mod/sev) (31)	69.48 (7.86)	15.63 (2.92)	RBANS total score	60.7 (7.48)	n/a	79	31/40 (78%)
	MCI (28)	66.02 (8.02)	14.68 (2.09)		80.72 (4.47)	n/a	91	
	Controls (30)	71.43 (8.99)	16.33 (3.19)		96.73 (8.61)	n/a	100	
Bortnik et al. (2013)	Good effort (119)	77 (7)	11.42 (4)	MMSE	20.8 (5)	n/a	**45** 78 **37** 91 **34** 95	20/28 (71%)
	Suspect effort (9)	72 (8.13)	10.44 (2.3)		17.9 (4.5)	**45** 100 **37** 78 **34** 67	n/a	
Rudman et al. (2011)	Dementia (mild) (20)	58.3	Not reported	CAMCOG	88.60 (5.03)	n/a	95	28/40 (70%)
	Dementia (mod/sev) (22)				62.45 (16.21)	n/a	36	
Dean et al. (2009)	Dementia (20)	Not reported	Not reported	MMSE	19.2 (4.4)	n/a	45	30/40

								(75%)
Merten et al. (2007)	AD (20)	73.5 (4.8)	11.7 (3.3)	MMSE	22.2 (2.9)	n/a	70	30/40 (75%)
	Controls (14)	76.6 (6.7)	11.3 (3.7)		28.9 (1.0)	n/a	100	
Teichner & Wagner (2004)	Dementia (21)	75.3 (6.1)	13.6 (3.3)	MMSE WAIS FSIQ	19.9 (2.8) MMSE 80.6 (12) WAIS	n/a	**45** 24 **42** 38 **40** 48	27/40 (68%)
	MCI (36)	70.6 (8.1)	14.2 (3.2)		25.6 (2.5) MMSE 90.8 (14.8) WAIS	n/a	**45** 91.7 **42** 94.4 **40** 97.3	
	Controls (21)	65.6 (8.6)	14.2 (3.6)		28.3 (1.7) MMSE 99.1 (15.3) WAIS	n/a	**45** 100 **42** 100 **40** 100	
Tombaugh (1997)	Dementia (37)	69.48 (7.86)	72.1 (7.6)	MMSE	Not reported	n/a	**45** 72.9 **40** 89.2	30/40 (75%)
	Controls (13)	66.02 (8.02)	45.9 (15)				**45** 100 **40** 100	

**Specificity means that the test correctly identified credible effort (i.e. the proportion of true negatives).*
±Sensitivity means that the test correctly identified non-credible effort (i.e. the proportion of true positives).

RBANS = Repeatable Battery for the Assessment of Neuropsychological Status
MMSE = Mini-Mental State Examination
CAMCOG = Cambridge Cognitive Examination
WAIS FSIQ = Wechsler Adult Intelligence Scale, Full Scale Intelligence Quotient

AD = Alzheimer's Dementia

2. Rey 15 Item Test (RFIT; Rey, 1964).

Alongside the TOMM, the RFIT is one of the most widely used effort tests in clinical practice (Slick, Tan, Strauss & Hultsch, 2004). The task consists of studying a card for 10 seconds which has five rows of three characters. The test consists of a free-recall and an optional recognition trial. A score lower than 9 on the free-recall trial and a score lower than 20 on the combination equation are said to be indicative of non-credible effort (Boone, Salazar, Lu, Warner-Chacon & Razani, 2002). The results of studies investigating the RFIT can be found in Table 2.

Bortnik, Horner and Bachman (2013) reported a high sensitivity to non-credible effort in their 'suspect effort' group (100%) but a very low specificity (28%) in their good effort group. Rudman, Oyebode, Jones and Bentham (2011) found a similar specificity level (27%) for their moderate/severe dementia group. The studies use different tools to assess cognitive impairment (MMSE and CAMCOG) however research suggests these screens are highly correlated (Heinik, Solomesh & Berkman, 2004) therefore it was possible to note that both Rudman and Bortnik's dementia samples were equally impaired (mean CAMCOG 62.45, mean MMSE 20.8). Rudman found far better specificity in their mildly impaired sample (85% specificity, mean MMSE 20.8). Both studies performed similarly on the quality rating checklists (70%).

Table 2. Rey 15 Item Test (RFIT; Rey, 1964) – Bortnik, Horner and Bachman (2013) use a cut-off of <9, Rudman, Oyebode, Jones and Bentham (2011) use a cut-off of <8.

Study	Sample (n)	Mean Age (SD)	Mean Education (SD)	Impairment measure	Mean Score on Impairment Measure (SD)	±RFIT Sensitivity (%)	*RFIT Specificity (%)	Quality Rating score
Bortnik et al. (2013)	Good effort (119)	77 (7)	11.42 (4)	MMSE	20.8 (5)	n/a	28	20/28 (71%)
	Suspect effort (9)	72 (8.13)	10.44 (2.3)		17.9 (4.5)	100	n/a	
Rudman et al. (2011)	Dementia (mild) (20)	58.3	Not reported	CAMCOG	Not reported	n/a	85	28/40 (70%)
	Dementia (mod/sev) (22)					n/a	27	

**Specificity means that the test correctly identified credible effort (i.e. the proportion of true negatives).*
±Sensitivity means that the test correctly identified non-credible effort (i.e. the proportion of true positives).

MMSE = Mini-Mental State Examination CAMCOG = Cambridge Cognitive Examination

3. The Coin in the Hand (CIH; Kapur, 1994).

The Coin in the Hand Test (CIH) was developed to be a 'simple, brief test designed to detect the presence of malingering in patients who are suspected of simulating poor memory performance' (Kapur, 1994, p.385). It is a stand-alone, forced-choice test in which the clinician holds a coin in their right or left hand in front of the examinee who then closes their eyes and counts backwards from 10 before opening their eyes to report which hand the coin is in. A score of 7 or less out of 10 trials is used as the cut-off for this test. See Table 3 for the results of studies investigating the CIH.

Two of the papers included in this review examined the utility of the CIH. Schroeder, Peck, Buddin, Heinrichs and Baade (2012) used a sample of 45 inpatients with a diagnosis of dementia. The Schroeder study performed fairly well on the CCAT (31/40). In order to ensure that their sample would put forth adequate effort on the CIH, Schroeder excluded participants involved in litigation or who were collecting disability payments and they also excluded any participant failing the RBANS Effort Scale. Using a cut-off score of ≤7, resulted in a specificity of 98% in the Schroeder sample. Using the same cut-off, Rudman et al. (2011) found specificity of 100% in their mild dementia sample compared to 77% in the moderate/severe sample. Specificity therefore is high with a CIH cut-off of ≤7 as long as the examinee is not too cognitively impaired.

Table 3. Coin in the Hand (CIH; Kapur, 1994) – all studies used a cut-off of ≤7/10.

Study	**Sample (n)**	**Mean Age (SD)**	**Mean Education (SD)**	**Impairment measure**	**Mean Score on Impairment Measure (SD)**	***CIH Specificity (%)**	**Quality Rating score (CCAT)**
Schroeder et al. (2012)	Dementia (45)	77.98 (7.05)	12.76 (3.02)	MMSE	21.47 (5.71)	98	31/40 (78%)
Rudman et al. (2011)	Dementia (mild) (20)	58.3	Not reported	CAMCOG	88.60 (5.03)	100	28/40 (70%)
	Dementia (mod/sev) (22)				62.45 (16.21)	77	

**Specificity means that the test correctly identified credible effort (i.e. the proportion of true negatives).*

MMSE = Mini-Mental State Examination CAMCOG = Cambridge Cognitive Examination

4. Word Memory Test (WMT; Green, 2003).

The WMT, MSVT and NV-MSVT, all devised by Green, are PVTs which are a departure from the traditional pass/fail effort tests in that they use profile analysis to determine an individual's level of effort. They are based on a hierarchical approach in line with the criteria devised by Slick et al. (1999). On these PVTs, a participant's score is first compared against a cut-off (a pass/fail approach) and then for those participants who fail, their profile of scores over several subtests is analysed. These tests are unique in that they allow the examiner to distinguish between failure due to poor effort and failure due to cognitive impairment. Green states that on the WMT, MSVT and NV-MSVT, people with genuine cognitive impairment produce a specific profile of results, different from the pattern of results produced by those exerting non-credible effort. To differentiate between the two, the difference between the mean scores on the easy and hard subtests is calculated. Individuals with a diagnosis of dementia invariably show easy-hard differences of at least 20 points on these subtests, whereas such significant differences are rarely present in people asked to feign impairment. This pattern of results is known as the 'dementia profile'. These three effort tests are described and the literature evaluated in turn.

The WMT is a word-list learning task that involves learning a list of 20 word pairs which are presented twice, either on a computer screen or spoken aloud by the examiner (as in the original oral version of the test). It contains multiple subtests of which the first two are specifically designed to measure effort (Immediate and Delayed Recognition), the remaining subtests are conventional memory subtests. As stated above, the profile of scores on the WMT subtests (particularly the difference between the effort subtests and the conventional memory subtests) can indicate whether individuals fail the test due to insufficient effort or to

the severity of their cognitive impairment. The results of the studies investigating the WMT can be found in Table 4.

Merten, Bossink & Schmand (2007) found that whilst all their controls passed the WMT effort subtests, almost all of the AD participants failed it (90% failed both the Immediate and Delayed Recognition trials). It is important to note however, that the authors were not able to analyse the profile of scores on the two effort subtests against those of the conventional memory tests because normative data for a Dutch population were not available at the time of the study. This is a significant limitation of the Merten et al. (2007) study since the researchers were only able to look at whether participants passed or failed the effort subtests and were not able to look at the profile of their scores to investigate whether they indicated a 'dementia profile'.

Green, Montijo and Brockhaus (2011) found that 41/65 (63%) of their dementia sample and 13/60 (21.6%) of their MCI sample failed the easy subtests of the WMT. Using profile analysis however they found that every participant with dementia exhibited a 'dementia profile' meaning there were no false positives. Regarding the MCI sample, only 2 of the 11 participants who failed the easy subtests of the WMT, indicated a profile suggestive of poor effort, which represents a false-positive rate of 3.3% for the total sample. The study however scored 28/40 on the CCAT due to inadequate reporting such as missing demographic data. It should perhaps also be noted that the participants were not screened for potential financial incentives, meaning that it cannot be concluded that the 2 participants who generated a poor effort profile were real false-positives.

Martins & Martins (2010) found that 67% of their MCI sample failed the easy subtests of the WMT however 95% produced a dementia profile therefore were not misclassified as exerting non-credible effort.

Table 4. Word Memory Test (WMT; Green, 2003).

Study	Sample (n)	Mean Age (SD)	Mean Education (SD)	Impairment measure	Mean Score on Impairment Measure (SD)	% Failed easy subtests (met Criterion A)	% Produced dementia profile (met Criterion A not Criterion B)	*WMT Specificity (%)	Quality Rating score
Green et al. (2011)	Dementia group 1 (42)	70.9 (8)	Not reported	CDR	1.05 (0.6)	71	100	100	28/40 (70%)
	Dementia group 2 (23)	65.2 (8)			0.83 (0.35)	48	100	100	
	MCI (60)	68.8 (8.9)			0.5 single domain (29) 0.5 multi domain (31)	21.6	96.7	96.7	
	Controls (19)	55.8 (7.5)			0	0	n/a	100	
Martins & Martins (2010)	MCI (21)	71.2 (2)	Not reported	WMS-III (Logical Memory Subtest)	Not reported	67	95.2	95.2	30/40 (75%)
**Merten et al.	AD (20)	73.5 (4.8)	11.7 (3.3)	MMSE	22.2 (2.9)	**50	n/a	**50	30/40 (75%)

(2007)	Controls (14)	76.6 (6.7)	11.3 (3.7)		28.9 (1.0)	**100		**0	

**Specificity means that the test correctly identified credible effort (i.e. the proportion of true negatives).* ***Based on subtests passed/failed – profile analysis not used.*

CDR = Clinical Dementia Rating MMSE = Mini-Mental State Examination AD = Alzheimer's Dementia

5. Medical Symptom Validity Test (MSVT; Green, 2004).

The MSVT is a shorter, modified and easier version of the WMT. Five of the papers in this review examined the utility of the MSVT in people with dementia, MCI, cognitively intact controls and volunteers asked to simulate dementia and the results of these studies can be found in Table 5. Two of the papers were able to report sensitivity levels of 60% (simulators) and 100% (suspect effort MCI sample). Performance on the easy subtests was variable across the studies from only 12.5% of Howe et al's (2007) MCI sample failing the subtests compared to 100% of Singhals' simulator sample. More importantly however is that by using profile analysis, specificity levels for the MSVT across the 4 studies ranged from 80 -100%, meaning that very few individuals were misclassified as exhibiting poor effort. The exception was Rudman et al. (2011) who solely calculated effort based on whether participants had passed or failed the easy subtests. They did not use profile analysis. Although these results are promising, they should be treated as preliminary as the studies involved very small sample sizes.

Table 5. Medical Symptom Validity Test (MSVT; Green, 2004).

Study	Sample (n)	Mean Age (SD)	Mean Education (SD)	Impairment measure	Mean Score on Impairment Measure (SD)	% Failed easy subtests (met Criterion A)	% Produce dementia profile (met Criterion A not Criterion B)	±MSVT Sensitivity (%)	*MSVT Specificity (%)	Quality Rating score
Suesse et al. (2015)	Dementia (15)	Not reported	Not reported	n/a	n/a	19.7	86.7	n/a	86.7	28/40 (70%)
Green et al. (2011)	Dementia group 1 (42)	70.9 (8)	Not reported	CDR	1.05 (0.6)	Not reported	Not reported	Not reported	Not reported	28/40 (70%)
	Dementia group 2 (23)	65.2 (8)			0.83 (0.35)	44	100	n/a	100	
	MCI (60)	68.8 (8.9)			0.5 single domain (29) 0.5 multi domain (31)	Not reported	Not reported	n/a	Not reported	
	Controls (19)	55.8 (7.5)			0	Not reported	Not reported	n/a	Not reported	
**Rudman et al. (2011)	Dementia (mild) (20)	58.3	Not reported	CAMCOG	88.60 (5.03)	35	Not analysed	n/a	*65	28/40 (70%)
	Dementia (mod/sev) (22)				62.45 (16.21)	72		n/a	*28	
Singhal et al.	Dementia (10)	81.7 (4.6)	10 (2.9)	MMSE	15.5 (5.3)	100	100	n/a	100	31/40 (78%)

(2009)	[a]Simulators (10)	36 (10)	17 (2)		n/a	100	40	60	n/a	
Howe et al. (2007)	**Disability:**			WAIS-III FSIQ	**FSIQ**					35/40 (88%)
	Dementia (5)	56.60 (4.95)	11.60 (2.30)		82.20 (5.63)	60	100	n/a	100	
	MCI (3)	44.33 (19.50)	13.33 (2.31)		89.33 (5.63)	66.7	0	100	n/a	
	Controls (1)	58	13		102	0	n/a	n/a	n/a	
	No Disability				**FSIQ**					
	Dementia early (13)	73.54 (9.03)	12 (2.42)		89.27 (14.54)	38.46	80	n/a	80	
	Dementia advanced (18)	76.39 (6.89)	13.83 (2.92)		86.77 (13.45)	83.33	86	n/a	86	
	MCI (16)	69.50 (9.60)	15.75 (2.93)		100.94 (10.34)	12.5	100	n/a	100	
	Controls (5)	57.20 (17.33)	15.40 (1.67)		107.80 (10.26)	0	n/a	n/a	n/a	

**Specificity means that the test correctly identified credible effort (i.e. the proportion of true negatives).*

±Sensitivity means that the test correctly identified non-credible effort (i.e. the proportion of true positives).

***Based on subtests passed/failed – profile analysis not used.*

CDR = Clinical Dementia Rating CAMCOG = Cambridge Cognitive Examination MMSE = Mini-Mental State Examination WAIS FSIQ = Wechsler Adult Intelligence Scale, Full Scale Intelligence Quotient

6. Non-Verbal Medical Symptom Validity Test (NV-MSVT; Green, 2008)

The NV-MSVT is the non-verbal equivalent of the MSVT. Three studies were found to investigate the diagnostic accuracy of this PVT in dementia and healthy controls and the results of these can be found in Table 6.

As noted with the MSVT, Rudman et al. (2011) found low specificity of the NV-MSVT (33% in their overall dementia sample), however they did not use profile analysis therefore their results are not complete or accurate. Henry et al. (2010) found specificity of 100% in their controls and dementia sample and 97.7% in their non-dementia (neurological) sample whilst Singhal, Green, Ashaye, Shankar and Gill (2009) found 100% specificity for their institutionalised dementia patients. Interestingly, Singhal's entire dementia sample failed the effort subtests of the NV-MSVT however they all showed a 'dementia profile' therefore they were not misclassified as malingering. They were also a particularly impaired sample (average MMSE of 15.5).

Table 6. Non-verbal Medical Symptom Validity Test (NV-MSVT; Green, 2008).

Study	Sample (n)	Mean Age (SD)	Mean Education (SD)	Impairment measure	Mean Score on Impairment Measure (SD)	% Failed easy subtests (met Criterion A)	% Produced dementia profile (met Criterion A not Criterion B)	±NV-MSVT Sensitivity (%)	* NV-MSVT Specificity (%)	Quality Rating score
**Rudman et al. (2011)	Dementia (mild) (20)	58.3	Not reported	CAMCOG	88.60 (5.03)	35	Not analysed	n/a	**65	28/40 (70%)
	Dementia (mod/sev) (22)				62.45 (16.21)	72		n/a	**28	
Henry et al. (2010)	Dementia (21) Without dementia (n=44)	59.1 (16.1)	13.1 (3.4)	MMSE	25.7 (3.9)	61 15	61 15	n/a	100 97.7	28/40 (70%)
	Controls (50)	62.8 (7.1)	15.8 (3)		28.7 (1.1)	2	2		100	
Singhal et al. (2009)	Dementia (10)	81.7 (4.6)	10 (2.9)	MMSE	15.5 (5.3)	100	100	n/a	100	31/40 (78%)
	[a] Simulators (10)	36 (10)	17 (2)		n/a	90	40	60	n/a	

**Specificity means that the test correctly identified credible effort (i.e. the proportion of true negatives).*
±Sensitivity means that the test correctly identified non-credible effort (i.e. the proportion of true positives).
***Based on subtests passed/failed – profile analysis not used. [a]Asked to simulate memory impairment.*
CAMCOG = Cambridge Cognitive Examination MMSE = Mini-Mental State Examination

7. Amsterdam Short Memory Test (ASTM; Schagen, Schmand, de Sterke & Lindeboom, 1998).

The ASTM is a forced-choice test in which five examples from a category (e.g. animals, colours etc.) are read aloud by the examinee. A relatively simple mathematical problem is then presented to the examinee after which time another five words from the same category are presented. The examinee's task is to recognise which three words are identical to those in the first list. The test maximum score is 90 and the cut-off is 84/85. Merten et al. (2007) found that all of their controls passed the ASTM however only 10% of their Alzheimer's dementia participants passed the test (i.e. specificity, 10%) (see Table 7). Given that 70% of the same AD sample passed the TOMM, it would appear that the ASTM is unlikely to be an appropriate choice of PVT for individuals with cognitive impairment due to a possible dementia because the false-positive rate is likely to be high.

Table 7. Amsterdam Short-Term Memory Test (ASTM; Schagen, Schmand, de Sterke & Lindeboom, 1997) – results based on cut-off of 84/85.

Study	**Sample (n)**	**Mean Age (SD)**	**Mean Education (SD)**	**Impairment measure**	**Mean Score on Impairment Measure (SD)**	***ASTM Specificity (%)**	**Quality Rating score (CCAT)**
Merten et al. (2007)	AD dementia (20)	73.5 (4.8)	11.7 (3.3)	MMSE	22.2 (2.9)	10	30/40 (75%)
	Controls (14)	76.6 (6.7)	11.3 (3.7)		28.9 (1.0)	100	

**Specificity means that the test correctly identified credible effort (i.e. the proportion of true negatives).*
MMSE = Mini-Mental State Examination

8. Dot Counting Test (DCT; Rey, 1941).

The DCT is a task in which the examinee is presented with twelve cards containing different numbers of dots. The first six cards contain dots arranged in a random order and the following six cards contain grouped dots. Examinees are asked to count the dots as quickly as possible. The three studies investigating the DCT in dementia and control samples used different scoring methods (see Table 8). Dean et al. (2009) and Boone et al. (2002) used a scoring system whereby the mean ungrouped dot counting time, the mean grouped dot counting time and the error score are summed. They found poor specificity for their dementia samples (33.4-75%) but perfect specificity for their healthy controls.

Rudman et al. (2011) used the scoring recommended by Lezak (2004) which is that suspect effort is considered if the time taken to count the grouped dots is equal to or more than the time required to count the ungrouped dots. They found good specificity (90%) for their mild dementia sample but poor specificity (68%) for their moderate/severe sample. They used this scoring method because the combined score uses two measures of reaction time which they proposed to be inappropriate index to use with dementia samples as individuals with dementia typically experience a reduction in their processing skills.

Table 8. Dot Counting Test (DCT; Rey, 1941)
Rudman et al. (2011) results based on "total ungrouped time< total group time".
Dean et al. (2009) and Boone et al. (2002) results based on 'mean ungrouped dots counting time + error grouped dot counting time + errors" ≤17.

Study	**Sample (n)**	**Mean Age (SD)**	**Mean Education (SD)**	**Impairment measure**	**Mean Score on Impairment Measure (SD)**	***DCT Specificity (%)**	**Quality Rating score (CCAT)**
Rudman et al. (2011)	Dementia (mild) (20)	58.3	Not reported	CAMCOG	88.60 (5.03)	90	28/40 (70%)
	Dementia (mod/sev) (22)				62.45 (16.21)	68	
Dean et al. (2009)	Dementia (80)	Not reported	Not reported	MMSE	18.8 (5)	50	30/40 (75%)
Boone et al. (2002)	Dementia (mild) (16)	75.4 (9.3)	12.7 (2.6)	MMSE	>20	75	21/40 (53%)
	Dementia (mod)	78.1 (10.4)	13 (2.4)		10-20	33.4	
	Controls	63.5 (8.8)	15.3 (2.4)		n/a	100	

**Specificity means that the test correctly identified credible effort (i.e. the proportion of true negatives).*
CAMCOG = Cambridge Cognitive Examination MMSE = Mini-Mental State Examination

9. Finger Tapping Test (FTT; Reitan & Wolfson, 1993).

The Finger Tapping Test (FTT) is used in neuropsychology both as part of the Halstead-Reitan Neuropsychological Test Battery (HRNB) and as a stand-alone PVT. See Table 9 for the results of the studies investigating the FTT in dementia and controls. The recommended cut-offs for dominant-hand finger tapping (women - ≤28, men ≤35) reveal poor specificity for the dementia groups across the two papers included in this review (69%-87%). Arnold et al. (2005), however provide various cut-offs for dominant hand, non-dominant hand, summed hands and difference between hands. They found 100% specificity for male dementia participants using a difference score of ≤-2 and 87% specificity for female dementia participants using a difference score of ≤5. Further analysis revealed differences in specificity values by dementia subgroup such that the dominant hand cut-off of ≤28 for women with Alzheimer's disease (AD) was 83% whilst for females with non-AD dementias the specificity reduced to 64%. Similarly, a cut-off of >41 for the dominant hand identified 100% of men with AD dementia whilst only 37% of men with non-AD dementias achieved the same score.

Table 9. Finger Tapping Test (FTT; Reitan & Wolfson, 1993).

Study	**Sample (n)**	**Mean Age (SD)**	**Mean Education (SD)**	**Impairment measure**	**Mean Score on Impairment Measure (SD)**	***FTT Specificity (%)**	**Quality Rating score (CCAT)**
Dean et al. (2009)	Dementia (80)	Not reported	Not reported	MMSE	18.8 (5)	*Women* **Dom ≤28** 69 *Men* ***Dom* ≤35** 69	30/40 (75%)
Arnold et al. (2005)	Dementia (31)	67.48 (10.95)	13.61 (3.87)	Not reported	n/a	*Women* **Dom ≤28** 75 **Non-dom ≤25** 87 **Sum ≤58** 80 **Difference ≤5** 87 *Men* **Dom ≤35** 87 **Non-dom ≤30** 80 **Sum ≤66** 71 **Difference ≤-2** 100	25/40 (63%)
	Controls (18)	66.67 (6.51)	17.89 (2.06)			*Women* **Dom ≤28** 100 **Non-dom ≤25** 87 **Sum ≤58** 100 **Difference ≤5** 100	

						Men **Dom ≤35** 100 **Non-dom ≤30** 100 **Sum ≤66** 100 **Difference ≤-2** 90	
**Specificity means that the test correctly identified credible effort (i.e. the proportion of true negatives).* *MMSE = Mini-Mental State Examination*							

Embedded PVTs

1. The Repeatable Battery for the Assessment of Neuropsychological Status (RBANS; Randolph, Tierney, Mohr & Chase, 1998).

The RBANS is a widely used neurocognitive battery commonly used in the assessment of dementia. Two RBANS embedded measures of effort have been developed: the Effort Index (EI; Silverberg, Wertheimer & Fichtenberg, 2007) and the Effort Scale (ES; Novitski, Steele, Karantzoulis & Randolph, 2012). Seven of the eight studies which investigated these embedded indices, examined the utility of the EI, four examined the utility of the ES and one paper compared both the EI and the ES against two novel RBANS embedded indices.

a. *RBANS Effort Index (EI; Silverberg, Wertheimer & Fichtenberg, 2007).* The EI combines performance on two subtests (Digit Span and List Recognition) to produce an 'effort index'. A score of greater than 3 is suggestive of non-credible effort. In the development of the EI, it was found that very poor performance on both these subtests was extremely rare in a mixed clinical sample of patients with neurological disorders. See Table 10 for results of studies investigating the EI.

All 4 of the studies which use a reference standard in order to divide participants into credible and non-credible effort groups, investigated the use of the EI. Sensitivity levels for the EI were 0.50, 0.89, 0.93 and 0.95. Unlike the TOMM studies, dementia severity does not appear to explain the differing sensitivity levels found for the EI. Bortnik et al. (2013) who report a sensitivity level of only 0.50 in their 'suspect effort' dementia sample also report that this group had a mean MMSE score of 17.9. Compare this to Paulson, Horner and Bachmann (2015) who found a sensitivity level of 0.93 in a dementia sample with a mean MMSE score of 23.1. It should be noted however that the Bortnik et al. (2013) study included a very small sample (n=9) and scored relatively poorly on the STARD (20/28) due to inadequate

reporting. Specificity levels were reported by all 7 papers and range from a low of 0.41 (Dunham, Shadi, Sofko, Denney & Calloway) to a relative high of 0.69 (Bortnik, Horner & Bachman, 2013). Only controls and MCI samples produced acceptable levels of specificity on the EI.

Table 10. RBANS Effort Index (EI; Silverberg, Wertheimer & Fichtenberg, 2007) – all results refer to a cut-off of >3.

Study	Sample (n)	Mean Age (SD)	Mean Education (SD)	Impairment measure	Mean Score on Impairment Measure (SD)	±Effort Index Sensitivity (%)	*Effort Index Specificity (%)	Quality Rating score
Paulson et al. (2015)	Good effort (189)	69.2 (9.4)	12.9 (3.2)	MMSE RBANS Total scores	27.5 (2.2) MMSE 85 (12.2) RBANS	n/a	63	27/28 (96%)
	Suspect effort (45)	64.2 (10.9)	12.3 (2.9)		23.1 (4.9) MMSE 59.9 (10.7) RBANS	93	n/a	
Burton et al. (2015)	AD (90)	Not reported	Not reported	CDR	Not reported	n/a	51	27/40 (68%)
	Non-AD dementia (55)					n/a	54	
Dunham et al. (2014)	Dementia (46)	76.44 (10.49)	11.17 (2.82)	RBANS Total scores	57.48 (11.70)	n/a	41	20/28 (71%)
	[a]Simulators (44)	27.82 (9.01)	16.41 (0.82)		48.52 (8.66)	89	n/a	
Bortnik et al. (2013)	Good effort (119)	77 (7)	11.42 (4)	MMSE	20.8 (5)	n/a	69.6	20/28 (71%)
	Suspect effort (9)	72 (8.13)	10.44 (2.3)		17.9 (4.5)	50	n/a	
Novitski et al. (2012)	aMCI (15)	80.61 (6.33)	Not reported	RBANS total scores	64.58 (12.89)	n/a	Not reported (AUC 0.608)	21/40 (53%)
	Probable AD (54)					n/a	Not reported	

	Controls (540)					n/a	Not reported	
Duff et al. (2011)	AD (126)	76.7 (6.8)	<9yrs=12 9-11yrs=14 12yrs= 38 13-15yrs=10 >15yrs=26	RBANS Total scores	65.9 (5.9)	n/a	67.1	31/40 (78%)
	aMCI (72)	72 (82.1)	<9yrs=0 9-11yrs=3 12yrs=17 13-15yrs=23 >15yrs=57		92.3 (9.1)	n/a	100	
	Controls (796)	73.4 (5.9)	<9yrs=5 9-11yrs=11 12yrs=26 13-15yrs=30 >15yrs=28		95.7 (13.3)	n/a	97.1	
Barker et al. (2010)	Suspect effort (45)	67.1 (10.3)	12 (3)	RBANS total index scores	IM 57 (16.4) DM 51.6 (14.7) A 60.7 (15.6) L 72 (20.6) VS 70.7 (20.9)	n/a	42.2	22/28 (79%)
	Good effort (258)	72.5 (8.8)	12.5 (3.5)		IM 75.1 (15.8) DM 76 (19.9) A 81.9 (16.1) L 88.3 (10.6) VS 84.3 (17.7)	95.3	n/a	

**Specificity means that the test correctly identified credible effort (i.e. the proportion of true negatives).*
±Sensitivity means that the test correctly identified non-credible effort (i.e. the proportion of true positives).

MMSE = Mini-Mental State Examination RBANS = Repeatable Battery for the Assessment of Neuropsychological Status CDR = Clinical Dementia Rating
aMCI = amnestic MCI AD = Alzheimer's Dementia IM = Immediate Memory DM = Delayed Memory A = Attention L = Language VS – Visuospatial

b. RBANS Effort Scale (ES; Novitski, Steele Karantzoulis & Randolph, 2012). The RBANS Effort Scale is based on the premise that when an individual has genuine memory impairment, their performance on free recall tasks (List Recall, Story Recall and Figure Recall subtests) will decline to close to zero before decline in List Recognition is seen. The authors of the ES propose a cut-off of <12 as being suggestive of non-credible effort (the ES calculation is: List Recognition – (List Recall + Story Recall + Figure Recall)) + Digit Span). See Table 11 for studies investigating the ES.

The studies investigating the ES scored relatively poorly on the quality checklists due to inadequate and/or missing information making it difficult to examine the results across studies. In the original validation study, Novitski et al. (2012) compared dementia and MCI participants with a poor effort group of mild head injury participants, they report an impressive area under the curve (AUC) of 0.908. The paper scored poorly however on the CCAT (21/40) due to inadequate information given regarding data collection, inclusion and exclusion criteria and participant demographic information.

Burton, Enright, O'Connell, Lanting & Morgan (2015) compared the ES in AD and non-AD dementias. They found an impressively high specificity level of 0.96 in their AD sample, however this fell to 0.69 in their non-AD sample. The authors conclude that the ES may be an appropriate effort index to use when assessing effort level in people with a dementia of the Alzheimer's type, however in clinical practice, effort tests are most likely to be given during the diagnostic process when a person's diagnosis is as yet unknown. It does raise the issue that the ES is based on the premise that a person's performance on tests of free recall will decline before their performance on tests of recognition. This profile of impairment is more likely to be seen in AD, which is characterised by a deficit in episodic memory, than in non-AD dementias.

Paulson et al. (2015) and Dunham et al. (2014) investigated the ES in both good and suspect effort groups. They found sensitivity of 0.71 and 0.88 and specificity levels of 0.42 and 0.81 respectively. It is not clear why the studies found very different specificity levels and due to inadequate reporting by the studies it was impossible to investigate whether type of dementia had a bearing on the discrepancy in the results found (i.e. if the reason for the high failure rate in Paulson's good effort group was due to the majority having a non-AD dementia).

Table 11. RBANS Effort Scale (ES; Novitski, Steele, Karantzoulis & Randolph) – all results refer to a cut-off of <12.

Study	Sample (n)	Mean Age (SD)	Mean Education (SD)	Impairment measure	Mean Score on Impairment Measure (SD)	±Effort Scale Sensitivity (%)	*Effort Scale Specificity (%)	Quality Rating score
Paulson et al. (2015)	Good effort (189)	69.2 (9.4)	12.9 (3.2)	MMSE RBANS Total scores	27.5 (2.2) MMSE 85 (12.2) RBANS	n/a	42	27/28 (96%)
	Suspect effort (45)	64.2 (10.9)	12.3 (2.9)		23.1 (4.9) MMSE 59.9 (10.7) RBANS	71	n/a	
Burton et al. (2015)	AD dementia (53)	Not reported	Not reported	CDR	Not reported	n/a	96	27/40 (68%)
	Non-AD dementia (36)					n/a	69	
Dunham et al. (2014)	Dementia (46)	76.44 (10.49)	11.17 (2.82)	RBANS Total scores	57.48 (11.70)	n/a	81	20/28 (71%)
	[a]Simulators (44)	27.82 (9.01)	16.41 (0.82)		48.52 (8.66)	88	n/a	
Novitski et al. (2012)	aMCI (15)	80.61 (6.33)	Not reported	RBANS total scores	64.58 (12.89)	n/a	Not reported (AUC 0.908)	21/40 (53%)
	probable AD (54)						Not reported	
	Controls (540)						84.9	

**Specificity means that the test correctly identified credible effort (i.e. the proportion of true negatives). MMSE = Mini–Mental State Examination CDR = Clinical Dementia Rating*
±Sensitivity means that the test correctly identified non-credible effort (i.e. the proportion of true positives). RBANS = Repeatable Battery for the Assessment of Neuropsychological Status
AD = Alzheimer's Dementia

b. *Performance Validity Index (PVI; Paulson, Horner & Bachman, 2015) and Charleston Revised Index of Effort for RBANS (CRIER; Paulson, Horner & Bachman, 2015).* Paulson and colleagues evaluated the EI and ES against two novel RBANS embedded indices of effort, the PVI and the CRIER. The PVI cut-off for detecting invalid responding is <42 and is calculated as follows:

RBANS PVI: List recall + story recall + figure recall + digit span + list recognition

The CRIER cut-off for detecting invalid responding is <24 and is calculated as follows:

RBANS CRIER: list recall + story recall + figure recall + digit span + list recognition – GDS

The PVI was found to have sensitivity of 0.82, specificity of 0.77 and AUC of 0.90 whilst the CRIER had sensitivity of 0.84, specificity of 0.90 and AUC of 0.94. These are promising results and warrant further research.

2. Reliable Digit Span (RDS; Greiffenstein, Baker & Gola, 1994).

Greiffenstein, Baker, and Gola (1994) originally derived the RDS from the Digit Span subtest of the Wechsler Adult Intelligence Scale–Revised (Wechsler, 1981) by 'summing the longest string of digits repeated without error over two trials under both forward and backward conditions' (pp. 219-220). Table 12 illustrates the results of studies investigating the RDS.

RDS specificity was found to be greatly impacted by severity of cognitive impairment. It had high specificity for Loring's dementia (87%), MCI (96%) and control samples (97%) and also for Zenisek's (2016) MCI sample (94.6%) whilst Kiewel, Wisdom, Bradshaw, Pastorek and Strutt (2012) found high specificity for their mild dementia sample (89%). All of these samples had MMSE scores of 23 or more (except for Zenisek et al. 2016 who used the MoCA). In contrast, with Kiewel's more impaired groups, the RDS yielded an unacceptably high level of false positives similar to those in the equally impaired Dean et al. (2009) study. Zenisek et al. (2016) divided their sample into subgroups based on type of dementia diagnosed and found differences in specificity levels with a high of 85.2 for dementia with Lewy bodies sample and a low of 73.3 for their frontotemporal dementia sample.

Table 12. Reliable Digit Span (RDS; Greiffenstein, Baker & Gola, 1994) – all results refer to a cut-off of ≤6.

Study	Sample (n)	Mean Age (SD)	Mean Education (SD)	Impairment measure	Mean Score on Impairment Measure (SD)	*RDS Specificity (%)	Quality Rating score (CCAT)
Zenisek et al. (2016)	AD (133)	72.6 (7.8)	14.8 (2.9)	MoCA	21 (4.8)	80.5	23/40 (58%)
	VAD (8)					75	
	DLB (27)					85.2	
	FTD (15)					73.3	
	MCI (168)					94.6	
Loring et al. (2016)	AD (178)	75.7 (7.5)	Not reported	MMSE	23.3 (2)	87	31/40 (78%)
	aMCI (365)	74.9 (7.2)			27 (1.8)	96	
	Controls (206)	76 (5)			29.1 (1)	97	
Kiewel et al. (2012)	Dementia (mild)	74.6 (8.8)	14.5 (2.7)	MMSE	23.4 (3.1)	89	30/40 (75%)
	Dementia (mod)	76.5 (9.4)	14.2 (2.7)		16.8 (2.9)	76	
	Dementia (sev)	71.2 (10.6)	13.5 (2.7)		7.7 (3.4)	17	
Dean et al. (2009)	Dementia (172)	Not reported	Not reported	MMSE	19.2 (4.4)	70	30/40 (75%)

**Specificity means that the test correctly identified credible effort (i.e. the proportion of true negatives). MMSE = Mini-Mental State Examination* a*MCI = amnestic MCI*
AD = Alzheimer's Dementia VAD = Vascular Dementia DLB = Dementia with Lewy Bodies FTD = Frontotemporal Dementia

Discussion

This review included 25 papers examining the diagnostic accuracy of 14 embedded and stand-alone tests of effort in dementia and MCI samples. In particular this review sought to establish which PVTs were most sensitive to non-credible effort whilst being insensitive to the cognitive impairment seen in individuals who meet criteria for these diagnoses. The effort tests which were found to be most sensitive to non-credible effort whilst being least sensitive to dementia-related cognitive impairment were the three PVTs devised by Green (2003, 2004, 2008). The WMT, MSVT and NV-MSVT go beyond the pass/fail approach of traditional effort tests and instead use profile analysis to determine if the pattern of results which an examinee produces are indicative of poor effort or if they show a 'dementia profile'. The vast majority of the studies which investigated these tests found 100% specificity. The only studies which found low levels of specificity for these PVTs were those which calculated Criterion A only and did not use profile analysis to determine effort level (i.e. determined effort level on whether the participants scored below cut-off on the effort tests but did not determine whether a dementia profile was present). The WMT, MSVT and NV-MSVT therefore appear to be the most appropriate effort tests for use in cognitive assessment when a dementia is queried. These results should be taken as provisional however since, although the evidence is promising there is not a wealth of literature to draw from. It should also be noted that producing a "dementia profile" is not in itself proof of adequate effort and that clinicians should aim to employ a multi-method approach to effort testing such as that devised by Slick et al. (1999) to highlight any inconsistencies between PVT results and information gathered from other sources. Some of the other tests explored in this review were found to have acceptable specificity levels when used with MCI or mild dementia samples such as the TOMM, the Coin in the Hand and the RDS. These PVTs do however lose specificity with more impaired samples.

Methodological limitations of effort testing literature

Compared to other areas of research in clinical neuropsychology, effort testing research faces significant methodological challenges since no 'gold-standard' exists with which to reliably determine performance validity. Five of the 25 papers included in this review attempted to independently assess effort by use of a reference standard (three used the TOMM, one used the MSVT and one used the RBANS Effort Scale). The remaining 20 papers either excluded participants who may have had an incentive to feign impairment or they assumed good effort by virtue of the participants having a diagnosis of dementia. This makes it difficult to know what the true false-positive rate actually is.

Diagnostic tests should be evaluated in samples that are representative of those with whom the test will be used in practice. The majority of the studies included in this review use a methodology whereby participants with an established dementia are compared to healthy controls. This is likely to lead to bias since the participants included have a more advanced stage of the disease (in this case dementia) than studies using a clinical sample of consecutive referrals to memory clinics. Indeed there appears to be a positive correlation between PVT specificity and severity of cognitive impairment, such that as scores on cognitive tests decrease, so too does the specificity level of a PVT.

Also, very few of the studies included samples with a diagnosis of MCI, a population of importance when evaluating the diagnostic accuracy of effort tests in older adults presenting with memory difficulties. Additionally, very few of the studies included in this review were able to assess sensitivity levels because they only included participants who were deemed to exert credible effort from the outset, either due to not being involved in litigation/claiming disability benefits or simply by virtue of having a diagnosis of dementia.

Finally, the studies included in this review largely approach the subject of effort in its 'malingering' definition whereby examinees are deemed to be feigning impairment. As

discussed previously, indications of poor or atypical motivation are not always the result of deliberate malingering, rather there are many reasons for an individual to put forth less than credible effort such as low mood, stress, conversion disorder, medication side effects and fatigue amongst others. The majority of the studies in this review did not assess for these factors.

Methodological limitations of the current review

This systematic review faced some methodological limitations. First, it was out with the scope of the review to include every study which has investigated the diagnostic utility of PVTs in dementia samples. The current study focused on papers which included dementia/MCI samples as their primary interest. There are, however, other studies which include a dementia sample alongside other clinical samples. Additionally, as previously mentioned, it was also out with the scope of the current review to examine every PVT investigated by the papers included therefore only the PVTs which are known to be most used in clinical practice were reviewed.

Finally, some of the studies examined the performance of the PVTs across different cut-offs. It was not possible to review each of these cut-offs therefore those most used in clinical practice were evaluated.

Conclusions

Future research on the diagnostic accuracy of PVTs in older adults should aim to focus on the recruitment of consecutive referrals to memory clinics and should employ a multi-method approach such as that proposed by Slick et al. (1999). As part of the cognitive test battery, the participants should receive a reference standard (such as the MSVT or NV-MSVT) and the PVT of interest (the index test). The index test should only be calculated

once the final diagnosis is known. This method would allow for both sensitivity and specificity of the PVT in question to be investigated since both good effort and suspect effort participants would be included. Future research should also seek to review the diagnostic utility of PVTs across various cut-offs in MCI/dementia samples. Future research may also seek to further evaluate the use of the two new RBANS embedded indices (PVI and CRIER) created by Paulson et al. (2015) as these showed good potential and the RBANS is a commonly used tool in older adult memory clinics. It is also important for future research to look at subtypes of dementia and how profiles of impairment impact PVT performance. It is likely that some of the variation in performance seen across the different studies may be partially accounted for not just by dementia severity but by dementia subtype. Burton et al. (2015) touched on this issue when they found that the Effort Scale had significantly higher specificity in their Alzheimer's sample (96%) compared to their non-Alzheimer sample (69%).

Following evaluation of the studies in this review, tests which take a hierarchical approach to effort testing such as the WMT, MSVT and the NV-MSVT may be the best PVTs to use in clinical practice given that these tests have been found to be particularly robust in dementia samples (i.e. they have very low false-positive rates). It should be noted however that Green's tests can be lengthy to administer which may be outwith the scope of some clinical contexts.

Finally it must be stressed that determining an individual's level of effort should not be judged on the basis of scores on an effort test alone. It should also be noted that the vast majority of older people referred for memory assessment will have no incentive to purposefully feign impairment.

Funding

This study was supported by NHS Greater Glasgow and Clyde.

References

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

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.

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

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

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

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

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L., . . . Cohen, J. F. (2015). STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies. *Radiology, 277*(3), 826-832.

Burton, R. L., Enright, J., O'Connell, M.,E., Lanting, S., & Morgan, D. (2015). RBANS embedded measures of subcredible effort in dementia: Effort scale has a lower failure rate than the effort index. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists, 30*(1), 1-6.

Crowe, M., & Sheppard, L. (2011). A general critical appraisal tool: an evaluation of construct validity. *International Journal of Nursing Studies, 48*(12), 1505–1516.

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

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

Dunham, K. J., Shadi, S., Sofko, C. A., Denney, R. L., & Calloway, J. (2014). Comparison of the repeatable battery for the assessment of neuropsychological status effort scale and effort index in a dementia sample. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists, 29*(7), 633-641.

Faust, D., Hart, K., & Guilmette, T. J. (1988). Pediatric malingering: The capacity of children to fake believable deficits on neuropsychological testing. *Journal of Consulting and Clinical Psychology, 56*(4), 578-582.

Faust, D. (1995). The detection of deception. *Neurologic Clinics, 13*(2), 255-265.

Green, P. (2003). *Word Memory Test for Windows: User's manual and program.* Edmonton: Green's Publishing.

Green, P. (2008). *Test manual for the nonverbal Medical Symptom Validity Test.* Edmonton: Green's Publishing.

Green P., Montijo J., & Brockhaus, R. (2011). High specificity of the word memory test and medical symptom validity test in groups with severe verbal memory impairment. *Applied Neuropsychology, 18*(2), 86-94.

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

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.

Hampson, N. E., Kemp, S., Coughlan, A. K., Moulin, C. J. A., & Bhakta, B. B. (2014). Effort test performance in clinical acute brain injury, community brain injury, and epilepsy populations. *Applied Neuropsychology: Adult, 21*(3), 183-194.

Heaton, R. K., Smith, H. H. J., Lehman, R. A., & Vogt, A. T. (1978). Prospects for faking believable deficits on neuropsychological testing. Journal of Consulting & Clinical Psychology, 46(5), 892-900.

Heilbronner, R. L., Sweet, J. J., Morgan, J. E., Larrabee, G. J., Millis, S. R., & Participants1, C. (2009). American academy of clinical neuropsychology consensus conference statement on the neuropsychological assessment of effort, response bias, and malingering. *The Clinical Neuropsychologist, 23*(7), 1093-1129. doi:10.1080/13854040903155063

Heinik, J., Solomesh, I., & Berkman, P. (2004). Correlation between the CAMCOG, the MMSE, and three clock drawing tests in a specialized outpatient psychogeriatric service. *Archives of Gerontology and Geriatrics, 38*(1), 77-84.

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

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

Kapur, N. (1994). The coin-in-the-hand test: A new "bed-side" test for the detection of malingering in patients with suspected memory disorder. *Journal of Neurology, Neurosurgery, and Psychiatry, 57*(3), 385-386.

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

Larrabee, G. J. (2012). Performance validity and symptom validity in neuropsychological assessment. *Journal of the International Neuropsychological Society, 18*, 1-7.

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

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

performance validity measures in amnestic mild cognitive impairment and early alzheimer disease. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists, 31*(4), 313-331.

Martins, M., & Martins, I. P. (2010). Memory malingering: Evaluating WMT criteria. *Applied Neuropsychology, 17*(3), 177-182. doi:10.1080/09084281003715709

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.

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: British Psychological Society.

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

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.

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

Paulson, D., Horner, M. D., & Bachman, D. (2015). A comparison of four embedded validity indices for the RBANS in a memory disorders clinic. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists, 30*(3), 207-216.

Randolph, C., Tierney, M. C., Mohr, E., & Chase, T. N. (1998). The repeatable battery for the assessment of neuropsychological status (RBANS): Preliminary clinical validity. *Journal of Clinical & Experimental Neuropsychology: Official Journal of the International Neuropsychological Society, 20*(3), 310-319.

Reitan, R.M., & Wolfson, D. (1985). *The Halstead-Reitan Neuropsychological Test Battery: Theory and Interpretation.* Tucson, AZ: Neuropsychology Press.

Rey, A. (1941). L'examen psychologique dans les cas d'encéphalopathie traumatique. (les problems.). = the psychological examination in cases of traumatic encepholopathy. problems. *Archives De Psychologie, 28*, 215-285.

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

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

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

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.

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

Scottish Intercollegiate Guidelines Network (SIGN). 2007. Methodology checklist 5: studies of diagnostic accuracy. In A Guideline Developer's Handbook. SIGN: Edinburgh, Annex B.

Silverberg, N. D., Wertheimer, J. C., & Fichtenberg, N. L. (2007). An effort index for the repeatable battery for the assessment of neuropsychological status (RBANS). *The Clinical Neuropsychologist, 21*(5), 841-854.

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

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

Slick, D. J., Tan, J. E., Strauss, E. H., & Hultsch, D. F. (2004). Detecting malingering: A survey of experts' practices. Archives of Clinical Neuropsychology: *The Official Journal of the National Academy of Neuropsychologists, 19*(4), 465-473.

Suesse M., Wong V.W., Stamper L.L., Carpenter K.N., & Scott, R. B. (2015). Evaluating the clinical utility of the medical symptom validity test (MSVT): A clinical series. *The Clinical Neuropsychologist, 29*(2), 214-231.

Teichner, G., & Wagner, M. T. (2004). The test of memory malingering (TOMM): Normative data from cognitively intact, cognitively impaired, and elderly patients with dementia. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists, 19*(3), 455-464.

Tombaugh, T.N. (1996). The test of memory malingering. Toronto, ON, Canada: MultiHealth 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.

van Gorp, W. G., Humphrey, L. A., Kalechstein, A., Brumm, V. L., McMullen, W. J., Stoddard, M., & Pachana, N. A. (1999). How well do standard clinical neuropsychological tests identify malingering? A preliminary analysis. *Journal of Clinical and Experimental Neuropsychology, 21*(2), 245-250.

van, d. W., Prins, J. B., Jongen, P. J. H., van, d. M., & Bleijenberg, G. (2000). Abnormal neuropsychological findings are not necessarily a sign of cerebral impairment: A matched comparison between chronic fatigue syndrome and multiple sclerosis. *Neuropsychiatry, Neuropsychology, & Behavioral Neurology, 13*(3), 199-203.

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

Zenisek R., Millis S.R., Banks S.J., & Miller, J. B. (2016). Prevalence of below-criterion reliable digit span scores in a clinical sample of older adults. *Archives of Clinical Neuropsychology : The Official Journal of the National Academy of Neuropsychologists*, *31*(5), 426-433.