Assessment of Effort in Clinical Testing of Cognitive Functioning for Adults
Working Party Members

Professor Tom McMillan (Chair), Psychological Medicine, University of Glasgow.
Dr Stuart Anderson, Sussex Rehabilitation Centre.
Professor Gus Baker, Head of Division of Neurosciences, University of Liverpool.
Professor Michael Berger, Representative of the Steering Committee on Test Standards.
Dr Graham Powell, Psychology Services, London.
Robert Knight, British Psychological Society Administrator.
Executive Summary

Purpose
To provide guidance and highlight key issues in the assessment of effort during cognitive testing in adults in clinical settings. In addition to identify issues relevant to the interpretation of findings. Finally to provide further reading and highlight areas for future research.

Readership
This document is for professional psychologists who are trained in clinical assessment of cognitive function in adults.

Key Points
- Cognitive test results are not valid if the testee does not try hard on the tests.
- Effort tests should be given routinely as part of clinical assessment of cognitive function.
- There are some exceptions where routine assessment of effort is not appropriate.
- Failure on effort tests requires careful interpretation. Although a number of causes are possible, deceit should always be considered.
- Clinicians should be aware of the sensitivity and specificity of the effort tests that they use and the base rates of sub-optimal performance in the population from which their testee comes and take these factors into account when interpreting findings.
- Interpretation of failure on effort tests needs to be reported as clearly as possible.
- There is little UK research literature on effort testing; there is a need for this to be developed.
**Purpose**

The aims of the present document are to highlight key issues in the assessment of effort, provide guidance and suggest further reading. Specifically, the document focuses on the assessment of effort as part of a clinical assessment of cognitive function in a wide range of settings. It further aims to identify issues relevant to the interpretation of findings and the potential courses of action when an individual fails cognitive tests of effort. The primary concern of this document is the validity of cognitive symptoms and not with a broader-based assessment of the validity of symptom complaint. The document will highlight important areas for future research.

**The scope of this document**

This document is written for professional psychologists who have been trained in clinical assessment of cognitive function. The majority of research on testing effort is on adults, and it is to this group that this guidance refers. Modification of this document or a further document would be needed if this guidance is to be extended to other client groups, such as children and young people and those with learning disabilities.
Introduction

Cognitive testing is used in many clinical contexts because it is judged that the resulting scores, together with information from other sources collected as part of the assessment process, will advance an understanding of the testee and their needs, and aid decisions about further action. Tests are given as part of purpose-driven psychological assessment processes. An assessment, in the form of a written or verbal statement, constitutes the psychologist’s informed conclusions in relation to the purpose, based on their interpretation of the test scores and the other information. The value of an assessment is therefore crucially dependent on the quality of the contributing test data. Users of psychological tests will be aware of the many factors that influence data quality in testing. These include the psychometric properties of the test, competence of the tester and the multiple influences affecting the test performance of the testee. Good practice in testing and assessment requires the issue of data quality to be addressed and, in particular, includes the testee’s motivation to adhere to the test requirements. Motivation that is at variance with test requirements can distort test findings, limit the relevance of the assessment and even invalidate it. The need for reliable and valid indices sensitive to distortions of motivation explains the development of ‘tests of effort’.

It is important to note that ‘effort’, a proxy for motivation, is a potentially misleading term, implying something that is uni-dimensional, uni-directional, and is encompassed by a test score that itself ranges along a uni-dimensional scale. In fact ‘effort’ is more likely to be a vector, having both magnitude and direction. For instance, a testee can put much effort into doing the opposite of what is required and little or no effort into what is. Further, like other concepts in psychology, ‘effort’ is a higher order concept, assessed and inferred from scores on special tests or test components, and other observations of behaviour. Although perhaps narrower in scope, it is no different in principle to concepts such as ‘intelligence’ and ‘personality’. Each can be conceptualised and tested in a variety of ways and each is the subject of multiple theories.

In applied settings, the test user is interested not only in the score but in particular, its meaning or interpretation. Indications of poor, atypical or unusual motivation can arise for one or more of many possible reasons, including disturbances in mental and physical health, fatigue, neuropsychological dysfunctions, negative beliefs about self-competence, poor communication or understanding of the demands of testing, prior tutoring, or deliberate non-adherence to the requirements of testing. Or the testee may perform to conform with a feigned condition, or to exaggerate a condition they have. Detecting and understanding such distortions of performance represent serious challenges for all test users who are required to reach sustainable conclusions about the pattern of performance. The assessment of effort is complicated by its multi-dimensional or multi-factorial nature, that might reflect possible co-action and interaction of several influences simultaneously and over time on test performance. The nature of the most salient of these will need to be identified and taken into account in order to judge their impact on the test data and the conclusions drawn from the assessment process. Part of the expertise of the test user comprises knowledge of such factors and how to identify and take them into account in an assessment. Tests of effort, either embedded in other tests or as stand-alone procedures,
properly understood and used, can assist this process. Technically, such tests aim to detect signs (patterns of test performance) and symptoms (verbal reports – what the client says), indicative of deliberate distortion and in this context are described as tests of symptom validity. Like other diagnostic tests, part of their psychometric evaluation involves demonstrating the fidelity with which they identify ‘true’ instances while minimising misclassifications. These and related issues are discussed below (see also Appendix 1 for further explanation of terms).

This document is intended to inform practitioners of the major advances and expanding literature concerned with testing effort and the assessment of symptom validity, evidenced for instance by the recent publication of the US National Academy of Neuropsychology position paper (Bush et al., 2005) and to draw on this literature and the experience of practitioners to provide a framework for good practice in relation to clinical testing.

Why is testing of effort an issue?

Psychometric testing is used widely to assess cognitive ability in clinical settings and a crucial skill in forming a judgement is interpretation of results. Scores on cognitive tests are usually considered in the context of published normative data, which in turn assume that the testee has conformed to the instructions of the tester to ‘try hard’ during testing. Clearly if effort is sub-optimal, then test data are likely to be invalid, erroneous conclusions may be reached and as a consequence, inappropriate recommendations given. There has been considerable interest in recent years in identifying such sub-optimal performance. This is evidenced by the burgeoning literature on the assessment of symptom validity, which in turn mirrors the development of tests of ‘effort’ that aim to be highly sensitive and specific, and that can be administered easily in a routine clinical situation (Bush et al 2005). In contrast, available evidence (although limited) suggests that clinical judgement is a poor means of detecting faking of cognitive impairment (Heaton et al., 1978; Faust et al., 1998).

Is sub-optimal effort common?

It is now well established that effort is an important determinant of performance on psychological tests and there is a substantial evidence base that supports the use of effort testing in neuropsychological assessment (e.g. Constantinou et al., 2005; Drane et al., 2006; Green et al., 2001; Green 2007a; Lynch 2004; Stevens et al., 2008). In an investigation of 904 individuals involved in litigation or claiming financial benefits for disability, Green et al. (2001) report that 53 per cent of variance on a composite neuropsychological test battery was explained by effort. This contrasted with 11 per cent of the variance shared by education, four per cent by age and five per cent by brain injury severity. Similar findings were reported in a UK study by Moss et al. (2003). These striking figures should be considered in the context of sample characteristics of these studies which tend to involve litigants.

Much of the research literature on the frequency of occurrence or base rates of sub-optimal effort has concentrated on purposeful exaggeration of cognitive impairment. Given the covert nature of deceit, base rates are difficult to determine with confidence. Larrabee (2003) reviewed 11 studies (N=1363) on mild head injury litigants and 15 to
64 per cent (average 40 per cent) were reported to be malingering. Overall, the figure may be higher given the likely prevalence of disability caused by mild head injury (five per cent) and the prevalence of complaints in litigation involving mild head injury (over 90 per cent) cases (Mittenberg et al., 2002). Malingering may also be common in disability claimants, with findings of 54 to 72 per cent probable or definite, reported in US studies (Miller et al., 2004; Chafetz 2008). In the criminal forensic field, 54 per cent of cases have been reported as probable or definite malingers (Ardolf et al., 2007). It is important that base rate data is considered when interpreting findings on effort tests (see Appendix 2 and Chapter 16 in Strauss, Sherman & Spreen, 2006). Such rates cannot be directly applied to the UK context, as all of the above studies are based on North American samples, although in general, base rates of malingering are likely to be higher in cases where financial reward is associated with the assessment than where assessment is purely for clinical purposes.

Can sub-optimal effort be detected?

The evidence that informs the importance of effort testing, has led to the creation of purpose-built tests designed to detect non-credible cognitive symptoms as well as the development of measures within existing cognitive tests (i.e. embedded tests). While effort testing will not guarantee detection of poor effort, it will help the assessor to make a judgement about the reliability of the collected data.

The literature on effort testing tends to evaluate the discriminative ability of tests in terms of sensitivity and specificity (see Strauss, Sherman & Spreen, 2006 and Appendices 1 and 2 for elaboration and further explanation of these and related terms). Generally speaking, the trend has been to set desired specificity at high levels (greater than 90 per cent) in order to minimise the risk of false positive errors (i.e. to identify effort as sub-optimal when good effort has been applied). Sensitivity and specificity data continue to be published, and the reader is referred to recent reviews (see Boone, 2007; Larrabee, 2007). In Appendix 3 a representative (rather than an exhaustive) list of purpose-designed tests of effort is given. The list is based on measures frequently used in North America and data are based on US samples (see also Sharland & Gfeller, 2007). However, many of these tests also seem to be used commonly in the UK (McCarter et al., 2009).

The use of embedded measures in conventional psychometric tests

There are a number of reasons why clinicians may need to consider alternatives and/or additions to purpose-designed tests of effort. At a practical level, the inclusion of effort tests increases assessment time. Furthermore, commonly used effort tests may be vulnerable to client coaching. Finally, there is an emerging consensus that clinicians should adopt a multi-method (and multi-test) approach to the testing of effort (Slick et al., 1999; Strauss, Sherman & Spreen, 2006; Boone, 2007; Larrabee, 2007).

Various conventional psychometric tests have been investigated for their sensitivity to suboptimal effort and the reader should consult Strauss, Sherman and Spreen (2006), Boone (2007) and Larrabee (2007) for listings and reviews of these various studies. Some of the more commonly used approaches are listed in Table 1. None of the indices listed in Table 1 are recommended for use as sole measures for detecting poor effort.
Table 1: Effort-sensitive measures embedded in conventional psychometric tests.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Representative Studies</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>WAIS-III Reliable Digit Span (RDS)</td>
<td>Iverson &amp; Tulsky (2003); Axelrod et al. (2006)</td>
<td>Computed by adding the longest string of digits forwards and backwards in which both trials at that level are passed, the RDS has been well validated for use as a measure of effort. A cut-off score of &lt;7 provides generally good specificity but variable sensitivity (Larabee, 2007).</td>
</tr>
<tr>
<td>WAIS-III Vocabulary minus Digit Span</td>
<td>Iverson &amp; Tulsky (2003); Miller et al. (2004); Larrabee (2003)</td>
<td>Iverson &amp; Tulsky (2003) suggest a low base rate (five per cent or less in general and clinical populations) for Vocabulary minus Digit Span difference scores of 5 or greater. Miller et al. (2004) found 90 per cent specificity for a difference score of 4 or more in non-litigating/compensation-seeking samples.</td>
</tr>
<tr>
<td>WMS-III Rarely Missed Index (RMI)</td>
<td>Killgore &amp; DellaPietra (2000)</td>
<td>Killgore &amp; DellaPietra (2000) identified six items in the Logical Memory recognition trial that were correctly endorsed at above chance levels by normal participants naïve to story content. A weighted scoring system forms the RMI which demonstrated excellent sensitivity (97 per cent) and specificity (100 per cent) in the validation study comparing analogue malingerers with neurological controls.</td>
</tr>
<tr>
<td>WMS-III Auditory Recognition Delay Index</td>
<td>Langeluddecke &amp; Lucas (2003)</td>
<td>Using a cut-off score of &lt;43 on the ARD and &lt;18 on word-list recognition, Langeluddecke and Lucas (2003) found sensitivity and specificity values of 80 per cent and 91 per cent for probable malingering (as identified by failed WRMT performance).</td>
</tr>
<tr>
<td>WMS-III Mental Control Test</td>
<td>Kelly et al. (2005)</td>
<td>This optional WMS-III task was investigated for its discriminative ability (TBI controls, simulators and non-simulators) in a study by Kelly et al. (2005). Using a cut-off score of &lt; 13.5, results revealed sensitivity and specificity values of 80 per cent and 82 per cent respectively.</td>
</tr>
<tr>
<td>RBANS Effort Index</td>
<td>Silverberg et al. (2007)</td>
<td>Recognising the value of word list recognition and digit span as measures of suboptimal effort, Silverberg and colleagues derived an RBANS Effort Index based on these subtest scores. A validation study using clinical TBI and malingering samples as well as three simulator groups found good to excellent discrimination with an overall classification accuracy of 86.9 per cent.</td>
</tr>
<tr>
<td>Measure</td>
<td>Representative Studies</td>
<td>Comment</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>----------------------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Rey Auditory Verbal Learning Test (RAVLT)</td>
<td>Barrash, Suhr &amp; Manzel (1998, 2004)</td>
<td>Delayed recognition score has been the most widely used index for detecting poor effort, although there is considerable overlap in the score distribution of probable malingerers and clinical controls (Larrabee, 2007). See Boone et al. (2005) and Barrash et al. (2004) for a combination approach using several AVLT indices.</td>
</tr>
<tr>
<td>Warrington Recognition Memory Test (WRMT)</td>
<td>Iverson &amp; Franzen (1998); Millis (2002)</td>
<td>In a simulator study paradigm, Iverson &amp; Franzen (1998) found that a score of &lt;32 on the Words subtest to be successful in detecting 90 per cent of malingerers with 100 per cent specificity. A score of &lt;30 on the Faces subtest showed 80 per cent sensitivity to malingering with 100 per cent specificity. Millis provides similar data on a TBI group.</td>
</tr>
<tr>
<td>Wisconsin Card Sorting Test (WCST)</td>
<td>Greve et al. (2002); Larrabee (2003)</td>
<td>Mixed findings are evident. The data appears to favour the use of multivariate formulae rather than a single index. An increased rate of false positive errors occurs in the face of brain injury severity and so caution should be exercised when using the WCST to detect poor effort in cases other than mild TBI.</td>
</tr>
<tr>
<td>Multiple measure contrast approach</td>
<td>Larrabee (2003); Nelson et al. (2003); Meyers &amp; Volbrecht (2003)</td>
<td>Some studies have attempted to operationalise the multi-test/ multi-method of detecting poor effort recommended by Slick et al. (1999). Larrabee (2003) investigated patterns of failure on embedded measures of symptom validity and found that sensitivity could be improved by aggregating across multiple measures without appreciably altering specificity. Meyers &amp; Volbrecht (2003) looked at the rate of failure on nine measures of response validity in various clinical groups, including litigating and non-litigating patients. They found that failure of more than two indicators was rare in non-institutionalised patients. However, they cautioned that their multiple indicator method was less appropriate for patients with stroke, dementia, mental retardation or those receiving 24-hour institutional care.</td>
</tr>
</tbody>
</table>

Many adult psychometric tests of personality, psychopathology and adjustment (e.g. *MMPI-II, MCMI-III, BRIEF-A*) incorporate built-in validity scales that may be sensitive to symptom augmentation, symptom denial and/or inconsistency. While it may be difficult to form an opinion about suboptimal effort using these scales, they are likely to be useful in providing evidence of non-credible or inconsistent performance.
What is optimal versus practical assessment?

In their recommendation for an ‘adequate assessment of response validity in neuropsychological assessment’ (p.425), Bush et al. (2005) recognise the contextual use of effort tests. Specifically, while they strongly recommend the use of effort tests in legal contexts, they recognise that their use in clinical settings may not always be indicated and/or appropriate (e.g. some individuals who require 24-hour supervised care; see below). They recommend that neuropsychologists in clinical settings outwith legal contexts assess symptom validity in the manner that is most appropriate for that context. Certainly, there is a strongly emerging view in North America, that effort testing should be routinely carried out, other than in exceptional cases (Bush et al., 2005). A multi-method and multi-test approach has also been recommended for the proper evaluation of effort in neuropsychological evaluations (Slick et al., 1999; Bush et al., 2005).

It is recognised that there is a diversity of practice settings in the UK and effort testing may be problematic in certain contexts and with certain client groups. For example, the reliability of effort testing can be adversely affected by sensory or motor impairment (as might be seen following stroke), in individuals with dementia or psychiatric illness or during early recovery from neurological injury or illness (e.g. during post-traumatic amnesia after head injury). Careful consideration of the usefulness of effort tests is needed in these situations.

If formal effort testing has been conducted and ‘failed’, particular care should be taken when interpreting findings because there can be several reasons for failure. Some propose alternative diagnostic categories for classification of individuals where there is insufficient evidence of intention to deceive or of symptom production to warrant a diagnosis of malingering (Delis & Wetter, 2007). There are published criteria that can aid clinicians in decision making about effort and malingering (Slick et al., 1999; Bianchini et al., 2005; Delis & Wetter, 2007).

What to do if detecting poor effort

(a) Discontinue session immediately? There is a need to obtain sufficient information in order to produce a more general formulation (i.e. beyond simply reporting cognitive test results). It is risky to discontinue as data may be inadequate. Advice is to continue the session.

(b) Discontinue testing? Generally there is a need to explore patterns of test performance and to consider the possibility (even if unlikely) that the testee underperformed only on the test of effort. Advice is to continue testing.

(c) Modify planned assessment? This is often appropriate; either to include further tests (e.g. a second test of effort) or to reduce the number of tests if the person seems to be generally underperforming. For the latter option, it may be important not to give some measures, in order to preserve them for a time when the person may be performing with more appropriate effort, but to give at least some tests as a baseline and to aid formulation.

(d) Continue as planned? This may not be justified; however it is recommended that the tester actively reconsiders their plan. They have to be able to form a view about why the person has failed the effort test.
Interpretation of failure on effort tests

Failure on a test of effort may be because of purposeful intent to deceive, but this is only one possibility and other causes must always be considered. Examples are abnormal arousal (e.g. during severe anxiety or drug induced hyperarousal or drug induced stupor), severe psychiatric disorder (e.g. acute psychosis), sensory impairment, poor compliance with testing, iatrogenic symptoms (see Appendix 4) or somatoform disorders. Of these, the last two can be the most difficult to distinguish from malingering because there is not motivation to underperform, but a belief in impaired capacity (Delis & Wetter, 2007). However, these possibilities should be balanced against the fact that effort tests are designed to be very easy with scores that are robust and not easily affected by factors such as anxiety or depression. For example, Boone (2007, p.297) sets out results of studies that examine the impact of anxiety and depression on one well known effort test, and reports that failure was rare.

Mittenberg et al. (2002) and Sharland and Gfeller (2007) suggest that there are a number of factors that are useful when considering the cause of underperformance. These include:

1. The severity of cognitive impairment in relation to the index event.
2. The pattern of cognitive test performance.
3. Scores that are below empirical cut-offs on forced choice tests.
4. Discrepancies among records, self-report and observed behaviour.
5. Implausible self-reported symptoms at interview.
6. Implausible changes in test scores across repeated examinations
7. Unusual or bizarre errors observed during the interview not captured by the test.

This array of possibilities emphasises the need to consider information from a variety of sources when interpreting results from tests of effort and achieving a formulation. Included here would be data from other cognitive test sessions, medical records, concurrent diagnoses, career and social background, self-report of complaints, psychological state and behaviour during interview, separate interview with a relative. It may become evident that further information should be obtained after the assessment in order to give an opinion and this needs to be actively considered.

Motivation to deceive should be considered and in doing so the context of the assessment must be taken into account (e.g. clinical, occupational, litigation). If it is believed that effort has been suboptimal, as clear a view as possible about the cause of this should be reported.

The assessor-testee interface: What should the testee be told?

There is a range of opinion with regard to this question as outlined below (see also Larrabee, 2007):

Advise testees to do their best at the start. This seems uncontroversial and good practice, providing it is in accord with test administration requirements.
Warn testees that effort on tests will be assessed. Whilst this could negatively affect rapport and might create anxiety, it has some benefits to the testee irrespective of whether best effort is intended by them. If intending to deceive however, the testee may be prompted to look to ‘spot’ the test of effort and this could reduce the sensitivity of the test. However, it is ethical to be open and to indicate the purpose of assessment. This may include a general statement along the lines that the tester will assess how hard the testee is trying on the tests.

Raise the issue with the testee if they fail an effort test. The assessor may not have sufficient information to make sense of this at the time. This could give the testee an opportunity to explain away their poor test performance. In some contexts this may be useful (e.g. a general clinical context) and in others potentially less so (e.g. a litigation context).

Introduce a general query before administering an effort test? For example, ask how the testee is feeling and so on; this would give them an opportunity (without specific prompting) to say they are feeling unwell or cannot concentrate.

Choice of language in reports and feedback session. In a litigation context the testee will have access to the report and there may be possibilities of re-test where rapport would need to be re-established. Hence language must be clear, but where possible non-confrontational. In the clinical situation there is a need to explain the formulation and not simply the test results to the testee and to have a plan about how to take matters forward. It is important to recognise that in the situation of assessment for legal purposes the testee is not the client, the primary duty of the assessor is to the Court and the assessor should not offer feedback without the permission of the instructing solicitor.

Contextual and situational factors

Should effort testing be restricted to any specific context (e.g. only forensic, only clinical, etc.)? There seems to be no rational reason for doing so as sub-optimal effort could arise in a variety of contexts.

Should effort tests always be given in a forensic or litigation context? There is likely to be a motivation to deceive in a forensic or litigation context and effort should be assessed when possible. Indeed this may protect some testees from unfair criticism. If not assessed the assessor should be prepared to justify this exclusion (e.g. see ‘Exceptional cases’ below).

Should effort test be given to all cases seen by neuropsychologists (other than exceptional cases)? There is likely to be a desire to establish a good rapport with the testee and a fear of unjustifiably accusing them of poor effort. These factors, together with historical practice not to give tests of effort routinely, may explain the current situation whereby some psychologists do not use effort tests at all, even in a litigation context (see McCarter et al., 2009). However, there are reasons to argue that tests of effort should be given more often than not. There are exceptions with regard to specific client groups or severities of impairment (see below), in an acute setting (e.g. soon after an injury) and if a relatively brief screening assessment is the purpose of testing. However, in other contexts, recommendations with regard to treatment, rehabilitation, medication or for social support may be made on the basis of cognitive tests. There is a need to judge whether a person has given their best effort or these recommendations could disadvantage the person (e.g. encouraging them towards a disabled role when poor test scores do not accurately reflect the person’s ability).
An example would be someone who was thought to have a severe head injury and this is seemingly confirmed by poor performance on cognitive tests. If admitted to a brain injury rehabilitation unit, their belief in their disability can be inappropriately strengthened, if in fact they sustained a mild head injury and under performed on the cognitive tests.

**Threats to effort testing: A cautionary note**

Given the nature and purpose of effort tests, it is important that reasonable precautions are taken to safeguard their use in order to maintain their validity. It is recognised that there are potential threats to test security. In litigation contexts, coaching by solicitors represents one possible threat. Internet and other media exposure is another (Bauer & McCaffery, 2006). Psychologists have an ethical and professional obligation to maintain test security by restricting access to actual tests or test data to those qualified to use and interpret them. They should also strive to minimise unnecessary test exposure and avoid the release of sensitive data (e.g. detailed descriptions of effort tests and their underlying rationale) into the public arena. During test administration, reasonable precautions need to be taken to avoid unnecessary exposure of test manuals and record forms to individuals undergoing assessment.

There are a number of other ways in which the security of symptom validity assessment can be maintained. In report writing for example, psychologists could omit the names and descriptions of effort tests in the body of the report. Instead, a general statement along the lines of ‘effort testing failed to indicate non-credible performance’ may suffice. Specific details of these procedures and actual test scores could subsequently be released to qualified professionals on request. Consideration of this practice is especially important in situations where individuals may complete several neuropsychological evaluations and where coaching is a possibility. The prevalence of coaching in the UK is unknown but in North America, a recent survey conducted by the National Academy of Neuropsychology and Association of Trial Lawyers, suggested that 75 per cent of attorneys spend 25 to 60 minutes preparing clients for psychological evaluation by providing them with information about tests and how they might respond (Victor & Abeles, 2004).

**Are there exceptional cases?**

While effort tests are generally designed to minimise the risk of false positive classification errors, some studies suggest that they can be failed for reasons other than poor effort. Examples are learning disability, severe amnesia, moderate to severe dementia, severe sensory or physical disability, perceptual disorders, neurological, psychiatric or physical conditions where initiation or alertness is diminished or attention severely impaired (e.g. confusional or low awareness states or acute psychosis), (Dean et al., 2008; Batt et al., 2008; Browndyke et al., 2008). Studies of this kind emphasise the need for further investigation of the performance characteristics of effort tests and their interpretation. Clearly, practitioners also need constantly to take account of issues relating to mental capacity and informed consent.

---

Cultural and ethnic issues

Assessment of effort needs to take account of ethnic and cultural differences that might render tests invalid. In many ways these issues are similar to those described more generally in textbooks on cognitive assessment (see Strauss et al., 2006). Specifically with regard to effort there have been US studies, including on Spanish-speakers and Chinese (see Salazar et al., 2007).

Future research needs

- Tests of cognitive function which simply include embedded assessments of effort need further development.
- Further evidence on UK base rates of cognitive impairment and failure on effort tests in a range of clinical presentations and service settings is needed.
- A better understanding of the sensitivity and specificity of tests of effort is needed using UK populations (especially in non-forced-choice measures (see Nitch & Glassmire, 2007).
- UK studies on ethnic and cultural issues in assessment of effort are needed.


Appendix 1: Explanation of terms

- **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.

- **Symptom validity**: The accuracy or truthfulness of the testee’s presentation (e.g. behaviour, self-report, test performance).

- **Response bias**: Presentation, including test performance that results from an attempt to mislead the examiner, including through sub-optimal effort.

- **Dissimulation**: The falsification or misrepresentation of symptoms by over representation or under representation, with an intention to appear different from the ‘true’ state.

- **Malingering**: The intentional presentation of false or exaggerated symptoms motivated by external gain.
Appendix 2: Statistical Issues

Sensitivity, Specificity and Base Rates

The diagnostic validity of a test concerns the ability of that test to detect the presence or absence of a defined characteristic in the person assessed. For example, in the context of the present document, this might be whether the person is simulating or not simulating. Gauging diagnostic validity involves understanding the sensitivity and specificity of a test and knowing the base rate of the characteristic in the population being examined.

A discussion of sensitivity, specificity and base rates in the context of effort testing can be found in chapter 2 of Larrabee (2007) and chapter 10 of Boone (2007).

Sensitivity is the ability of a test to detect a characteristic in those who truly do possess that characteristic. For example, a test might pick out 90 per cent of people who are truly simulating. (The precise percentage will of course depend on the cut-off score chosen). Technically sensitivity is the ratio of true positives to true positives plus false negatives. Therefore, in the example given, the sensitivity of the test will be 90/(90+10)=0.9.

Specificity is the ability of a test to detect the absence of a characteristic in those who truly do not possess that characteristic. For example, the same test using the same cut-off might pick out 95 per cent of those who are truly not simulating. Technically, specificity is the ratio of true negatives to true negatives plus false positives. Therefore, in the example given the specificity is 95/(95+5)=0.95.

However, the precise interpretation of a test result depends on the base rate of the characteristic in the population being assessed, i.e. the percentage of that population that truly posses the characteristic.

If in the example given, clinician A sees a population in which 50 per cent are truly simulating. The result of 200 assessments on that test will be:

<table>
<thead>
<tr>
<th>Truly simulating (N=100)</th>
<th>Test says simulating</th>
<th>Test says not simulating</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>90</td>
<td>10</td>
</tr>
<tr>
<td>Truly not simulating (N=100)</td>
<td>5</td>
<td>95</td>
</tr>
</tbody>
</table>

In this scenario, if the test indicates simulating, the likelihood that the person is truly simulating is 90:5; i.e. a probability of 90/95=0.947.

However, suppose clinician B sees a population in which only 10 per cent are truly simulating. In this case the results of 200 assessments on the test will be:

<table>
<thead>
<tr>
<th>Truly simulating (N=20)</th>
<th>Test says simulating</th>
<th>Test says not simulating</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>18</td>
<td>2</td>
</tr>
<tr>
<td>Truly not simulating (N=180)</td>
<td>9</td>
<td>171</td>
</tr>
</tbody>
</table>
In this scenario, if the test indicates simulating, the likelihood that the person is truly simulating is 18:9, which is a probability of 18/27 = 0.667.

Hence, in this example, the lower the base rate of simulating, the less certain is the meaning of a positive result on the test.

**The Binomial Test:**

In multiple-choice testing, the testee is required to select one from among a set of given alternative answers. In such a situation, by guessing alone (chance), a proportion of the answers are likely to be correct. To help compensate for this in interpreting test results, it is useful to have a procedure for taking into account the effect of correct guesses. The Binomial Test is a statistical procedure useful for determining the probability of achieving a particular score or less. For a given score (i.e. number correct) the standard score \((Z)\) can be computed, and can be converted into a probability or percentage using tables of the normal distribution. The formula is \(Z = (X - NP) / \sqrt{NPQ}\), where \(Z\) is the standard score (a \(Z\) of 1.98 is significant at the .05 level, two tailed), \(X\) is the obtained score (i.e. the number of correct answers achieved by the client), \(N\) is the number of items, and \(P\) and \(Q\) are the proportion of correct and incorrect options (e.g. both \(P\) and \(Q\) will be 0.5 in a two alternative forced-choice situation, or 0.25 and 0.75 in a four alternative forced-choice situation). A correction factor is applied for small \(N\), i.e. \(Z = (|X - NP| - 0.5) / \sqrt{NPQ}\), where the expression \(|X - NP|\) means the modulus of \((X - NP)\), i.e. the value of \((X - NP)\) regardless of sign. Therefore if a client scored 1 out of 10 on a two-choice, forced-choice test the corrected \(Z\) would be \(|1 - 10 \times 0.5| - 0.5) / \sqrt{10 \times 0.5 \times 0.5} = 3.5 / 1.58 = 2.2.\) Hence the probability of obtaining a score of 1 out of 10 by chance is \(p < 0.05\). Therefore, in this example, it is highly likely that the score was not obtained by chance. The score is so poor that the possibility that the individual deliberately gave wrong answers is raised.
## Appendix 3: Commonly used purpose-designed tests of effort

<table>
<thead>
<tr>
<th>Measure and year of publication</th>
<th>Paradigm</th>
<th>Administration Mode</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test of Memory Malingering (TOMM) Tombaugh (1996)</td>
<td>Picture recognition (forced choice)</td>
<td>Manual or Computer (20 min.)</td>
<td>Relatively unaffected by age, low education, cultural factors, anxiety, depression, pain. Suitable for use with children (Donders, 2005) but high misclassification rates in adults with dementia (Teichner &amp; Wagner, 2004). Questions have been raised about sensitivity relative to other tests (e.g. Green, 2007b).</td>
</tr>
<tr>
<td>Word Memory Test (WMT) Green (2003)</td>
<td>Word recognition (forced choice)</td>
<td>Manual or Computer (20 min.)</td>
<td>Assesses effort and verbal memory. Performance not affected by education, IQ, neurological or psychiatric conditions, pain or age (suitable for use with children). Appears to easily identify simulators and those not putting forward maximal effort (Green et al., 2002).</td>
</tr>
<tr>
<td>Rey 15-Item Test (and recognition trial) Rey (1964)</td>
<td>Letter/Number/Symbol recall and recognition</td>
<td>Manual (5 min.)</td>
<td>Boone et al. (2002a) extended the original 15-Item test to include a recognition</td>
</tr>
<tr>
<td>Measure and year of publication</td>
<td>Paradigm</td>
<td>Administration Mode (Time)</td>
<td>Comment</td>
</tr>
<tr>
<td>---------------------------------</td>
<td>----------</td>
<td>--------------------------</td>
<td>---------</td>
</tr>
<tr>
<td>Dot Counting Test Rey (1941)</td>
<td>Processing of visual stimuli</td>
<td>Manual (10 min.)</td>
<td>Empirical studies are equivocal about sensitivity and specificity. Variable performance across clinical groups increases the risk of false positive errors. Use of combination effort index score (Boone et al., 2002b) improves sensitivity. Performance relatively unaffected by age and education but may be affected by psychosis and ethnicity (Larrabee, 2007).</td>
</tr>
<tr>
<td>Computerised Assessment of Response Bias (CARB) Conder, Allen &amp; Cox (1992)</td>
<td>Digit recognition (forced choice)</td>
<td>Computer (administration time variable depending on version used)</td>
<td>Performance is relatively unaffected by severity of head trauma and reasonably resistant to coaching efforts in simulated studies</td>
</tr>
<tr>
<td>Measure and year of publication</td>
<td>Paradigm</td>
<td>Administration Mode (Time)</td>
<td>Comment</td>
</tr>
<tr>
<td>---------------------------------</td>
<td>----------</td>
<td>-----------------------------</td>
<td>---------</td>
</tr>
<tr>
<td>Victoria Symptom Validity Test (VSVT) Slick et al. (1997)</td>
<td>Digit recognition (forced choice)</td>
<td>Computer (15 to 20 min.)</td>
<td>Easy to administer. Format promotes test-taker perception of varied level of difficulty. Loring, Lee and Meador (2005) found poorer VSVT scores in individuals over 40 years with low performance on cognitive tests. Failure rate in clinical groups has been found to be substantially lower than compensation-claiming samples (Larrabee, 2007).</td>
</tr>
<tr>
<td>Amsterdam Short-Term Memory Test (ASTM) Schmand &amp; Lindeboom (1998) – in collaboration with Merten &amp; Millis</td>
<td>Word recall and recognition</td>
<td>Computer (10 to 30 min.)</td>
<td>Merten et al. (2005) found good separation between individuals asked to try their best and those coached in cognitive exaggeration with reported sensitivity and specificity values of 100 per cent. They found the ASTM to be approximately equivalent to the</td>
</tr>
<tr>
<td>Measure and year of publication</td>
<td>Paradigm</td>
<td>Administration Mode (Time)</td>
<td>Comment</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>----------</td>
<td>---------------------------</td>
<td>---------</td>
</tr>
<tr>
<td>Medical Symptom Validity Test (MSVT) Green (2004)</td>
<td>Word recognition (forced choice) and free recall</td>
<td>Computer (15 to 20 min.)</td>
<td>MSVT in its ability to differentiate between simulators and those demonstrating good effort. Preliminary findings are encouraging and the test manual includes information to help identify dementia profiles in failed protocols. Interestingly, respondents tested in a foreign language can achieve near ceiling scores on the MSVT (Richman et al., 2006). Howe et al. (2007) provide data comparing diagnostic groups of individuals attending a memory disorders clinic.</td>
</tr>
</tbody>
</table>
Appendix 4: Iatrogenic factors

One of the processes in presentation and symptom complaint is for the individual to come to report symptoms that they are exposed to during a clinical or litigation process (iatrogenic symptoms); e.g. when reading expert or clinical reports, or as described to them by therapists or lawyers, or in educational programmes in rehabilitation units, or during their own investigations by reading books or searching the internet. An individual can come to believe that they have a severe brain injury and severe cognitive problems even when the brain injury itself was mild and their early history was of good recovery with only mild symptoms. It is important that experts and psychological services do not reinforce the disabled role or reinforce erroneous and maladaptive health beliefs.