INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

The Victoria Symptom Validity Test: Development of a New Clinical Measure of Response Bias

by

Daniel Joseph Slick

B.Sc., University of Alaska, 1989
M.Sc., University of Victoria, 1992

A Dissertation Submitted in Partial Fulfillment
of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

in the Department of Psychology

We accept this dissertation as conforming to the required standard:

Dr. Esther H. Strauss, Supervisor (Department of Psychology)

Dr. Frank J. Spellacy, Department Member (Department of Psychology)

Dr. Catherine A. Mateer, Department Member (Department of Psychology)

Dr. Max R. Uhlemann, Outside Member (Department of Psychological Foundations)

Dr. Grant L. Iverson, External Examiner (Department of Psychiatry, University of British Columbia)

Supervisor: Dr. Esther H. Strauss

Abstract

This dissertation describes the development of the Victoria Symptom Validity Test (VSVT). The VSVT was designed to assist in screening for non-optimal performance during neuropsychological evaluation due to malingering, psychiatric disturbance, or other environmental or dispositional factors. Specifically, the VSVT is a test of recognition memory that uses the forced-choice paradigm for detecting biased or random responding. Response latency is also recorded. Results from pilot and follow-up normative studies with experimental and clinical populations are presented. The VSVT was found to have excellent divergent and adequate convergent validity in samples of compensation-seeking and non-compensation-seeking patients. Classifications of experimental participants using below chance performance as a cutoff were consistent with the majority of previous studies in finding 100% specificity but poor sensitivity. A new system wherein a third category is added for *questionable* (at chance) performance showed greatly increased sensitivity, with no decrement in specificity. Although scores in the questionable range are not unequivocal indicators of malingering, findings suggest good utility for screening or corroborative purposes. A Bayesian diagnostic probability matrix that takes base-rates into account was also provided as a more flexible alternative to absolute cutoff scores. Response latency as a measure of symptom validity was shown to have adequate sensitivity for screening, but less acceptable discriminant validity and lower specificity. Limitations of experimental results and clinical applications of symptom validity tests are discussed.

Examiners:

Dr. E.H. Strauss, Supervisor (Department of Psychology)

Dr. P.J. Spellacy, Department Member (Department of Psychology)

Dr. C.A. Mateer, Department Member (Department of Psychology)

Dr. M.R. Uhlemann, Outside Member (Department of Psychological Foundations)

Dr. G.L. Iverson, External Examiner (Department of Psychiatry, University of British Columbia)

Table of Contents

List of Tables

List of Figures

Acknowledgments

The author would like to thank the following people for their valuable contributions to this project:

Dr. Esther Strauss, Dr. Merrill Hiscock, Dr. Frank Spellacy, Dr. David Berry, Dr. Grant Iverson, Dr. Max Uhlemann, Dr. Catherine Mateer, Glen Slick, Dr. Galia Artzy, Dr. Michael Hunter, Dr. Paul Craig, Dr. Travis White, Tom Allen, Dr. Adele Hern, Dr. Deborah Allison, Grace Hopp, the staff at the Cornett Computing Centre, and all research participants.

Dedication

To Lee and Jean Slick, and to Elisabeth Sherman.

Introduction

The Problem of Malingering

Within the last twenty years, research on techniques and instruments for detecting feigned or exaggerated cognitive and psychiatric dysfunction has grown dramatically. The burgeoning interest in assessing symptom validity derives in large part from the substantial growth in both number and cost of claims within systems of limited resources. Under such conditions, the costs of malingering are not insignificant. For example, the Insurance Corporation of British Columbia reported that the cost of claims arising from car accidents in 1994 was $1,960,898,000 (ICBC, 1994). If just 0.5% of that amount was accounted for by faked impairment, nearly ten million dollars was lost to fraud. Such considerations have led to steadily increasing demands on neuropsychologists to provide evidence that deficits observed in assessments of compensation-seeking patients are not exaggerated or faked

A Closer Look at Malingering

According to the American Psychiatric Association (APA), malingering is "the intentional production of false or grossly exaggerated physical or psychological symptoms, motivated by external incentives such as avoiding military duty, avoiding work, obtaining financial compensation, evading criminal prosecution, or obtaining drugs" (1994, p. 683). Malingering becomes a diagnostic consideration whenever (1) readily identifiable and commonly recognized incentives for exaggeration or fabrication of dysfunction exist, (2) subjective complaints or test results are not consistent with neurological or functional status, (3) patient cooperation is questionable, or (4) a history of sociopathic behavior is obtained. However, these conditions alone do not justify a diagnosis of malingering. For making a diagnosis, the clinician must have a full understanding of what malingering is, and whether a patient fits that category beyond a reasonable doubt. To that end, the following sections expand on the necessary criteria for a diagnosis of malingering.

Conscious and Deliberate Intent

To malinger, a person must be aware that she or he is fabricating or exaggerating a deficit *and* have the capacity to understand that her or his actions are in some way wrong or contravene established social mores or laws. Patients who are unable to form or understand such concepts are not capable of malingering. Diagnoses of malingering are also excluded in cases were patients are unable to exercise volitional control over their behavior. Therefore, young children or individuals with certain neurologic or psychiatric disturbance

are not usually considered capable of malingering. Conversion disorder, for example, shares with malingering a presentation of neurological "deficits" that are the product of psychological processes. Conversion disorder may even present in situations where material incentives exist for disability. However, conversion disorder is differentiated form malingering by a "lack of conscious intent in the production of the symptoms" (APA, 1994, p. 457).

External Incentives

Malingering can only occur when patient behavior can be directed toward either escape from formal obligation or some type of material gain commonly accepted as having value. The conscious faking of a deficit in order to meet a pathological need to play the sick role (e.g., Factitious Disorder) is not normally considered malingering. Likewise, people may consciously fabricate or exaggerate impairment in order to receive attention, release from informal obligation, or otherwise manipulate their environment. This type of behavior, although usually viewed as maladaptive, is not normally considered malingering.

Other Possibilities

Malingering must also be distinguished from poor or inconsistent effort, and defensive, hostile, or oppositional approaches to test taking that are motivated by factors other than conscious attempts to obtain readily identifiable and commonly accepted external incentives. Finally, although various legal and nosological systems encourage black and white diagnostic decisions, consideration should also be given to the coexistence of fabricated and real deficits.

Alternative Approaches

Like all things in psychology, the definition of malingering and its clinical value as a diagnostic entity are not agreed upon by all. Taking an extreme stand against malingering as a diagnostic entity, Erickson (Pankratz & Erickson 1990, p. 381) argues that "the diagnosis of malingering is a weak diagnosis of exclusion that serves to justify the denial of treatment and benefits;" and that "were it not for some medicolegal expectations, we could do without the diagnosis entirely." At the other end of the spectrum are those like Pankratz (Pankratz & Erickson 1990), who proposes loosening the necessary criteria for a diagnosis of malingering. He suggests that the determination of intent – usually considered crucial – is the least important aspect of diagnosing malingering, arguing instead that "intentions,

awareness, conscious purposes, and psychodynamics should not be the main focus of the diagnostic process" (p. 386). Pankratz asserts that conscious intent and volition can not be reliably assessed, and therefore the diagnosis of malingering can and should be made with little attention to a patient's internal states, observing that: "we use the labels of arsonist and shoplifter without regard to the actor's control or awareness; we should be able to use psychiatric labels similarly" (p. 386). Rather than abolish malingering as a diagnostic entity or loosen its boundaries to make it more inclusive, Rogers (1983) proposes that we instead reframe our attitudes toward malingering: "considering what is at stake for many patients in an adversarial evaluation, it would be interesting to reconceptualize 'malingering' in terms of coping strategies, good judgment, and survival. This is not an argument toward discarding malingering as a viable concept; rather it is against the somewhat oversimplified and moralistic assumption that 'you should never lie.'"

There is something to be said for each of these divergent opinions on malingering. Although most neuropsychologists would probably reject Erickson's (1990) call for the abolition of malingering as a diagnostic entity, he correctly notes that such diagnoses may have drastic consequences for patients, including the denial or termination of treatment or support. This is especially problematic in cases where exaggerated or fabricated deficits coexist with real impairments or disorders that may be amenable to treatment.

Pankratz (1990) makes an important point by calling for an emphasis on objective indicators of malingering. However, his proposal that the criteria for malingering should be relaxed does not represent an acceptable solution to the difficulties of diagnosis. It is true that making judgments about internal states is difficult, but it is not impossible if one is willing to accept some uncertainty. Indeed, making judgments about a client's internal states is essential to much of clinical practice. This reality is nowhere more apparent than in the DSM-IV. Many diagnostic entities in the DSM-IV *require* some judgment about internal states that are supportable by observational evidence (e.g., presence or absence of behavior suggestive of hallucinations). Consider one of Pankratz's examples: compare the *behavior* of shoplifting with the *diagnosis* of kleptomania. The diagnosis requires not only behavior that can be directly observed (stealing), but also requires a judgment about internal states (overwhelming impulse to steal). A patient may present in a forensic setting with features of kleptomania that may be real or malingered, and the diagnostic decision may turn on an evaluation of intent, motivation, and ability to conform behavior. Most

clinicians would agree that kleptomania and malingering are two different conditions that require significantly different clinical responses, but Pankratz's recommendations for dropping conscious intent from the diagnostic criteria for malingering blurs this critical difference to the point of abolishing a useful clinical distinction.

Rogers (1983) provides a useful perspective on malingering. Clinicians need to think carefully about how the current medicolegal system encourages malingering, about how it can be changed to reduce what may be in some senses an adaptive response to the system, and about how people may be assisted to find more socially acceptable and adaptive ways of managing their environment.

A Brief History of Methods for Detecting Poor Effort and Biased Responding

Clinically, malingering is detected like any other "disorder;" by a combination of thorough record review, interviewing with the patient and third parties, testing, and behavioral observation. To conclude that a person is malingering, one must rule out the alternatives. To do this, test data are essential. However, psychometric methods for detecting malingering are still in a relatively early stage of development. One of the most pressing questions in malingering research today is whether it is possible to tease apart legitimate from exaggerated impairment in cases where both may be present.

Two approaches are available to researchers and clinicians who need measures of biased responding. In the first approach, validity scales or indices are developed for use with conventional neuropsychological. The second approach is to develop new instruments designed specifically for detecting dissimulation.

Validity Scales and Indices for the MMPI and MMPI-2

The best example of the first approach is probably embodied by the MMPI (Hathaway & McKinley, 1983) and its successor, the MMPI-2 (Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989). These instruments are perhaps the most widely used tests in clinical psychology. The MMPI was designed from the outset to include a set of validity scales; a variety of official and unofficial validity scales were subsequently developed through independent research.

The F Scale

As the original fake-bad scale developed by the authors of the MMPI, the F (infrequency) scale is perhaps the oldest, most well researched, and most widely used measure of

dissimulation in clinical psychology. The F scale, like the entire MMPI-2, is not without flaws (see, Helmes and Reddon, 1993 for a recent review). Two criticisms are especially troublesome: (1) the F scale was developed by identifying items that were infrequently responded to in the keyed direction within the original MMPI normative sample; with the exception of dropping four items, the F scale remains unchanged in the MMPI-2 and T scores are again derived from the normative sample. It is therefore not known whether the items that make up the F scale are infrequently endorsed within various clinical populations. (2) the F scale shares items with many clinical scales, reducing its specificity as a dissimulation scale. This confound is evident in the MMPI-2 manual (Butcher, Dahlstrom, & Graham, 1989), where legitimate psychopathology is listed as a possible interpretation of F scale T-Score elevations between 56-90, while scores above this range are considered invalid. The dual nature of the F scale suggests a cautious approach to the application of cut-scores. Indeed, the specificity of commonly suggested cut-scores for the F scale has been questioned, with legitimate cases of psychopathology presented as a viable alternative explanation for extreme F scale elevations (Anthony, 1971; Butcher & Williams, 1992; Greene, 1991; Grillo, Brown, Hilsabeck, Price, & Lees-Haley, 1994). Nonetheless, extreme F scale elevations, when legitimate, are invariably associated with severe psychopathology of a kind uncommon among personal injury claimants seen for neuropsychological assessment. Therefore, the F scale remains a popular method of comparing reported to observed psychopathology. Following a meta-analysis of 28 studies, Berry, Baer and Harris (1991) found that the F scale was as good or better than other MMPI validity scales for detecting malingering and was associated with a substantial effect size ($d = 2.34$). A large effect size ($d = 3.03$) was also reported in a more recent meta-analysis of studies with MMPI-2 (Rogers, Sewell, & Salekin, 1995). Thus, studies with the revised F scale in the MMPI-2 continue to support the contention that highly elevated F scale scores require endorsement of a wider range of problems than are typically associated with specific psychiatric illnesses, and suggest instead a deviant approach to the test.

Unfortunately, most of the empirical support for the revised F scale as a measure of dissimulation, including the above cited meta-analyses, is based on statistical differences between groups rather than clinical utility as measured by sensitivity, specificity, and predictive power of specific cut-scores. These findings encourage the use of the revised F scale, but leave the clinician with little practical information. Furthermore, considerable

variability is found among the studies that have suggested specific F scale cut-scores (see Rogers, Sewell, & Salekin, 1994 for a review). Therefore, caution is warranted when applying the revised F scale as an indicator of malingering. To be fair, it must be recognized that a considerable amount of the variability in suggested cutoff scores for the revised F scale is related to between-study population differences (i.e., as manifested by varying scale score base-rates). Thus, in order to control false-positives, a higher cutoff is required for discriminating real from faked schizophrenia (e.g., T > 120: Rogers, Bagby, & Chakraborty, 1993) than for discriminating real from malingered personal injury claims (e.g., T > 65: Lees-Haley, 1991). For applications with compensation-seeking neuropsychological patients, there are now several studies providing information on the utility of the MMPI-2, and the revised F scale in particular (Berry et al., 1995; Greiffenstein, Gola, and Baker, 1995; Lamb, Berry, Wetter, & Baer, 1994; Lees-Haley, 1991; Wetter et al., 1992). However, only one of these studies, using a small sample of patients (Lees-Haley, 1991), provides suggested cutoff scores with associated sensitivity and specificity. Using larger samples of analog and clinical samples, Greiffenstein et al. (1995) found that the F scale scores cold not reliably discriminate real from feigned traumatic brain injury. In summary, research provides qualified support the use of the revised F scale as a measure of validity of reported symptoms. However, it has yet to be demonstrated that cutoff scores can provide adequate sensitivity and specificity for distinguishing between psychopathology and malingering in medicolegal neuropsychological assessments.

The F-K Index

Another popular dissimulation scale obtained from the MMPI and MMPI-2 is the F minus K dissimulation index (F-K) first proposed by Gough in 1950. The index is derived by subtracting the raw K score from the raw F score. Though not an original validity scale, the common use and acceptance of F-K is acknowledged by its appearance on outputs of standard computer scoring systems for the MMPI-2. Because it is based in part on the F scale, the F-K index is subject to most of the criticisms that apply to the F scale. Primary among these criticisms are the confounding of validity assessment with measurement of psychopathology. Scores that are rarely obtained by healthy individuals are common among psychiatric patients. For example, F-K scores greater than zero are found in less than 5% of the normative sample, but were present in over 50% of a sample of 456 psychiatric inpatients (Rothke, Dahlstrom, Greene, Arredondo, & Mann, 1994). Among the

latter sample, only F-K scores in excess of 23 for men and 18 for women were rare (occurring in 5%or less of the sample). These scores greatly exceed recommended cutoffs for invalid profiles (Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989). Nonetheless, many studies support the use of the F-K index for evaluating profile validity. Berry, Baer and Harris's (1991) found an average effect size of (d = 1.89) in their meta-analysis of studies of F-K as a measure of profile validity. This finding was replicated by a meta-analysis on studies with the revised F-K from the MMPI-2 (d = 2.46: Roger, Sewell, & Salekin, 1995). As with the revised F scale, however, most studies supporting the use of F-K have based their conclusions on statistical rather findings rather than demonstrations of sensitivity, specificity, and predictive power. When suggested cutoff scores are considered, considerable variability is found (Rogers, Sewell, & Salekin, 1994). As with the F scale, this finding is due in part to between-study differences in populations and associated scale score base rates. Only one study with a small sample has provided suggested cutoff scores with associated sensitivity and specificity for purposes of evaluating compensation-seeking individuals, (F-K > -4: Lees-Haley). A larger study found that the F-K index was ineffective for discriminating real from malingered traumatic brain injury (Greiffenstein et al., 1995). In sum, the clinical utility of the F-K index as an indicator of patient veracity for medicolegal neuropsychological assessments has yet to be adequately demonstrated.

The Obvious minus Subtle Index

The Wiener and Harmon Obvious and Subtle subscales (Wiener, 1948) have a history similar to that of the F-K index. They were developed by bifurcating 5 of the clinical scales into subscales based on rational analyses of items. Essentially, the Obvious subscales contain items which are face-valid, while Subtle subscales do not. Because of this dichotomy, the Obvious and Subtle subscales quickly became popular for assessing profile validity as feigned or exaggerated psychopathology was assumed to elevate Obvious but not Subtle scores. A validity index was subsequently developed by subtracting the total of the Subtle T scores from the total of the Obvious T scores. Like F-K, the continued popularity of the Obvious and Subtle subscales amongst clinicians is reflected by their appearance on the output of standard scoring programs. One of the purported strengths of the Obvious-Subtle (O-S) index as a measure of symptom exaggeration is that it provides an assessment of validity that is independent of the presence of psychopathology (i.e.,

exaggeration manifested by Obvious-Subtle discrepancies might be removed from scale elevations to reveal "true" psychopathology). Despite the apparent logic of its construction, O-S has not demonstrated consistent efficacy as a validity index and is therefore among the most controversial of MMPI validity scales (Graham, 1992). Although some recent studies continue to support the use of O-S for identifying over-reporting of symptoms (Dannenbaum & Lanyon, 1993; Greene, 1991; Lees-Haley, 1991b; Lees-Haley, 1992; Lees-Haley & Fox, 1990), other studies have cast serious doubt on the utility of O-S based on high correlation with legitimate psychiatric disorder (Schretlen, 1990), poor reliability of Subtle subscales (Boone, 1994), and poor prediction of external criterion (Weed, Ben-Porath, & Butcher, 1990). In fact, Graham (1992) suggests that most subtle items would not have survived an adequate cross validation of the original MMPI, and therefore amount to little more than noise which attenuates the reliability and validity of their respective clinical scales. Berry, Baer, and Harris, 1991 also found a small effect size for O-S in their meta-analysis, and recommend caution in its use. However, a meta-analysis on studies with the revised O-S from the MMPI-2; Roger, Sewell, & Salekin (1995) reporting a large *d* value (3.09). Like the revised F scale and F-K index, considerable variability exists among suggested cutoff scores (Rogers, Sewell, & Salekin, 1994), due in part to population differences. In addition, a recent study found that the O-S index was ineffective for discriminating real from malingered traumatic brain injury (Greiffenstein et al., 1995). In sum, of the MMPI-2 validity scales and indices, O-S has some of the weakest empirical support, particularly for use in neuropsychological assessment of compensation-seeking patients.

The Lees-Haley Fake-Bad Scale

In 1991, Lees-Haley, English, and Glenn published a newly developed dissimulation scale for the MMPI-2. The *Fake-Bad Scale* (FBS) was derived from an item analysis of protocols from malingering and non-malingering personal injury claimants. Interestingly, the FBS contains a mix of items from scales such as L and F that are usually used as indicators of either fake-good or fake-bad response sets. Lees-Haley et al. explain this apparent paradox by noting that unlike situations where persons may attempt to dissimulate extremes of mental health (e.g., child custody vs. forensic assessments), malingering in personal injury assessments is more likely to present as a mix of fake-good and fake-bad. That is, personal injury claimants attempting to feign distress and dysfunction on the MMPI-2 are likely to endorse items that present themselves as persons who: (1) are and

have always been exceedingly virtuous, (2) have always been psychological healthy prior to the injury on which the claim is based, and (3) are now disabled by the psychological and/or physical consequences of the injury. Encouraging initial results were obtained when the FBS was evaluated with small samples of credible and malingering personal injury claimants, and simulated malingering participants. A second, larger study (Lees-Haley, 1992) again found the FBS to have good sensitivity (.74) and specificity (.94), after the cut-off suggested from the first study was to be adjusted upward to reduce false-positives. In contrast, Rogers, Sewell, and Ustad (1995) found that the FBS scale was ineffective for distinguishing between profiles produced by psychiatric outpatients patients under standard and fake bad conditions (sensitivity=.49; specificity=.81). However, the findings of Rogers et al. may not be directly relevant to the use of the FBS with litigating patients. Whereas the FBS was designed to detect exaggeration among personal injury claimants, the patients in the Rogers et al. study (all with verified psychiatric disturbance) were instructed to produce profiles consistent with an immediate need for hospitalization. Clearly the FBS is a potentially useful addition to the MMPI-2, although adequate cross-validation will be necessary before it should be used routinely.

The F(p) Scale

One of the newest validity scales to be developed for the MMPI-2 was created in response to the limitations of the F scale. Arbisi and Ben-Porath (1993, 1995) developed the *Infrequency-Psychopathology,* or F(p) scale by identifying 27 items that had rates of endorsement of ten percent or less both within the normative sample and within an additional sample of 706 acute and chronic care psychiatric inpatients. Given a non-random response set, high scores on the F(p) scale result from endorsement of a wider range of symptoms than might be expected from any normal or psychiatric patient, and strongly suggest the possibility of exaggerated or faked distress. This assumption has recently been supported by two studies. Employing a population of chronic psychiatric outpatients that were administered the MMPI-2 under honest and feigned psychopathology conditions, Rogers, Sewell, & Ustad (1995) found a sensitivity of .88 when specificity was set to a minimum of .80. Arbisi and Ben-Porath (1995) demonstrated significant incremental validity when F(p) was compared to F in ability to discriminate healthy adults who were instructed to fake psychopathology from legitimate psychiatric inpatients. To date, no investigators have examined the utility of the F(p) scale in the detection of

malingered head injury in neuropsychological assessment, but its sound psychometric development, and initial cross validation findings suggest that it may be a useful tool.

Neuropsychological Application of MMPI-2 Validity Scales

Although widely used in neuropsychological evaluations to assess the extent and validity of reported psychiatric symptoms, few studies have assessed the sensitivity and specificity of MMPI/MMPI-2 validity scales for detecting malingered reporting of symptoms associated with closed head injury (CHI). Heaton, Smith, Lehman, and Vogt (1978) found that normal participants attempting to simulate CHI produce greater elevations than real CHI patients on the F scale. More recently, Lamb, Berry, Wetter, and Baer (1994) carried out an analog study with normal participants attempting to feign CHI on the MMPI-2 with the purpose of evaluating the effects of coaching. They also found that naïve feigning participants produced elevated F scales, and that relative to this group, feigning participants who received information on the typical sequelae of CHI tended to produce higher elevations on the F scale, while the opposite trend was observed among those participants who received information about the nature of the validity scales. Feigning participants who received both types of information typically produced valid profiles. Unlike the above studies which focused on group differences in elevations, Greiffenstein, Gola, and Baker (1995), directly evaluated the classification accuracy of F, F-K , and O-S. They found that none of these scales could adequately differentiate patients complaining of persistent post-concussive symptoms who were identified as probable malingerers from those who were identified as probable non-malingerers. While the results of these studies indicate that feigning participants tend to perform differently from controls and legitimate patients, and that coaching significantly affects feigning performance, it has yet to be demonstrated that malingered CHI symptoms can be accurately detected with the MMPI-2.

Validity Indices for Standard Neuropsychological Tests

Researchers have for years attempted to empirically derive malingering cutoff scores or "response profiles" for conventional (standardized, performance-based) neuropsychological instruments. If successful, this approach has the advantage of greatly increasing the efficiency of assessment because specialized tests of malingering may then be dispensed with. However, developing such indices can be very difficult. Conventional measures of cognitive function are typically designed and normed with the assumption that those being evaluated will perform to the best of their ability. This approach results in tests which may

be useful for distinguishing between "normal" and "impaired" performance, but not between real and fabricated deficits. For example, the utility of "malingering" cutoff scores for conventional neuropsychological tests is often limited by floor effects. That is, the easiest items are hard enough that many patients with legitimate deficits can obtain scores at floor or near floor level. Because of this, it is difficult to develop cutoffs that adequately balance sensitivity and specificity. The only exceptions to this limitation are conventional neuropsychological tests of recognition memory that use a forced-choice format. As will be described in a later section, cutoff scores with adequate sensitivity and specificity can be derived for detecting malingering with these instruments.

A more sophisticated approach has been exemplified by the development of malingering indices based on multiple measures from the same or different tests. This approach – often referred to as the *pattern of performance hypothesis* – is probably the most effective way to detect malingering with conventional neuropsychological measures. For example potentially useful malingering indices have been developed by evaluating serial position effects in list learning and other memory tests (Bernard, 1991; Russell, Spector, & Kelly, 1993), comparing recall to recognition (Beetar & Williams, 1994; Bernard, 1990 & 1991; Binder, Villanueva, Howieson, & Moore, 1992; Brandt, 1988; Knight & Meyers, 1995), comparing indices of attention to indices of memory (Mittenburg, Azrin, Millsaps, and Heilbronner, 1993), comparing performance on easy and hard items (Baker, Hanley, Jackson, Kimmance, & Slade, 1993; Fredrick & Foster, 1991; Tenhula & Sweet, 1996), and by combining multiple sources of data in statistical procedures such as discriminant function analysis (Bernard, Houston, Natoli, 1993; Bernard, McGrath, & Houston, 1996; Fredrick & Foster, 1991; Fredrick, Sarfaty, Johnston, & Powel, 1994; Hayward, Hall, Hunt, & Zubrick, 1987).

A slightly different, but equally promising approach is to develop diagnostic checklists for malingering. These checklists of diagnostic criteria are similar to, but more precise and clinically useful than those found in *Diagnostic and Statistical Manual of Mental Disorders, 4th Edition* (APA, 1994). Rogers first proposed such a multiple criterion based system in 1990. His classification model incorporates multiple sources of data from different domains, including self report, tests-scores, behavior during assessment, and collateral information. A minimum number of criteria from each domain (e.g., endorsement of an unusually high number of rare symptoms, contradictory collateral information, and evidence from

standardized tests) are required for a patient to be classified as a malingerer. In addition, contraindications, such as evidence of factitious or hysterical disorders must be absent for a malingering classification. More recently, Greiffenstein, Baker, and Gola (1994) described a very similar index of 'overt' malingering developed for use with litigating post-concussive patients. In two studies, (Greiffenstein et al., 1994; Greiffenstein, Gola, & Baker, 1995) they demonstrated clinically significant associations between classifications made by their index, and scores on malingering measures, including the PDRT. These results demonstrate not only the value of Greiffenstein et al.'s heuristic, but also the strong link between performance on some malingering tests and patient behavior outside the testing room. Undoubtedly, diagnostic systems such as those proposed by Greiffenstein et al. and Rogers will began to see increasing clinical use, as their more precise and standardized nosology provides a considerable advance over the more simplistic systems currently in common use.

Overall, attempts to develop malingering indices for conventional neuropsychological tests have met with mixed success. Many of the recent studies using performance pattern indices have shown considerable promise, but other studies report that conventional neuropsychological tests and test batteries may be ineffective in distinguishing malingered from legitimate impairment (Bernard, 1990, Faust & Guilmette, 1990; Faust, Hart, & Guilmette, 1988; Faust, Hart, Guilmette, & Arkes, 1988; Heaton, Smith, Lehman, & Vogt, 1978). Of those studies that have demonstrated potential utility for particular malingering indices, most remain unreplicated or cross validated.

Tests Designed Specifically to Detect Dissimulation

Perhaps the classic example of a test designed for detecting dissimulation is Rey's 15-Item Visual Memory Test (R-15), which was introduced in 1964, and remains popular despite decidedly mixed opinion about its utility. The R-15 is a brief and easy to administer immediate free-recall test in which "the patient need only remember three or four ideas to recall most of the items" (Lezak, 1995, p. 802). Unfortunately, the obvious ease and simplicity of Rey's Test reduces its face validity as a legitimate memory measure, potentially tipping off its true nature, especially with more sophisticated patients. In this regard, the R-15 may be less effective at detecting dissimulation than tests that are designed to appear difficult when they are in fact quite easy (Binder et al., 1993). Variable results have been reported from studies using Rey's Test, and some differences exist in the literature regarding clinically useful cut-scores for indicating the likelihood of malingering or other motivational

problems (Bernard & Fowler, 1990; Davidson, Suffield, Orenczuk, Nantau, & Mandel, 1991; Goldberg & Miller, 1986; Guilmette, Hart, Giuliano, & Leininger, 1994; Lee, Loring, & Martin, 1992; Millis & Kler, 1995; Schretlen, Brandt, Krafft, & Van Gorp, 1991). A variety of other methods for detecting poor effort and malingering have been developed since the initial publication of the R-15 (See Lezak, 1995; Franzen & Iverson, 1994; and Nies & Sweet, 1994 for recent reviews); by far the most popular type have been variants of the symptom validity test.

Symptom Validity Testing

The term *symptom validity testing* (SVT) was first used by Lezak in 1976. Since then it has become the generally accepted designation for an increasingly popular method of screening for poor effort or dissimulated deficits (although such tests are also called *measures of response bias* in some quarters). Essentially, SVT is a probabilistic analysis of patient performance on forced-choice tests of sensory or cognitive function. Most symptom validity tests use a two-choice response format, as this arrangement requires the least number of trials to reliably detect dissimulation. In two-choice recognition tests, for example, the probability of responding correctly on all items by chance alone (i.e., guessing) is 50%. Overall performance should therefore approximate 50% correct in the most severe cases of memory impairment (i.e., given an adequate number of trials, overall scores are expected to be at chance level when a patient cannot recognize the target and has to guess on all items). Assuming random responses to a given number of two-choice items, confidence intervals for chance-level performance can be calculated (see Appendices I & II). Scores within the confidence interval around chance level performance therefore reflect either severe impairment or, possibly, exaggeration of deficits. High or low scores that are outside a large confidence interval around chance are highly unlikely by chance alone, and are assumed to be the product of *purposeful* selection of correct or incorrect answers (in either case depending on intact memory), with the latter being indicative of exaggerated or faked memory deficits.

The Early History of Symptom Validity Testing

The symptom validity technique was originally developed by physicians as a method of screening for conversion disorder. Grosz and Zimmerman (1965) were the first to report on the use of symptom validity testing, employing a three-choice procedure for detecting

hysterical blindness. Other reports followed, as the technique was modified for detecting conversion disorder manifesting as visual field deficits (Miller, 1968; Theoder & Mandelcorn, 1973), deafness (Pankratz, Fausti, & Peed, 1975), and somasthesia (Miller, 1986; Pankratz, 1979; Pankratz, Binder, & Wilcox, 1986). Symptom validity really started to become popular in neuropsychology after the publication in 1983 of a report by Pankratz on an adaptation of the technique for testing questionable memory complaints. It is instructive to note that the earliest proponents of symptom validity testing cautioned that the technique, although useful for detecting non-organic influences on performance, was incapable of determining whether patients were malingering or hysterical (Grosz and Zimmerman, 1965).

The Sensitivity Problem

Following the initial popularization of the SVT technique by Pankratz and others in the early 80's, increasingly refined variations of the procedure have appeared. These instruments are of two types: (1) adaptations of conventional instruments (Fredrick, Hilliard, & Foster, 1991; Millis, 1992, Iverson, 1993; Iverson & Franzen, 1994; Iverson & Franzen, 1996), and (2) entirely new tests (e.g., Binder, 1989; Hiscock & Hiscock, 1989; Iverson, Franzen, & McCracken, 1991; Slick, Hopp, & Strauss, 1992). Aside from the need to standardize symptom validity tests for clinical use, investigators have also been motivated by the need to increase the relatively poor sensitivity of these measures. The earliest symptom validity tests utilized very simple stimuli and procedures (e.g., presenting a finger to one visual field or the other) that were generally found to be sufficiently sensitive to detect functional sensory deficits. When the SVT technique was subsequently adapted for detecting dissimulated cognitive deficits, very simple stimuli and procedures were again employed. For example, Pankratz (1983), used two different coloured lights as the stimuli, with the Symbol Digit Modalities Test as a distracter during 15 second retention intervals. Although Pankratz's case study demonstrated the potential usefulness of using SVT for detecting biased responding in neuropsychological assessment, malingering is likely associated with more sophisticated approaches to test taking than conversion disorder. Obviously easy tests are less likely to fool malingerers who are cautious or neuropsychologically sophisticated. These individuals are unlikely to miss enough items to incriminate themselves. As this fact became apparent to investigators, efforts were made to produce more sensitive SVT's.

Ideally, tests of symptom validity should be maximally sensitive to poor effort or feigned deficits and minimally sensitive to real deficits. To achieve that goal, such tests need to appear as difficult as possible while in fact maintaining a trivial level of difficulty for the majority of non-malingering patients. This provides the widest separation of scores between the ceiling level performance of non-malingering patients and the exaggerated deficits of malingerers who have been fooled by the apparent difficulty of the test. Recent developments of the forced-choice paradigm have attempted to increase apparent difficulty by employing more complex stimuli, increasingly difficult distracter tasks, increasing retention intervals, and potentially deceptive instructions that stress the difficulty of the task (e.g., Binder, 1990; Hiscock & Hiscock, 1989; Iverson, Franzen, & McCracken, 1991, 1994). For example, Bickart, Meyer, & Connell (1991) explored the effects of manipulating apparent SVT difficulty on rates of detecting malingering. Although the actual probability of choosing correctly remained the same for items on "easy" and "difficult" versions of the test, subjects instructed to feign brain damage generally performed significantly worse when the test was apparently harder. As a result, the proportion of obvious malingering (i.e., detection hit-rate) among the subjects was higher when the difficult as opposed to easy version of the test was employed. Thus, when feigning subjects believe a test to be difficult, they are more likely to overplay their "impairments." Care must be taken, however, in designing more face-valid SVT's, as increases in real difficulty may lead to floor effects which can decrease the specificity of any normative based cutoff scores developed for such tests (Binder, 1993; Iverson et al., 1991, 1994; Wiggens & Brandt, 1988). Nevertheless, increases in perceived or actual difficulty are associated with detecting higher proportions of malingerers by below-chance performance (Bickart et al., 1991; Binder, 1990, 1992, 1993a, 1993b; Prigatano & Amin, 1993; Slick et al., 1994). Unfortunately, the sensitivity of symptom validity tests still has considerable room for improvement, as substantial proportions of malingerers in simulation studies do not obtain below-chance scores, even newer SVT's with increased face validity (Beetar & Williams, 1995; Binder & Willis, 1991; Brandt, Rubinsky, and Lassesn, 1985; Fredrick & Foster, 1991; Guilmette, Har, & Giuliano, 1993; Iverson & Franzen, 1996; Iverson, Franzen, & McCracken, 1991, 1994; Prigatano & Amin, 1993; Slick et al., 1994; Wiggins & Brandt, 1988). Clearly, more potent approaches to detecting malingering need to be evaluated, including the derivation of additional, norm-

supported scoring systems to augment the standard probabilistic scoring systems for symptom validity tests.

Development of the Victoria Symptom Validity Test

The Victoria Symptom Validity Test (VSVT: Slick, Hopp, & Strauss, 1992, 1996) is a substantial refinement of the Hiscock Digit-Memory Test (HDMT: Hiscock & Hiscock 1989). The HDMT was developed as a less obvious alternative to simpler measures of response bias, such as Pankratz's original symptom validity test and the Rey 15-Item Test. In the HDMT, a five-digit number is presented on a card for a five second-study period, followed after a brief delay by another card, containing the correct choice and a foil. The correct answers can always be discriminated from foils by recognizing the first or last digit. In order to increase the face validity of the test, the period between study and recognition is overtly increased to alter perceived item difficulty. The retention interval begins at 5 seconds and is overtly increased to 10, and then 15 seconds. Malingering subjects are cued to perform poorly by being told that the test is difficult for those with memory problems, and that the level of difficulty increases with increases in retention interval. Hiscock and Hiscock (1989) assume that the increase in retention interval does not increase the actual difficulty of the test. A significant decline in scores across retention interval is therefore considered a suggestive of biased responding.

Although more face valid as a "test of memory" than the Rey's 15 Item Test or earlier SVT's, the HDMT is still an obviously trivial task–healthy subjects as young as five years old have obtained perfect scores (Hiscock, Branham, & Hiscock, 1993). Thus, despite attempts to increase face validity, the true purpose of the HDMT is likely to be apparent to many malingerers. Another shortcoming of the HDMT (which is common to the PDRT as well) is the 30-40 minute administration time which makes this test quite onerous for patients, who find it extremely tedious, and clinicians, who often have to drop other tests to fit it in to their battery.

In light of these problems with the HDMT, two major modifications were made to the basic design, producing the Victoria Revision (Slick, Hopp, & Strauss, 1992). First, the number of items was reduced from 72 to 48, effectively halving administration time while still providing enough items for reliably detecting below-chance performance. Guilmette, Hart, Giuliano, Leininger (1994) have reported a similar modification of their version of the HDMT, as has Binder (1993c). Second, perceived item difficulty was manipulated by

making half of the foils highly similar to targets, while the other half was made completely dissimilar. A pilot study (Slick et al., 1994) showed that the new test was effective for detecting malingering. However, the results of the study suggested that sensitivity could be greatly increased by developing a normative based scoring system to augment probabilistic cut scores.

Relationship Between Symptom Validity Tests and MMPI-2 Validity Scales

Few studies have directly examined the relationship between symptom validity test scores and MMPI scores. Villanueva and Binder (1993) used scores from the Portland Digit Recognition Test (PDRT) to divide compensation-seeking mild-head trauma patients into high- and low-motivation groups, and compared their MMPI-2 validity scales to those of non-compensation-seeking traumatic brain injured patients. The Obvious-Subtle subscale of the MMPI-2 was found to reliably distinguish between all three groups, while the F scale only discriminated high from low motivation groups. These two scales also showed negative correlations with the PDRT (r = -.38 and r = -.40 respectively), indicating that patients who performed most poorly on the PDRT tended to have the highest validity scale elevations on the MMPI-2. However, Youngjohn, Burrows, and Erdal, (1995) found a substantially smaller correlation of -.18 between the PDRT total correct and F Scale scores in a sample of litigating post-concussive patients. Greiffenstein, Gola, and Baker (1995) found that MMPI-2 Validity Scales, including F, F-K, and O-S, all loaded highly on a single factor separate from a second factor on which the PDRT, the RMT, and several other performance-based malingering measures were highly loaded.

Methodological Considerations for Studies of Malingering

As Rogers (1988) and others have pointed out, the methods commonly employed for developing and validating instruments and procedures for detecting malingering, whether as stand-alone tests or for use with conventional measures, are not without problems. Some of these problems (e.g., demonstrating adequate test-retest reliability) are common to the development of any psychological test. Other problems (e.g., generalizing from analog samples and coaching) are particular to the development of measures of symptom validity.

Analog Studies

The use of healthy participants that are instructed to feign deficits (i.e., as analogs to real malingerers) is commonplace in research on malingering due to a lack of identifiable

malingering populations (see criterion validity, below). This method is generally recognized as a requirement for developing cutoff scores for malingering indices and establishing elements of reliability and validity. However, as Rogers (1983) points out, analog studies introduce the following paradox: participants are asked to *comply* with instructions to *fake* in order to estimate the performance of patients who *fake* when asked to *comply*. This results in a state of affairs where "it is difficult to know the degree of compliance under either the experimental or naturalistic condition." (p.447). In addition to this issue, the usual considerations for normative populations apply: the analog sample should be as similar as possible to typical patient populations with which the test will be used, and the test should be administered the test in standardized fashion in conditions that do not depart significantly from typical clinical settings. However, several other issues must be addressed when designing analog studies or reviewing findings to determine how well analog data will generalize to the clinical setting.

Motivation

Rogers (1983, 1988) has repeatedly pointed out one of the major potential challenges to the validity of analog studies on malingering: the effects of motivation on performance. Real life situations in which persons malinger are typically those in which both incentives for success and penalties for failure are substantial (e.g., civil and criminal litigation). Participants in simulation studies on malingering, even those who are offered financial incentives, are obviously operating under different parameters. Most often, participants in analog studies are university students who receive course credit for participation regardless of their level of effort. It seems clear that "the incentive of a few dollars to a college student, the incentive of monthly income (in some instances, lifetime support), and the incentive of freedom from incarceration cannot fairly be construed as equivalent incentives for faking bad" (Fredrick et al., 1994, p. 124). Despite the obvious inability of simulation studies to approximate the magnitude of rewards for success and penalties for failure that await real-world malingerers, researchers should strive to provide the best possible approximation of these influences on behavior. To that end, Rogers (1988) recommends that participants be given some type of material incentive for performing well, such as a financial reward for successful dissimulation, and that some type of censure be provided for being detected, such as publicly posting the names of those participants who were poor malingerers.[1] Recent research has provided qualified support for the first of Rogers' two suggestions. Frederick,

Sarfaty, Johnston and Powel (1994) found reduced detection rates among experimental participants who were offered financial incentives ($20.00) for successfully malingering cognitive impairment (including memory deficits). Using a variety of measures, including forced-choice tests and performance-pattern indices they found a decrease in the detection rate from 94% to 81% percent in naïve malingerers, and from 74% to 49% among coached malingerers. In contrast to Fredrick, et, al.'s findings, Bernard (1990) found that incentives affected participants' perceptions of how well the were able to fake memory deficits, but not their actual performance. Group means did not differ between Participants who were given a financial incentive for successful faking ($50.00 each to the two best fakers) and those who were given no incentive. Martin, Bolter, Todd, Gouvier, and Niccolls (1993) also found that monetary incentive did not affect performance, although they used a very small sum ($2.00) as a reward for adequate faking. Despite the negative findings, it seems clear from Frederick and Foster's work that incentives can significantly impact performance under some conditions. Studies which do not supply incentives may therefore overestimate the sensitivity of malingering measures.

Participant Self-Preparation and Background Knowledge

Another potential problem with analog studies is that "subjects are rarely given any preparation time or opportunities to plan their strategy of deception" (Rogers, 1988). Although some malingering undoubtedly occurs on a spur-of-the-moment basis, many persons who malinger do so in a premeditated fashion. However, participants in analog studies are typically given little time to prepare themselves and think about the task of dissimulating. Allowing participants in analog studies time to plan, and perhaps research their approach to dissimulating, especially in conjunction with material incentives for success, may provide a more realistic estimate of the distribution of performance of malingerers, and thus a more realistic estimate of test sensitivity.

Malingerers come to assessment with varying degrees of knowledge about what "real" neuropsychological disorders and deficit look like. People acquire this knowledge through experience, self-study, and the counsel of others. When developing tests for detecting malingering, it is helpful to have an idea of how much lay persons typically know about the behavioral manifestations of cognitive deficits following head injury, and to what extent education in these matters facilitates malingering. Many lay persons probably know that even mild-head trauma can result in lasting cognitive deficits, but the proportion of the

population who understand the functional manifestations of such deficits (i.e., how they would translate into test performance) is unknown.

What can we assume about the knowledge of typical participants in the studies? Surprisingly, only a small number of studies have evaluated the level of public knowledge about the sequelae of head injury and strategies for malingering. The data suggest that although naïve individuals are capable of endorsing symptoms consistent with head injury, they are also susceptible to endorsement of highly unusual items that distinguish them form legitimately head injured individuals. For example, Aubry, Dobbs, and Rue (1989) administered a symptom checklist containing probable and improbable symptoms to a small sample of undergraduates to test their knowledge of sequelae of minor head trauma. They found that most participants had good knowledge about physical symptoms that typically follow car accidents sufficient to produce minor head trauma or whiplash. However, participant knowledge about typical cognitive and psychiatric symptoms was poor, with many participants endorsing highly unusual symptoms, such as uncontrollable laughter. In fact, unusual symptoms were endorsed with the same frequency as much more likely cognitive sequelae such as difficulty remembering phone numbers. Similar results were also reported by Gouvier, Prestholt, and Warner (1988).

In contrast to these findings, are reports of good understanding of cognitive sequelae of head injury among naïve participants reported by Lees-Haley and Dunn (1994) and Mittenberg, DiGiulio, Perrin, and Bass (1992). Lees-Haley and Dunn administered a series of symptom checklist questionnaires to 97 undergraduates to assess their knowledge of psychiatric and neurological syndromes. One checklist contained ten symptoms common in mild brain injury. When asked to check those symptoms they felt were common to mild brain injury, 63% of the participants checked 5 or more of the 10 items on the head injury symptom checklist. In a similar, but expanded study Mittenberg et al administered a 30 item symptom checklist (including affective, somatic, and memory items) to a community sample of 223 adults who had no history of head injury or contact with a head injured person. The checklist was also administered to a sample of 100 patients seen for neuropsychological assessment following head injury. Healthy participants were asked to check any symptoms they regularly experienced within the last six months, and to separately check any symptoms they would expect to experience following a mild head injury sustained in an automobile accident. Head injury patients were instructed to check current symptoms, and to separately

check any symptoms they experienced regularly in the past. A very high rate of agreement was found between the symptoms endorsed by healthy and head injured participants. Mean number of reported symptoms did not differ between the groups, nor did frequency of endorsement of 22 of the 30 symptoms. Rank analysis showed a correlation of .82 between the groups when symptoms were ranked by endorsement frequency. Interestingly, head injury patients endorsed significantly fewer pre-injury symptoms than healthy individuals endorsed, suggesting that they may have underestimated their pre-injury status and attributed pre-injury symptoms to their head injury. One limitation of this study is that the composition of head injury patients with respect to litigation or other compensation-seeking activities was not reported. If the sample included a substantial number of these patients, then it is reasonable to assume that the data were biased by some amount of exaggeration. In addition, neither of these studies used distracter items, so the reported accuracy rates of naïve participants are quite possibly inflated relative to procedures that employ distracter items or require participants to generate a list of symptoms rather than choosing them from a checklist. Nonetheless, Lees Haley and Dunn make the point that their approach may be more applicable to typical clinical practices because clinicians often give patients checklists that do not contain improbable symptoms. Furthermore, in interview, clinicians often ask about common head injury symptoms but rarely ask about improbable or contradictory symptoms. Finally, there is concern for the practice of giving interviews and checklists without querying improbable symptoms at the beginning of assessments, as such practices may serve to cue malingering patients about deficits to report and portray. Clearly, clinicians need to consider their approach to interviewing, and also whether they should employ or develop neuropsychological symptom checklists that include improbable symptoms.

Although there are now several studies on public knowledge about the sequelae of head injury, only one study to date has systematically evaluated actual test-taking and self-presentation strategies that persons use to malinger. Iverson (1995), obtained self-report information of malingering strategies used by experimental study participants including university undergraduates, community volunteers, psychiatric inpatients, and federal inmates. Strategies for both preparation and test-taking were reported. Less than 4% of the respondents described any individual method for preparation, such as studying the effects of head injury, or engaging in corroboratory behaviors, such as missing appointments. The test-taking strategy reported most often (16%) was to fake total amnesia. Other reported

strategies included "poor cooperation, aggravation and frustration, slow response latencies and frequent hesitations, and general confusion during the testing process" (p. 37). Responses were notable both for limited numbers of strategies and limited descriptions of how strategies would be operationalized. Unfortunately, no break down of strategy prevalence was provided for the different groups. Interestingly, no study has directly measured the relationship between participant knowledge and ability to believably fake deficits.

Instructions, Coaching, and Test Security

Anecdotes abound about insurance fraud networks and unscrupulous lawyers providing coaching to personal-injury litigants. Persons with health-care backgrounds or other exposure to neuropsychological knowledge are also an inevitable component of the patient population. Indeed, neuropsychologists may be their own worst enemy in this regard, as the impressive body of publicly available literature on neuropsychology, and neuropsychological assessment in particular, is an open invitation to fraud for the reasonably intelligent and highly motivated. For this reason, many researchers consider providing coaching or test-taking instructions to analog participants as a means of estimating sensitivity of tests of malingering under "worst-case scenario" conditions.

Several recent studies have systematically explored the effects of coaching on malingering ability. There is evidence that coaching significantly facilitates malingering of head injury symptoms on both self-report measures (Berry, Wetter, & Baer, 1994), and performance measures such as the PDRT (Fredrick et al., 1994; Rose, Hall, & Szalda-Petree, 1995). For example, Martin et al. (1993) provided dramatic evidence of the efficacy of coaching. They found that whereas more than 80% of naïve malingering participants performed below chance on a version of the HDMT, only around 40% of those who received coaching performed below chance. Thus coaching was associated with a 50% reduction in detection rate. Coaching included the following three instructions "1) miss more of the difficult items than easy ones, 2) try to be fairly consistent in your responses by not missing easy items and then getting more difficult ones, and 3) be sure to perform at a level better than chance" (p. 880). Interestingly, although coaching lowered detection rates, it actually increased participants' susceptibility to manipulation of apparent difficulty. Scores of coached malingerers declined as retention interval increased, but scores of uncoached malingerers and brain-injured patients were unaffected by apparent difficulty. As described above,

however, the increased susceptibility to apparent difficulty found among coached malingerers was more than offset by their resistance to producing overly exaggerated deficits.

Although coaching analog participants seems like a logical and perhaps necessary component of test development, there are serious ethical considerations to such a course of action. The basic question is, "do we want to expose the public to information that may encourage and assist them to malinger?" Coached participants are under no formal ethical obligation to maintain test security or keep secret the content of what they learned during the experiment. Performance-based rewards provide feedback to participants about the relative effectiveness of various strategies. Articles that describe research on malingering and the efficacy of various coaching strategies provide highly distilled information on how to malinger (coaching instructions) as well as analyses on how well particular coaching strategies work (e.g., Martin et al., 1993). Even listing general symptoms for participants to simulate without giving specific feigning strategies provides a list of target behaviors for potential malingerers. In addition, most published research on tests of malingering describe the instruments and procedures in enough detail for readers to recognize such measures or tell others how to. Clearly all research on response bias needs to be reconciled with the code of ethics:

> Psychologists make reasonable efforts to maintain integrity and security of tests and other assessment techniques consistent with the law, contractual obligations, and in a manner that permits compliance with the requirements of this ethics code (Ethical Standard 2.10 of the American Psychological Association's Ethical Principles of Psychologist and Code of Conduct; APA, 1992).

Berry, Lamb, Wetter, Baer, and Widiger (1994) list a variety of options for dealing with the ethical dilemmas surrounding research on dissimulation:

> (a) taking the stance that the researcher has no ethical responsibility in this matter, (b) placing the burden of ethical judgment entirely on the researcher, (c) prohibiting publication of this type of work in American Psychological Association (APA) journals, (d) limiting the amount of information included in such publications in APA journals, (e) limiting access to such publications, or (f) charging editors with making an

evaluation of the costs and benefits of publications in terms of the impact on a test's integrity.

The first option is clearly in contradiction with the APA code of ethics. Options *b* and *f* combined probably capture the current state of affairs, with the primary emphasis on author responsibility. Options *c* and *e* represent extreme and draconian approaches to the problem that would protect test security but seriously hinder advances in test development. Perhaps some combination of options *d* and *e* might produce a workable solution. For example, articles on malingering could be published with limited details about the tests and coaching instructions employed. Detailed information could than be provided by some means to those persons ethically obligated to protect test security. As increasing attention is drawn to research on malingering, pressure will no doubt grow to the point where they are addressed formally by the APA. Until such time, Berry et al. (1994) recommend that "producers and consumers of research on coached malingering should carefully weigh the issues and keep their ethical responsibilities well in mind as they work in this complicated arena."

Criterion Validity

Although symptom validity tests and other methods for detecting malingering can demonstrate adequate construct validity in a variety of ways (e.g., divergent validity, convergent validity, and internal reliability consistency), perhaps the most important measure of validity for this type of test, congruence with external criterion, is generally very difficult or impossible to ascertain. Obviously, malingering patients will rarely self identify as such. As we have no way of knowing or even estimating the proportion of patients who malinger effectively enough to pass themselves off as legitimately impaired, the true base-rate and sensitivity of our current methods for detecting malingering are thus difficult to estimate, and may be poorer than we like to think. When patients are grouped by the weight of test evidence and clinical judgment into categories like "probable" or "definite" malingering, we can rarely be completely sure the grouping includes no false positives. Furthermore, if scores from the malingering test being evaluated enter into decisions about patient classification, then estimates of test criterion validity will almost always be favorably biased.

Clinical Considerations for Use of Measures for Detecting Biased Responding

Like any clinical instrument, the competent use and interpretation of malingering tests depends primarily on a clear understanding of the psychometric properties of the instrument, as well as any ethical or pragmatic considerations regarding its use.

Cut-scores vs. Diagnostic Probability

The idea of a cut-score, a point at which scores above and below signify meaningful differences in ability or status, is appealing in its ease of use and intellectual simplicity. This appeal is manifest in the wide acceptance and use of cut-scores in neuropsychology and by the fact that they are provided in the manuals of many neuropsychological tests in use. However, although cut-scores may be a useful, and sometimes necessary tool for classifying patient performance or ability, diagnostic labels like malingering should be qualified by estimates of diagnostic probability. Even below chance performance – typically considered an unequivocal sign of negative response bias – is still a probabilistic statement that leaves open some possibility, albeit remote, of diagnostic error. The use of normative based cutting scores for detecting malingering is even more problematic.

Meehl and Rosen (1955) were probably the first people to point out to psychologists the limitations of actuarial methods for detecting conditions with low or unknown base-rates. Since that time, similar warnings concerning the psychometric properties of clinical instruments have periodically resurfaced, but the issues are still poorly understood by many clinicians. For example, Vecchio (1966) demonstrated that a measure with sensitivity and specificity of .95 would have a false positive rate of 32% in a population where the base-rate for the target disorder is .10. This apparently paradox is lucidly explained in a recent article by Elwood (1993). He notes that people often focus exclusively on the weakest measure of criterion validity, statistical comparisons of mean scores from reference groups. Much less often, effect sizes are also presented as evidence of a test's ability to discriminate between groups of interest. However, neither p nor d values provide adequate information about a test's ability to make clinically useful discriminations. Recently, more attention has been paid to the evaluation of cutting scores through examinations of *sensitivity* (the probability of a positive test result when a patient has the disorder) and *specificity* (the probability of a negative test result when a patient does not have the disorder). These statistics provide more clinically useful measures of a test's general performance. However, Elwood points out that when clinicians use tests they are typically interested in the inverse of sensitivity: the

probability that a patient has the disorder given a positive test result (a test's *positive predictive power*).

Clinicians are also usually interested in the inverse of specificity: the probability that a patient does not have the disorder given a negative test result (a test's *negative predictive power*). Like Vecchio, Elwood shows that contrary to popular assumption, positive and negative predictive power are not equal to their inverses, sensitivity and specificity, but vary with prevalence rates in accordance with Bayes' theorem. Elwood provides an example of the application of Bayes' theorem showing that the positive predictive power of tests may be much poorer than their sensitivity, particularly in the context of low base rate disorders. Because of the general lack of awareness of these factors, large numbers of clinicians are probably overestimating the discriminative ability of many of their tests. Therefore, Elwood concludes with the following recommendations: 1) test authors should provide classification outcome data so that positive and negative predictive power may be derived; 2) cutoff scores should be adjusted to reflect the best estimates of local prevalence for a condition; 3) test interpretations should not exceed the limit of their predictive power; and 4) estimates of positive and negative predictive power are only as reliable as the test scores on which they are based. Other recommendations for working with low base-rate conditions are provided by Derogatis and DellaPietra (1994) and Lindeboom (1989).

Derogatis and DellaPietra (1994) propose a practical method for increasing positive predictive power in which the "base-rate" of the target disorder (e.g., malingering) is increased through the use of *sequential screening*. In this procedure, a highly sensitive screening test is used to eliminate those individuals who do not have the disorder. Persons identified by this procedure now form a higher base-rate population who can be further screened by measures with high specificity (and ideally, high sensitivity as well). Positive predictive power increases as a function of the increase in base-rate among the patients given the second screening. This increases confidence in positive findings from the second screening.

Lindeboom (1989) describes a useful application of Bayes' theorem for producing an alternate to standard cutoff scores. Lindeboom points out that psychological testing usually produces estimates of ability or status in the form of scores that fall along continua. Therefore, interpretations of scores vary in confidence as a function of their closeness to the extreme ends of the scale. However, cutoff scores reduce diagnosis to an either-or

proposition, and preclude more appropriate statements of diagnostic probability. Further, cutoff scores, as typically presented, do not make allowances for differing base-rate conditions. As an alternative, Lindeboom provides a method for using Bayes' theorem to compute classification probabilities for a range of test scores given normative data from two populations of interest (e.g., normal elderly and demented). These probabilities can then be adjusted for varying base-rates so that a classification confidence matrix can be constructed for given scores and assumed base-rates (see Appendix II). Lindeboom concludes that because test results carry considerable implications, and because base-rates may differ significantly from practice to practice, test authors should strongly consider publishing such tables for use with their tests.

False-Negatives and False-Positives

When cut-scores are recommended by test authors, consideration must be taken of the cost of both false-negatives and false-positives. Although the cost to the system of false-negatives is high, and continues to increase with the size of settlements, the cost to the individual of a false positive is devastating. Cut-scores at the boundary between chance and below-chance performance run a low risk of producing false-positives. However, norm-based cut scores, which are generally somewhere at or above chance, run a higher risk of producing false positives. Any time a clinician wishes to use norm-based cut-scores, special attention should be paid to converging evidence. Here especially, the diagnostic probability matrix technique proposed by Lindeboom (1993) may be more appropriate.

Deception of Patients

It is nearly impossible to develop an effective malingering test that does not involve some type of deception. Symptom validity tests, for example, are presented as tests of perceptual or cognitive function, rather than as tests of biased responding. Of course, this practice is not really deceptive, as some level of intact function is required for very good or very poor performance. Examinees are therefore given incomplete rather than incorrect information. In this respect, symptom validity tests are no different from other commonly used neuropsychological tests. The Rey Complex Figure Test, for example, is introduced as a copying task; examinees are not informed in advance that delayed recall will be assessed (Spreen & Strauss, 1991). Most malingering measures do differ significantly from conventional neuropsychological measures, however, in that the stated or implied level of

difficulty significantly exceeds the actual difficulty level for most patients (e.g., Binder, 1992; Hiscock & Hiscock, 1989). Such deceptive procedures introduce an ethical dilemma in which professional responsibilities must be balanced against responsibility to respect and protect examinees (c.f. American Psychological Association Ethical Standards 1.07, 1.21, & 1.14, 1992). Nevertheless, when contracted to perform a medicolegal assessment, a psychologist's paramount professional responsibility is to provide a report which thoroughly and impartially addresses the referral question. Only protecting individuals from physical harm is of higher concern. Thus, symptom validity testing may be a required part of psychological assessments in cases where poor motivation, oppositional behavior, malingering, or other psychological factors are a legitimate possibility. Many clinicians adopt the attitude that only those clients who malinger will be significantly impacted by the use of deceptive measures for detecting dissimulation, and in these cases, it is the clients and not the clinician that bear the primary burden of responsibility for the outcome. Others, like Prigatano and Amin (1993), argue for a more conservative approach. They recommend only non-disclosure of the true nature of symptom validity tests; any deceptive statement to the client concerning test difficulty is considered inappropriate.

Warning Patients About the Nature of Testing

It is advisable on both ethical and legal grounds to routinely provide a general warning to medicolegal patients prior to assessment that (1) a necessary component of assessment involves screening for poor effort and malingering, (2) that findings of poor effort or malingering are likely to affect diagnostic accuracy and influence legal proceedings, and (3) that the instruments employed are sensitive to poor effort and fabrication of deficits. The practice of providing a general warning to examinees is consistent with APA ethical standards concerning disclosure and clarification of the nature of services and relationships between parties, and taking reasonable steps to avoid and minimize harm to examinees (see APA Ethical Standards 1.07, 1.21, & 1.14; APA, 1992). At the same time, providing a general warning about the nature and purposes of testing rather than individual warnings for specific tests adheres to ethical principles governing the maintenance of test security and integrity and the need for psychologists to maintain professional standards for impartiality and thoroughness of assessments (see APA Ethical Standards 2.01, 2.02, 2.10, 7.01, & 7.02; APA, 1992). In this regard, malingering tests do not differ from many other tests such as

the MMPI-2 (Butcher, et al., 1989) in which examinees are encouraged to respond honestly, but are not explicitly warned that the test contains dissimulation scales.

Providing explicit warnings prior to administration of highly specific tests is also an option, but puts the clinician in the bind of possibly tipping off malingering patients to perform well only on such measures. The situation becomes even more complicated when questionable performance is evident at some point part way through testing. The decision about whether or not to provide additional reminders or warnings depends to a large degree on the clinician's focus (detection vs. prevention).

Other Factors Influencing Effort

Whenever non-optimal patient performance is suspected, it becomes necessary to determine the cause(s). Malingering is but one of many possible explanations for poorer than expected performance. Experienced clinicians know that a patient's performance may not be reflective of his or her highest level of function or ability for a variety of reasons other than the calculated seeking of financial rewards or escape from work. Pain, fatigue, psychiatric disturbance, or distrust of health-care professionals may impact performance or cooperation. In these cases too, it is important to have a valid and reliable index of whether, and by how much, a patient may be under-performing. Therefore, a test designed to detect response bias or poor effort may be useful whenever malingering or other reasons for substantially less than maximum performance are suspected (e.g., depression). However, it is of the utmost importance to recognize that data from symptom validity tests such as the VSVT are, at best, only capable of indicating that factors other than cognitive impairment may be influencing patient performance. Even in cases where financial or other external incentives exist and a patient's performance is in the questionable or invalid range, the patient may be legitimately impaired and/or acting without conscious intent. Consider the case of a patient with executive dysfunction who is disinhibited and evidences impaired foresight and judgment. Suppose the patient impulsively attempts to malinger or approaches the test oppositionaly without fully appreciating the nature of his or her actions or anticipating the likely consequences. Is such a patient a true "malingerer", or are her or his scores on the VSVT evidence of legitimate cognitive dysfunction? Likewise, severe disturbances of arousal or attention secondary to head injury and/or psychiatric disorders may manifest as chance-level performance (i.e., in the questionable range). Clearly, the

ethical responsibility of the clinician is to thoroughly explore all of the alternative explanations when patient performance is in the questionable or invalid range.

Research Development of the Victoria Symptom Validity Test

Following the creation of the VSVT in 1992, its utility was evaluated in a pilot study. The results of the study are presented below.

Pilot Study

A pilot study was carried out to assess the basic psychometric properties and utility of the original VSVT (then known as the Victoria Revision of the Hiscock Digit Memory Test). Goals of the study were to (a) obtain estimates of sensitivity and specificity, and (b) determine if any substantial modifications were required to improve the utility of the test.

Method

Participants

The VSVT was administered to the following groups of normative research participants:

1. **Control**

   This group consisted of twenty young adults recruited from the undergraduate population at the University of Victoria. Participants were excluded if they self-identified as having uncorrected sensory-perceptual disorders, history of head injury requiring hospitalization, or current treatment of psychiatric disorder.

2. **Feigning**

   This group consisted of twenty young adults recruited from the undergraduate population at the University of Victoria. Participants were excluded if they self-identified as having uncorrected sensory-perceptual disorders, history of head injury requiring hospitalization, or current treatment of psychiatric disorder.

3. **Closed-Head Injury Patients**

This group consisted of ten patients who were seen following closed-head injuries (CHI) suffered in automobile accidents. At the time of testing, all CHI patients were seeking claims or litigating for damages sustained in their accidents. The patients were tested 6-48 months post-injury. Exclusionary criteria for CHI patients included behavior or performance consistent with overt symptom fabrication (e.g., test scores significantly worse than expected given reports on or observations of everyday function; provision of bizarre or highly improbable symptoms or symptom history such as detailed recall of the accident together with claims of significant antero- or retrograde amnesia); Wechsler Adult Intelligence Scale-Revised (WAIS-R) Full Scale IQ less than 75; severe memory deficits at testing, positive history of psychosis; and severe disturbance in language or visuo-perceptual functions.

Measures

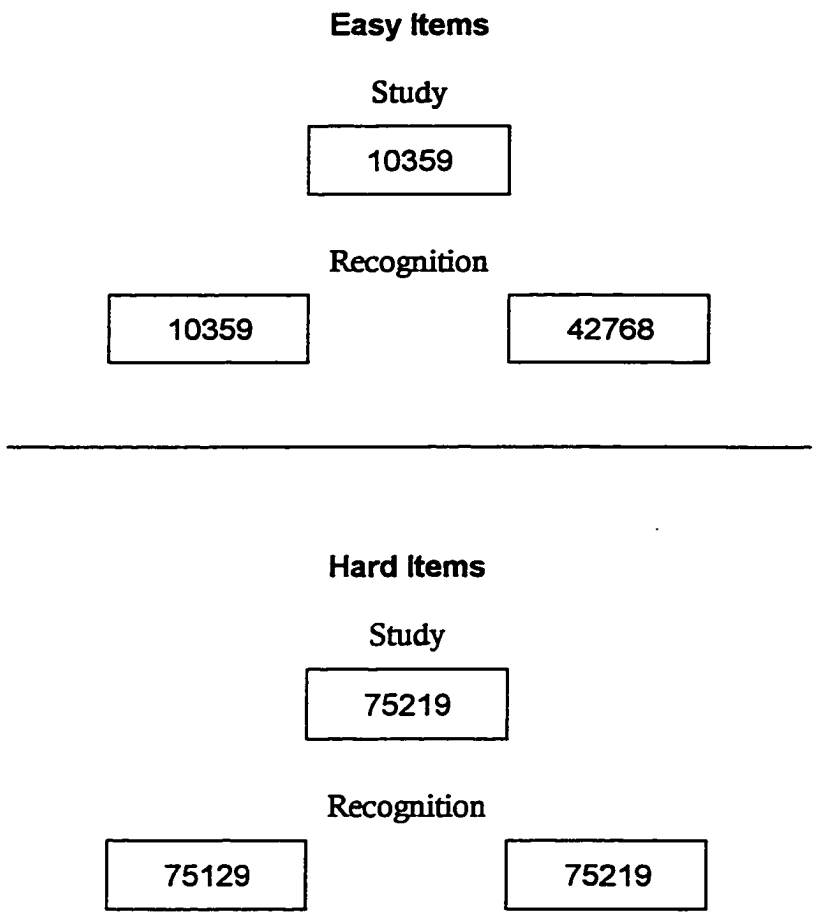
Victoria Symptom Validity Test

The VSVT includes a total of 48 items, presented in three blocks of 16 items each. In each block, 5-digit numbers are individually presented for five seconds of study at the center of a computer monitor. The presentation of study numbers is followed by a blank-screen retention interval, after which the previously shown study number and a 5-digit foil are displayed, one to each side of screen center (at equal height and offset). Patients respond by striking one of two keys on a conventional PC keyboard. Location of correct choices and foils (left/right) is counterbalanced and pseudo-randomized on a preset, invariant order. The retention interval is 5 seconds in the first block, 10 seconds in the second block, and 15 seconds in third block. Stimuli are displayed horizontally in a large (approximately 2 cm) screen font.

Item difficulty in the Victoria Revision is operationalized as follows: in easy items, the foil and correct choice share no common digits (see Figure 1). Recognition of the first, last, or any other digit or pattern of digits from the study number will facilitate a correct choice. In hard items, the foil is identical to the correct choice with the exception of a transposition of the second and third, or third and fourth digits (see Figure 1). To choose correctly on hard items, the order of the middle digits must be remembered. Recognition of the first or last digit of the study number will not aid in choosing. An equal number (12 each) of

second-third and third-fourth place transpositions are included in the set of hard foils. All three sections contained an equal number (8 each) of easy and hard items. Within sections, the order of easy and hard items and screen location (left / right) of foils are pseudo-randomized on a preset, invariant order.

Figure 1. Example VSVT stimuli



A version of the VSVT program was developed to run on a LAN network in a computer lab at the University of Victoria, allowing testing of up to 16 participants at a time. Control and feigning participants were tested in small groups and individually. Participants with CHI were tested individually in a clinical setting. As a part of the standard instructions, all participants were told that they would be "taking a test of memory that requires concentration," and that "people with memory problems often find this test to be difficult." Prior to both increases in retention interval, participants were told that the test becomes more difficult because the retention interval increases.

Procedure

Informed consent was obtained from all participants prior to data collection. CHI patients were given standard administrations of the VSVT as a part of full neuropsychological assessments. Control and feigning participants were administered a brief self-report questionnaire to obtain demographic information and pertinent health history. Control subjects were then administered the VSVT with standard instructions. Feigning participants also received the standard instructions, but were first given the following scenario and additional instructions:

> I want you to pretend that you were involved in a serious automobile accident six months ago. You are currently attempting to sue the insurance company for damages and hope to increase the size of your claim by faking brain damage. This is one of the tests that the neuropsychologist gives you to determine the authenticity of your claim. It is important that the results of this test do not make it obvious that you are faking as this could result in a loss of settlement and severe court penalty.

Following completion of the test, those participants who were instructed to feign memory impairment were also asked to indicate on a scale from 0 to 4, how successful they felt they were at "faking brain damage."

Dependent Measures

Basic demographic data (age, sex, education) was obtained from all participants. VSVT data obtained from all participants included number correct from each condition (2 difficulty levels x 3 retention intervals). Additional data available from participants with CHI included Full Scale IQ (FSIQ) from the Wechsler Adult Intelligence Scale-Revised (WAIS-R: Wechsler, 1981), total and recognition scores from the Rey Auditory Verbal Learning Test (RAVLT: Spreen & Strauss, 1991), and scores from the Logical Memory I and Logical Memory II subtests (LM-I & LM-II) from the Wechsler Memory Scale Revised (WMS-R: Wechsler, 1987).

Results

Basic demographic data from all three participant groups are presented in Table 1.

Table 1. Participant demographics

	N	Age	Sex (M/F)	Yrs Ed.
Control	22	28 (11)	6/16	15 (1)
Feigned	20	27 (9)	5/15	15 (3)
CHI	10	28 (8)	7/3	11 (1)

Note. *SD* in parentheses.

The participant groups did not differ in age. Control and malingering groups were equivalent in years of education and male/female ratio, whereas the CHI group included more males than females [$\chi^2(2, N = 52) = 6.87, p < .05$] and was less educated on average [$F(2, 49) = 15.25, p < .001$]. Additional data from the CHI group is presented in Table 2.

Table 2. CHI patient mean test scores

	Patient Score	Norm
WAIS-R FSIQ	92 (16)	100 (15)
RAVLT Total Score	47 (13)	55 (7)
RAVLT Recognition Score	12 (3)	14 (1)
WMS-R LM-I	12 (6)	23 (7)
WMS-R LM-II	9 (6)	19 (7)

Note. *SD* in parentheses; norms taken from Spreen & Strauss (1991).

The mean self-rating of success for feigning participants was 2.9 out of 4 ($SD = .9$). Number of correct responses per condition was submitted to a mixed-model ANOVA. Group membership defined the three-level between-subjects factor; item difficulty (two-levels), and retention interval (three-levels) defined within-subject factors. Both second-order interactions involving group membership were significant.[2] Overall, the performance of the groups differed as a function of item difficulty [$F(2,49) = 24.1, p < .001$], and as a function of retention interval [$F(4,98) = 6.8, p < .001$]. No other interactions were significant. The number of items correct for participant groups by item difficulty and retention interval are presented in Table 3. It can be seen that the distribution of control

participants' scores was marked by very small variance and restricted range; typically one or less items was missed out of the entire 48. It can also be seen that although the interaction with retention interval was found to be significant, differences in performance across time intervals were not substantial except in the feigning group. To examine the possibility that group differences in education may have accounted for variance in participant performance, the ANOVA described above was reevaluated with years of education as a covariate. The main effect and interactions of education level were not significant, nor did its inclusion substantively change any of the other effects observed in the original analysis. Thus, education level was not included in any further analyses.

Table 3. Means and standard deviations for VSVT scores

		Control	Feigning	CHI
Total	(Max = 48)	47.4 (0.9)	30.3 (9.2)	41.9 (4.8)
Total Easy	(Max = 24)	24.0 (0.2)	19.5 (5.5)	32.5 (0.8)
Total Hard	(Max = 24)	23.5 (0.9)	10.8 (5.0)	18.4 (4.7)
Easy 5s	(Max = 8)	8.0 (0.2)	7.0 (0.2)	8.0 (0.4)
Easy 10s	(Max = 8)	8.0 (0.3)	6.7 (0.3)	7.8 (0.4)
Easy 15s	(Max = 8)	8.0 (0.3)	5.8 (0.3)	7.7 (0.4)
Hard 5s	(Max = 8)	7.7 (0.3)	4.3 (0.4)	6.7 (0.5)
Hard 10s	(Max = 8)	7.9 (0.3)	3.2 (0.3)	5.8 (0.5)
Hard 15s	(Max = 8)	7.9 (0.4)	3.4 (0.4)	5.9 (0.4)

Although the above analysis indicated that the groups performed differently, the practical implications in terms of clinical utility of the test were not sufficiently evaluated. For clinical purposes, the most accurate classifications of clients are sought, with an emphasis on limiting false-positive malingering classifications. Ideally, such classifications are obtained through the simplest combination of test-derived cut-scores. It can be seen from the preceding analysis that the group's scores were maximally divergent on difficult items, and on items with 15-s retention intervals. To simulate the cross-validity of participant classification using those scores, separate jack-knifed discriminant functions were calculated using the total number of difficult items correct, and the total number correct at the 15-s retention interval.[3]

The function using scores on difficult items successfully classified 77% of the participants, and the function using scores at the 15-s retention interval successfully classified 83% of the participants. The results of the latter function are presented in Table 4. Specifically, 100% of controls, 80% of feigning participants, and 50% of Participants with CHI were correctly classified. Twenty percent of the feigning participants succeeded in obtaining CHI classifications, and one of the ten brain-injured patients (10%) was classified as feigning.

Table 4. Group membership (rows) by discriminant classification (columns)

	n	Control	Feigning	TBI
Control	22	22 (1.00)	0 (.00)	0 (.00)
Feigning	20	0 (.00)	16 (.80)	4 (.20)
TBI	10	4 (.40)	1 (.10)	5 (.50)

Note. Proportion by row in parentheses.

To investigate the utility of chance-based classification criteria, and compare it with the results of the discriminant function classifications, binomial probability z scores were calculated for the number of difficult items correct for each participant, as suggested by Hiscock and Hiscock (1989). Two z score cutoffs were applied to determine their efficacy for distinguishing between participant groups. A conservative z score cutoff for detecting feigning was established at -1.64; scores below this level are significantly below chance at $p < .05$ (one-tailed). Twenty percent of the feigning participants performed below chance at the established criterion on difficult items. No CHI or normal participants performed below the criterion. A less conservative z score cutoff was set at -1.05; scores below this level represent performance below chance at $p < .15$ (one-tailed). Forty percent of the feigning participants performed below chance at the less conservative criterion. Again, no CHI or normal participants performed below the criterion.

Discussion

Hiscock and Hiscock's Digit-Memory Test (1989) was constructed to discriminate overt, naive malingerers from both normal participants and clinical patients. However, the very low difficulty level of the original measure reduces the amount of clinical data it provides and may also alert participants to its nature. In addition, the large number of items may be

excessive for some applications. Therefore, a revised measure was developed with a reduced number of items and the addition of two levels of item difficulty. It was hoped that the Victoria Revision would show enhanced sensitivity to real deficits while maintaining the ability to detect motivational problems.

Data from our study indicate that although probabilistic analysis is unlikely to produce false-positives, it may be ineffective for detecting all but the most overt attempts at dissimulation. Even when the cutoff probability for below chance performance was set at .15, less than half of the feigning participants were correctly identified. However, no false-positives were generated--an expected finding given the considerable unlikelihood from a probabilistic standpoint that a person with even severe memory dysfunction would ever score so poorly when performing at his or her best. One can therefore be quite confident that some secondary factors are influencing performance when confronted with scores below chance at a conservative p value.

Although probabilistic criteria produced a low hit-rate, the data indicate that norm-based criteria may be more effective at distinguishing cases of motivational problems, particularly if scores on items that appear more difficult are examined. Jack-knifed discriminant analysis using scores at the longest (15-s) retention interval correctly classified 83% of participants from a pool including verified CHI patients, simulated malingerers, and normal controls. Similar findings using more difficult items on the PDRT to discriminate participant groups have been reported by Binder (1993a).

Follow-up Normative Study

Modifications to the VSVT were suggested by the pilot study and other sources. Changes were made to the program (reaction times were recorded) and instructions (potentially deceptive statements were changed). An initial normative study was then completed, with data collected from 1993 through 1996. Results from this follow-up study are presented below.

Objectives of the Study

In the current study, clinical and experimental populations were enlisted to (a) establish normative based cutoff scores for the VSVT, (b) evaluate the construct validity of the VSVT, (c) evaluate the test-retest reliability of the VSVT.

Method

Participants

The VSVT was administered to the following groups of normative research participants:

1. Control

   This group consisted of young adults ($n = 95$) recruited from the undergraduate population at the University of Victoria. This group includes the 22 control participants from the pilot study.[4] All participants reported English as their first language. Participants were not asked to provide their race; by observation, the great majority were white. Participants were excluded if they self-identified as having uncorrected sensory-perceptual disorders, history of head injury requiring hospitalization, or current treatment of psychiatric disorder.

2. Feigning

   This group consisted of young adults ($n = 43$) recruited from the undergraduate population at the University of Victoria. This group includes the 20 feigning participants from the pilot study[2]. All participants reported English as their first language. Participants were not asked to provide their race; by observation, the great majority were white. Participants were excluded if they self-identified as having uncorrected sensory-perceptual disorders, history of head injury requiring hospitalization, or current treatment of psychiatric disorder.

3. **Non-Compensation-Seeking Patients (Non-Comp)**

This group consisted of non-compensation-seeking patients with seizure disorder (n=20), together with a smaller number of patients with head injury or other brain dysfunction (n=12). Participants were not asked to provide their race. All patients spoke English as a primary language or were fluent for the purposes of testing. All patients received the VSVT as part of a full neuropsychological assessment. All patients presented with spontaneous complaints of significant everyday memory dysfunction. Memory deficits as measured by standard neuropsychological tests varied from mild to severe within this group.

4. **Compensation-Seeking Patients (Comp)**

This group consisted of patients (n = 206) seen in private practice for medico-legal neuropsychological evaluations following possible head injury. Ninety four percent of the patients reported no or only momentary loss of consciousness at the time of the accident; 3% reported 5-60 minutes of unconsciousness; and 3% reported loss of consciousness in excess of one hour. The median time since accident at assessment was 688 days (range = 86-7,637). Eighty six of these patients reported their race as White; the remaining patients were roughly evenly distributed among Asian, Black, East Indian, First Nations, Hispanic and other. Eighty-eight percent of the patients reported English as their native language; English was a second or third language among the remaining patients. At the time of testing, all patients were actively seeking monetary claims or litigating for damages including psychological distress and/or cognitive dysfunction resulting from accidents (almost all involving automobiles). Memory deficits as measured by standard neuropsychological tests varied from mild to severe (performance below the first percentile on some tests) within this group. This group includes the 10 participants from the pilot study.

Measures

Revised VSVT

Following the pilot study, modifications were made to the VSVT in an attempt to increase its efficacy. Other research (Holden & Kroner, 1992), direct observations of patients, and clinical anecdotes all suggested that response latency might be a useful adjunct measure for detecting malingering. Therefore, the revised VSVT included measures of

response latency (average latency and *SD* for easy and hard items). The instructions were modified accordingly; test-takers were encouraged to respond as rapidly as possible without making mistakes, but were not told explicitly that response latencies were being recorded. It was hoped that this would cue feigning participants to realize that responses were being timed, and to significantly extend their response latencies in an attempt to appear impaired.

A second modification to the VSVT was made in the interest of avoiding ethical problems when using measures of response bias. Prigatano and Amin (1993) found that the majority of a sample of patients with legitimate memory impairment did not make more errors with increases in retention interval on the HDMT. They pointed out that telling patients that the test becomes more difficult as retention intervals increase (standard instructions for the HDMT) is therefore deceptive, and suggested that these statements be eliminated from the instructions. In consideration of this point, slight changes in wording of instructions were incorporated in an attempt to limit deception. Whereas the instructions previously *stated* that item difficulty increased with longer retention intervals, the new instructions *suggested* that participants might find items more difficult as retention intervals increased.

Other Measures of Symptom Validity

The MMPI-2 (Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989) was administered to 20 of the control and 20 of the feigning participants in the same session as the they received the VSVT. Clinical participants received individually tailored assessments. MMPI-2 data was available for 6 non-compensation-seeking and 92 compensation seeking participants.

Neuropsychological Measures

Selected scores from other neuropsychological tests were used in various analyses. These tests included: WAIS-R Full Scale IQ & Digits Forward (Wechsler, 1981); total correct from the North American Adult Reading Test (NAART: Blair & Spreen, 1989); Peterson Trigrams total scores (Peterson & Peterson, 1959); Logical Memory I and II subtests from the WMS-R (Wechsler, 1987); total and recognition scores from the RAVLT (Spreen & Strauss, 1991); copy and recall scores from the RCFT (Spreen & Strauss, 1991); color, word, and color-word scores from the Stroop Test (Regard, 1981); the time scores from Trail Making Test, parts A and B (Reitan & Wolfson, 1985); number of categories and

perseverative errors from the Wisconsin Card Sorting Test (WCST: Heaton, 1981); and 5 of the content scales from the MMPI-2 (Butcher et al., 1989)).

Procedure

Informed consent was obtained from all participants prior to data collection. Compensation-seeking and non-compensation-seeking patients were tested individually in a clinical setting. They received standard administrations of the VSVT and MMPI-2 during full neuropsychological assessments. Control and feigning participants were administered a brief self-report questionnaire to obtain demographic information and pertinent health history. Small groups of control and feigning participants were then administered the VSVT followed by the MMPI-2. Control and feigning participants received the standard instructions for the VSVT and MMPI-2. Prior to testing, feigning participants were also given the following scenario and additional instructions:

> I want you to pretend that you were a passenger in a serious automobile accident six months ago. In the accident, you received a blow to the head resulting in a concussion and a momentary loss of consciousness. Now you are suing the insurance company for $250,000 in damages, which you plan to obtain by faking problems with headaches, fatigue, memory, concentration, anxiety, and depression resulting from "brain damage" suffered in the accident. The two tests you are about to take are part of the examination by the insurance company psychologist to determine the authenticity of your claim. Your goal is to appear realistically impaired on these tests. It is very important that your performance on these tests does not make it obvious that you are faking as this will result in a loss of settlement and a $10,000 fine.

Results

Demographic Data

Demographic data from the participants are presented in Table 5. The control and feigning participants were about ten years younger and had about two years more of education on average than the patients.

Table 5. Demographic statistics for participant groups

	Control	Feigning	Non-comp	Comp
n	95	43	32	206
Gender (M / F)	42 / 53	17 / 26	14 / 18	120 / 86
Age (years)	22.7 (8.1)[a]	23.9 (7.3)[a]	35.3 (12.1)[b]	32.1 (12.4)[b]
Education (years)	13.7 (1.4)[a]	14.5 (2.2)[a]	12.6 (3.0)[b]	11.9 (2.3)[b]

Note. Non-Comp = Non-Compensation-Seeking; Comp = Compensation-seeking; *SD* in parentheses; *a* significantly different from *b* (Tukey-H*SD* at .05 or less)

IQ and Memory Test Scores

Memory test score data from the participants are presented in Table 6 along with results of t-tests for group mean differences. With the exception of scores on digits forward – where compensation-seeking patients obtained raw scores about two points lower than non-compensation-seeking patients – the patient groups did not differ significantly in performance on memory tests. Both patient groups contained individuals who performed at floor level on at least one memory test.

Table 6. Clinical groups: Selected mean test scores with *SD's* and ranges

	Non-Comp	Comp	sig.
WAIS-R FSIQ	97 (12) 80-136	98 (14) 73-142	ns
Digits Forward (WAIS-R)	8 (3) 4-14	6 (1) 3-12	.006[1]
Logical Memory - I	17 (8) 3-35	18 (6) 3-35	ns
Logical Memory - II	12 (10) 0-34	14 (7) 2-27	ns
RAVLT Total	47 (12) 24-67	47 (10) 19-68	ns
RAVLT Recognition	12 (3) 5-15	12 (3) 1-15	ns
RCFT Recall	17 (8) 4.5-23	20 (7) 3.5-35	ns

Note. FSIQ = Full-Scale IQ; RAVLT= Rey Auditory Verbal Learning Test; RCFT= Rey Complex Figure.
[1] *df* adjusted for unequal variance

VSVT Scores

Mean scores for number of items correct are presented in Table 7, along with standard deviations. For the control group, range was restricted due to ceiling effects. Ceiling effects are also present in data for the non-compensation-seeking patients. Significant group effects were found for each dependent variable (Kruskal Wallis tests for k independent means; all $\chi^2 > 58$; all $p < .001$). It can be seen by inspection that there were no meaningful effects of increasing retention interval for any group (i.e., all mean differences in number of items correct across intervals were less than 1). In contrast, item difficulty was associated with meaningful differences in performance for feigning participants and compensation-seeking patients (approximately 9 and 3 more easy items correct, respectively). These differences in performance for easy and hard items were also statistically significant (feigning: t=11.3, $p < .001$; non-comp: t=11.2, $p < .001$).

Table 7. Means and standard deviations for VSVT scores

		Control	Feigning	Non-Comp	Comp
Total	(Max = 48)	47.4 (0.9)	31.3 (9.1)	46.2 (2.6)	43.5 (6.1)
Total Easy	(Max = 24)	24.0 (0.2)	20.3 (4.4)	23.5 (1.2)	23.3 (2.0)
Total Hard	(Max = 24)	23.4 (0.9)	10.9 (6.1)	22.6 (1.8)	20.1 (4.8)
Block 1	(Max = 16)	15.7 (0.6)	11.3 (3.1)	15.3 (2.0)	14.7 (2.0)
Block 2	(Max = 16)	15.8 (0.4)	10.2 (3.1)	15.3 (1.2)	14.6 (2.1)
Block 3	(Max = 16)	15.8 (0.4)	9.8 (3.5)	15.6 (1.0)	14.3 (2.4)
Easy 5s	(Max = 8)	8.0 (0.1)	7.2 (1.5)	7.8 (1.0)	7.9 (0.6)
Easy 10s	(Max = 8)	8.0 (0.0)	6.8 (1.7)	7.8 (0.6)	7.8 (0.6)
Easy 15s	(Max = 8)	8.0 (0.1)	6.3 (1.8)	7.9 (0.3)	7.7 (1.0)
Hard 5s	(Max = 8)	7.8 (0.6)	4.0 (2.2)	7.5 (1.0)	6.8 (1.7)
Hard 10s	(Max = 8)	7.8 (0.4)	3.4 (2.1)	7.5 (0.9)	6.8 (1.8)
Hard 15s	(Max = 8)	7.8 (0.4)	3.5 (2.4)	7.7 (1.0)	6.6 (1.8)
Easy item response latency		1.29 (0.38)	2.48 (1.05)	1.66 (0.59)	2.04 (0.99)
Hard item response latency		1.93 (0.51)	4.08 (2.12)	2.61 (1.40)	3.41 (1.92)

Note. response latencies in seconds.

Table 8 provides distribution data in the form of selected percentiles with corresponding easy and hard item scores. This data is useful for determining how unusual a given score is relative to specified reference groups. Table 8 also illustrates the positive skew found for all score distributions except those of the feigning group.

Table 8. Means, *SD's*, and percentiles for easy and hard items correct

		Control		Feigning		Non-Comp		Comp*	
		Easy	Hard	Easy	Hard	Easy	Hard	Easy	Hard
Mean		24.0	23.4	20.3	10.9	23.5	22.6	23.6	20.1
SD		0.2	0.9	4.4	6.1	1.2	1.8	1.0	3.7
Percentiles	99	24	24	24	24	24	24	24	24
	95	24	24	24	23	24	24	24	24
	50	24	24	21	10	24	23	24	22
	5	23	21	11	2	21	18	21	12
	1	23	20	0	0	18	18	18	9

* does not include participants who scored below chance ($n = 10$)

Cut-Scores: Patient Classification

Traditionally, dichotomous classification rules have been used to interpret data from symptom validity tests such as the VSVT. Protocols in which the number of correct items is less than expected by chance at a small probability (e.g., $p < .05$) are considered invalid or malingered. Protocols where the number of correct items are at or above chance are considered valid. This cutoff score has the advantages of a negligible false-positive rate, even for the most severely impaired patents (i.e., patients reduced to random responding by severe cognitive impairments are only expected to perform below chance on 5 out of 100 administrations). However, when this stringent cutoff score (less than 9/24 correct on easy or hard items) was applied to VSVT scores, sensitivity was found to be poor, with only 39% of the feigning participants correctly classified. As expected, specificity was excellent, with no false-positives among controls and non-compensation-seeking participants.

Boxplots in Figures 2 and 3 show the distributions of the participant groups with respect to chance performance. Center lines represent median scores; box ends delimit quartiles. Figure 1 shows the ceiling effect found in all groups except feigners. It can be seen from Figure 3 that the majority of the feigning participants did not perform below chance on hard

items, but rather performed at chance. At the same time, all of the non-compensation seeking patients – even those with severe memory impairments – performed above chance. This post hoc observation led to the development of a different approach to classifying patient performance on the VSVT, a new, three-category classification system.

Figure 2. Easy Items Correct by Group (Max = 24)

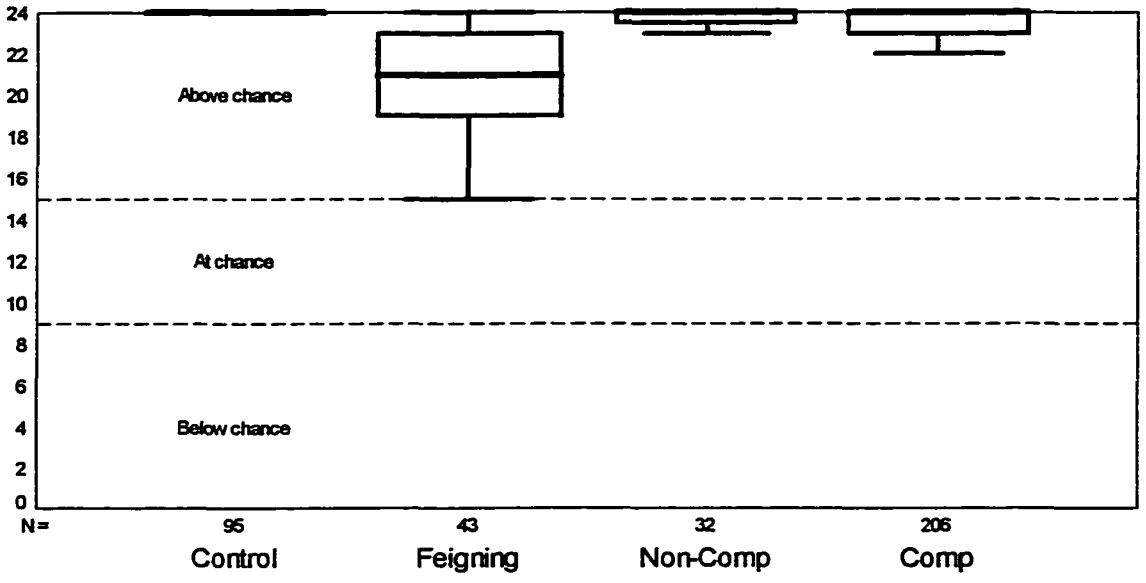
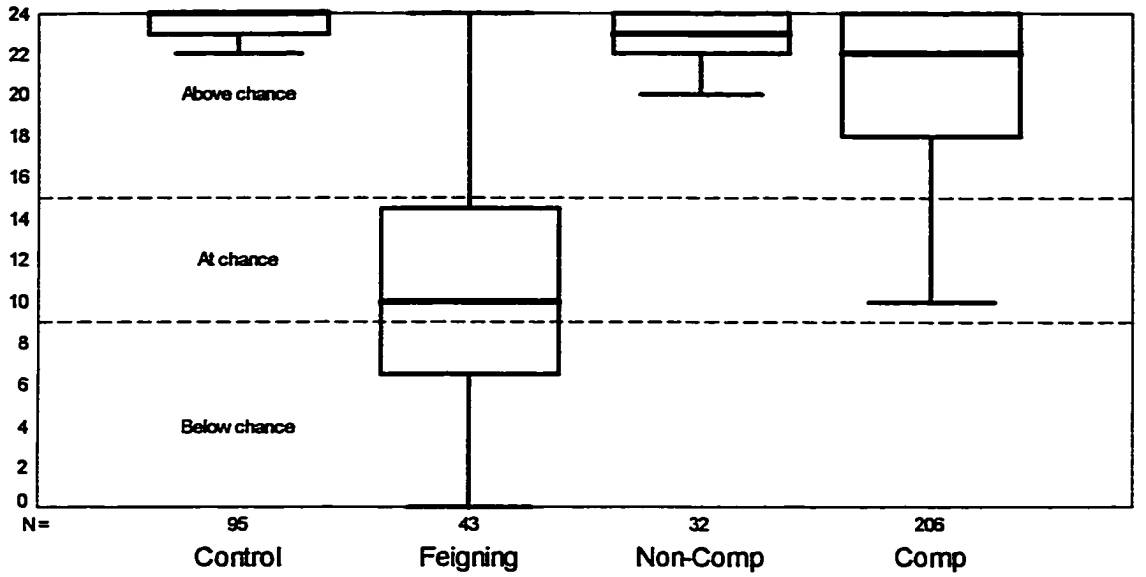


Figure 3. Hard Items Correct by Group (Max = 24)



Under the new three-category system, below chance performance continues to be labeled as unequivocally invalid/malingered. Performance significantly above chance is likewise labeled as unequivocally valid. The new, third category applies to scores that fall within the remaining confidence interval around chance level performance. Protocols with this level of performance were labeled "questionable," as they represented a level of impairment (i.e., scores resulting from guessing) that is expected to be highly infrequent among outpatient medico-legal cases with mild to moderate head-injuries. The decision rules for the three category system are elaborated in Table 9.

Table 9. Classification rules for VSVT scores

Valid	Questionable	Invalid / Malingering
Above chance: $p \leq .052$	*At chance:* within 89.6% confidence interval	*Below chance:* $p \leq .052$
total correct > 28, and easy correct > 15, and hard correct > 15.	29 > total correct > 19, or 16 > easy correct > 8, or 16 > hard correct > 8.	total correct < 20, or easy correct < 9, or hard correct < 9.

The results of applying the three-category cut-off to easy and hard item scores obtained by participants are presented in Table 10.

Table 10. Classification of participants by VSVT scores on easy and hard items using three-category system

	Valid	Questionable	Invalid / Malingering
	above chance ($p \leq .052$)	*at chance (89.6% C.I.)*	*below chance ($p \leq .052$)*
Control	95 (100)	0	0
Feigning	8 (19)	18 (42)	17 (39)
Non-Comp	32 (100)	0	0
Comp	176 (85)	20 (10)	10 (5)

Note. percentage per group in parentheses

Using the new three-category classification system, sensitivity was greatly improved with minimal negative effects on specificity. Zero-level false-positive rates were obtained among the control and non-compensation-seeking participants (i.e., all performed above chance). This finding is unlikely to be an artifact of group constituency, as several of these patients had severely impaired memory function (e.g., scores below the 1st percentile) as measured at assessment and confirmed by independent report. Within the feigning group, 42% were labeled questionable. An additional 10% of compensation seeking patients were labeled as questionable. To better quantify the usefulness of the three-level system, clinical judgments were obtained about status of the 20 compensation-seeking patients who obtained scores in the questionable range. Based on all available data (including VSVT scores) , five of these patients (25%) were judged to be non-malingerers. Deficits displayed by four patients (20%) were judged to be of questionable validity, but without sufficient evidence to make conclusive determinations. Eleven of the patients (55%) were judged to be cases of exaggeration or outright fabrication of deficits.

Additional normative data bearing on the issue of scores in the questionable range have been made available by Dr. David Berry and colleagues (personal communication, July, 1995). A slightly modified version of the VSVT was administered to 30 moderate to severely closed-head-injured adults (20 male, 10 female) who were not receiving or seeking financial compensation at the time of testing. Administration was modified to use retention intervals of 2.5, 5, and 10 seconds for the first, second, and third sets respectively. Data from these cases are provided below in Table 11 and Table 12.

Table 11. Descriptive data for sample Provided by David Berry, et al.

	Mean	*SD*	Range
Age	33.5	10.6	18 - 55
Years of education	12.2	3.1	7 - 20
Years since injury	4.8	6.8	.05 - 25
Days of unconsciousness	20.5	28.6	0 - 115
Days of post-traumatic amnesia	63.1	74.7	0 - 330
Days of retrograde amnesia	54.9	178.6	0 - 730

Table 12. Means, *SD's*, and percentiles for easy and hard items correct

		Easy	Hard
Mean		23.8	22.5
SD		.7	2.6
Percentiles	99	24	24
	95	24	24
	50	24	24
	5	22	15
	1	20	12

The three-level classification criteria were applied to the data. Twenty-nine (97%) of the head-injured patients obtained scores in the valid range, while one patient (3%) obtained a score in the questionable range. It is possible Berry et al.'s patients may have benefited from the shorter, non-standard delay interval used in their study. However, scores in this sample was not meaningfully affected by retention interval. For example, the mean difference between hard item scores from 5 and 10 second retention intervals was +0.03 points. Recall also from Table 7 that changes in retention interval did not significantly impact scores in the normative clinical samples, even though some patients in the non-compensation-seeking group had severe memory impairments. This strongly suggests that the nonstandard (shorter) retention intervals employed by Dr. Berry have minimal impact on the generalizability of his findings to applications of standard retention intervals.

Thus, Dr. Berry's data support the contention that scores in the questionable range are unlikely, even following moderate to severe head injury.

Classification Matrix

Using the technique suggested by Lindeboom (1992), a classification confidence matrix was created for scores on easy items (Table 13) and hard items correct (Table 14). The sample of TBI patients provided by Berry et al. and feigning participants from the current study served as reference populations. To use the matrix, find the row corresponding to a score of interest (number of easy or hard items correct). Next, select a column for the assumed prevalence of malingering in your population (5%-50%).[5] The junction of the row and column provides the probability that the score belongs to a feigning participant. This is the estimated probability that the person taking the test is malingering. For example, Table 14 shows that a score of 12 hard items correct is associated with a probability of malingering of only .57 if the assumed base rate is .15. Given an assumed base rate of .35, the same score is associated with a probability of malingering of .80. Clearly, the effects of base rates are not trivial for certain scores.

On the other hand, base rate adjustments sometimes make little difference. It can be seen from Table 13 that regardless of estimated base rates, scores below 20 easy items correct were associated with a high probability of malingering. Scores of 20 or greater were at most associated with a 60% probability of malingering, and then only when a base rate of 50% was assumed. This marked trend is the result of limited overlap between feigning and TBI distributions; adjustments for base rates decrease as the separation of reference population distributions increases. Thus, in cases where distributions are widely separated, as they were for easy item scores from feigning and TBI samples, diagnostic probability matrices may offer little practical advantage over unadjusted cutoff scores.

Table 13. Classification confidence matrix for easy items: Data from Berry's TBI patients (*n*=30) and feigning participants (*n*=43)

Easy Items	TBI	Feigning	LR*	Probability of Feigning for Assumed Base-Rate (p')									
# correct	cum. %ile	cum. %ile		5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
0	.00	1.00	99.74	.84	.92	.95	.96	.97	.98	.98	.99	.99	.99
1	.00	1.00	99.74	.84	.92	.95	.96	.97	.98	.98	.99	.99	.99
2	.00	1.00	99.74	.84	.92	.95	.96	.97	.98	.98	.99	.99	.99
3	.00	1.00	99.74	.84	.92	.95	.96	.97	.98	.98	.99	.99	.99
4	.00	1.00	99.74	.84	.92	.95	.96	.97	.98	.98	.99	.99	.99
5	.00	1.00	99.74	.84	.92	.95	.96	.97	.98	.98	.99	.99	.99
6	.00	1.00	99.74	.84	.92	.95	.96	.97	.98	.98	.99	.99	.99
7	.00	1.00	99.74	.84	.92	.95	.96	.97	.98	.98	.99	.99	.99
8	.00	1.00	99.74	.84	.92	.95	.96	.97	.98	.98	.99	.99	.99
9	.00	1.00	99.74	.84	.92	.95	.96	.97	.98	.98	.99	.99	.99
10	.00	1.00	99.74	.84	.92	.95	.96	.97	.98	.98	.99	.99	.99
11	.00	.98	99.69	.84	.92	.95	.96	.97	.98	.98	.99	.99	.99
12	.00	.98	99.69	.84	.92	.95	.96	.97	.98	.98	.99	.99	.99
13	.00	.98	99.69	.84	.92	.95	.96	.97	.98	.98	.99	.99	.99
14	.00	.98	99.69	.84	.92	.95	.96	.97	.98	.98	.99	.99	.99
15	.00	.95	99.56	.84	.92	.95	.96	.97	.98	.98	.99	.99	.99
16	.00	.95	99.56	.84	.92	.95	.96	.97	.98	.98	.99	.99	.99
17	.00	.88	98.68	.84	.92	.95	.96	.97	.98	.98	.99	.99	.99
18	.00	.88	98.68	.84	.92	.95	.96	.97	.98	.98	.99	.99	.99
19	.00	.81	97.01	.84	.92	.94	.96	.97	.98	.98	.98	.99	.99
20	.33	.72	1.50	.07	.14	.21	.27	.33	.39	.45	.50	.55	.60
21	.33	.65	1.44	.07	.14	.20	.27	.32	.38	.44	.49	.54	.59
22	.33	.47	1.22	.06	.12	.18	.23	.29	.34	.40	.45	.50	.55
23	.67	.35	.71	.04	.07	.11	.15	.19	.23	.28	.32	.37	.41
24	1.00	.16	.38	.02	.04	.06	.09	.11	.14	.17	.20	.24	.27

*Likelihood ratio (see Lindeboom, 1992)

Table 14. Classification confidence matrix for hard items: Data from Berry's TBI patients (*n*=30) and feigning participants (*n*=43)

Hard Items	TBI	Feigning	LR*	Probability of Feigning for Assumed Base-Rate (p')									
# correct	cum. %ile	cum. %ile		5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
0	.00	1.00	99.74	.84	.92	.95	.96	.97	.98	.98	.99	.99	.99
1	.00	.98	99.69	.84	.92	.95	.96	.97	.98	.98	.99	.99	.99
2	.00	.98	99.69	.84	.92	.95	.96	.97	.98	.98	.99	.99	.99
3	.00	.93	99.35	.84	.92	.95	.96	.97	.98	.98	.99	.99	.99
4	.00	.91	99.06	.84	.92	.95	.96	.97	.98	.98	.99	.99	.99
5	.00	.84	97.65	.84	.92	.95	.96	.97	.98	.98	.98	.99	.99
6	.00	.79	96.29	.84	.91	.94	.96	.97	.98	.98	.98	.99	.99
7	.00	.74	94.56	.83	.91	.94	.96	.97	.98	.98	.98	.99	.99
8	.00	.70	92.50	.83	.91	.94	.96	.97	.98	.98	.98	.99	.99
9	.00	.61	87.25	.82	.91	.94	.96	.97	.97	.98	.98	.99	.99
10	.00	.56	84.01	.82	.90	.94	.95	.97	.97	.98	.98	.99	.99
11	.00	.49	78.42	.80	.90	.93	.95	.96	.97	.98	.98	.98	.99
12	.03	.47	7.43	.28	.45	.57	.65	.71	.76	.80	.83	.86	.88
13	.03	.42	7.00	.27	.44	.55	.64	.70	.75	.79	.82	.85	.88
14	.03	.33	5.99	.24	.40	.51	.60	.67	.72	.76	.80	.83	.86
15	.03	.26	5.09	.21	.36	.47	.56	.63	.69	.73	.77	.81	.84
16	.03	.19	4.05	.18	.31	.42	.50	.57	.63	.69	.73	.77	.80
17	.03	.16	3.67	.16	.29	.39	.48	.55	.61	.66	.71	.75	.79
18	.10	.16	1.46	.07	.14	.21	.27	.33	.39	.44	.49	.54	.59
19	.10	.14	1.30	.06	.13	.19	.25	.30	.36	.41	.46	.52	.57
20	.13	.09	.75	.04	.08	.12	.16	.20	.24	.29	.33	.38	.43
21	.23	.09	.50	.03	.05	.08	.11	.14	.18	.21	.25	.29	.33
22	.27	.09	.45	.02	.05	.07	.10	.13	.16	.20	.23	.27	.31
23	.50	.07	.24	.01	.03	.04	.06	.07	.09	.12	.14	.17	.20
24	1.00	.05	.14	.01	.02	.02	.03	.04	.06	.07	.08	.10	.12

*Likelihood ratio (see Lindeboom, 1992)

Response Latencies

Response latency parameters for the sample are provided in Table 15 (unless specifically noted, all further references to patients do *not* include the sample provided by Berry et al.). All participants were collapsed into a single group, then regrouped by their classification under the three-level classification system. It can be seen by inspection of the 95% confidence intervals that those participants who produced questionable or invalid protocols took significantly longer to respond compared to those who produced valid profiles. This difference is particularly evident when response latencies for easy items are examined. The response latency distributions of the questionable and invalid groups were found to have substantial overlap, indicating similar performances of participants from these two groups.

Table 15. Means, *SD's* and 95% confidence intervals for response latencies (in seconds) to easy and hard items for participant with valid, questionable, and invalid protocols

		Easy Items			Hard Items		
		$\bar{x}$	*SD*	95% CI	$\bar{x}$	*SD*	95% CI
Valid	*n*=135	1.73	(0.84)	1.58 - 1.87	2.82	(1.44)	2.58 - 3.07
Questionable	*n*=20	2.62	(1.46)	1.94 - 3.30	5.43	(4.27)	3.79 - 7.79
Invalid	*n*=15	3.22	(1.14)	2.58 - 3.85	4.36	(1.75)	3.52 - 5.46

Effects of Age, Education, and Gender on VSVT Scores and Response latencies

Effects of age and years of education on VSVT scores and response latencies were evaluated with Spearman rank correlations. Table 16 shows that that among all patients obtaining scores in the valid range, age and education were significantly correlated with numbers of items correct and response latencies. However, correlations with age were small (less than 5% of variance shared), suggesting minimal clinical significance of this finding. Correlations between education and response times were in a range (13-15% of variance shared) suggestive of possible clinical significance.

Table 16. Correlations of age and education with VSVT scores among all participants producing valid protocols (n=208)

	Number Correct		Response Latency	
	Easy	Hard	Easy	Hard
Age	-.13*	-.12*	.21**	.22**
Education	.08	.19**	-.39**	-.36**

Note. * $p < .05$; ** $p < .001$

Table 17 shows that among patients scoring below chance on the VSVT, age and education were not significantly related to VSVT measures. Although small sample size precluded significant findings for all but very large correlations, all observed correlations were of a size not suggestive of clinically significant relationships.

Table 17. Correlations of age and education with VSVT scores among participants producing invalid protocols (n=10)

	Number Correct		Response Latency	
	Easy	Hard	Easy	Hard
Age	.19	.17	.15	.21
Education	-.17	-.10	-.21	-.16

Note. no correlation significant; all $p > .15$

Effects of gender were evaluated by determining the proportions of female and male patients classified as valid, questionable, and invalid. For women, proportions classified as valid, questionable, and invalid (88%, 8%, 4%, respectively) were not substantially different than those found for men (87%, 9%, and 4%, respectively). Thus, it appears that gender is not a significant predictor of response bias as measured by the VSVT

Effects of Time Since Injury on VSVT Scores

For all patients (comp and non-comp pooled) who obtained scores in the valid (above chance) range, time since injury was not significantly correlated (Spearman) with number of easy or hard items correct, or response time to easy or hard items (all $r < .16$; all $p > .05$). Among patients who performed at or below chance, time since injury was also not

significantly correlated (Spearman) with number of easy or hard items correct, or response time to easy or hard items. However the lack of significance was in part an artifact of small sample size, as medium sized correlations were obtained between time since injury and number of easy items correct (r = -.31), hard items correct (r = -.40), and response time to easy items (r = .33). It appears that for those patients scoring in the questionable and invalid range, length of time since injury may be moderately associated with magnitude of bias in performance.

Internal Reliability Consistency

The entire clinical sample (Comp and Non-Comp, including patients performing at or below chance) was pooled for estimating internal reliability consistency. Alphas for the 24 easy items, 24 hard items, and the entire set of 48 items were .82, .87, and .89, respectively. These findings confirm the assumption that the test items are homogenous.

Test-Retest Reliability

The VSVT was administered twice to subsets of control participants and compensation seeking patients. Demographic information for these subsets are presented in Table 18. The test-retest interval was 14 days for control participants. Median test-retest interval for compensation-seeking patients was 31 days (range = 1 - 550 days). The test-retest samples are comparable to the larger samples from which they were drawn age and education.

Table 18. Demographic statistics for test-retest participant groups

	n	Sex (M / F)	Age *	Years of Education *
Control	30	12 / 18	21.9 (8.0)	13.6 (1.5)
Comp	27	9 / 19	34.6 (12.0)	11.0 (1.7)

* *SD* in parentheses

Selected scores for the VSVT obtained by the test-retest groups are presented in Table 19. Mean scores for the test-retest samples are comparable to the larger samples from which they were drawn.

Table 19. Means, *SD's* and ranges for VSVT scores for the test-retest groups

	Test 1	
	Control	Comp
Total	47.3 (0.9) 45-48	42.1 (7.0) 19-48
Total Easy	24.0 (0.0)	23.3 (1.7) 16-24
Total Hard	23.3 (0.9) 21-24	18.9 (5.9) 3-24
	Test 2	
Total	47.1 (1.3) 43-48	43.1 (6.6) 20-48
Total Easy	24.0 (0.2) 23-24	23.5 (1.2) 19-24
Total Hard	23.1 (1.4) 19-24	19.6 (5.8) 2-24

Test-retest correlations for selected measures from the VSVT are presented in Table 20. Because of restriction in ranges, the test-retest correlations for number of items correct for the control samples do not accurately reflect the reliability of the VSVT with this sample. As can be seen from Table 20 however, average performance at retest did not meaningfully differ from performance at initial testing for the control sample. Test-retest correlations for response latencies for the control participants were moderate in size. Because of non-normal distributions, Spearman correlations were computed for the compensation-seeking sample. A considerable range of time between first and second testing was present in the sample (1 - 550 days). To evaluate the effects of length of time on test-score change, test-retest interval (in days) was correlated with the difference between scores at first and second testing. Test-retest interval was also correlated with the difference between classifications at first and second testing. All obtained Spearman correlations were small (less than .10 in absolute magnitude) and nonsignificant.

Table 20. Test-retest correlations

	Control	Comp
Total	*	.73
Total Easy	*	.57
Total Hard	*	.72
Easy Response Time	.54	.60
Hard Response Time	.53	.72

* correlation not interpretable due to restricted range of sample

Table 21 shows the test-retest changes in easy and hard item scores for the compensation-seeking patients. Changes in scores for easy items were generally small; 96% of the sample obtained retest scores within ± 2 points of original scores. In contrast, scores on hard items showed greater variability. Only 60% of the participants obtained retest scores within ± 2 points of original scores. A substantial proportion of patients (22%) obtained retest scores differing from original test scores by ± 6 or greater.

Table 21. Distribution of test-retest score differences*

Easy Items			Hard Items		
Score	Frequency	Percent	Score	Frequency	Percent
-2	1	4	-7	3	11
-1	1	4	-2	1	4
0	19	70	-1	3	11
1	3	11	0	7	26
2	2	7	1	4	15
3	1	4	2	1	4
			3	3	11
			4	2	7
			6	1	4
			9	2	7

* positive score = gain at retest

One hundred percent of the control participants obtained the same classification at retest (valid) as they did at first testing. Table 22 shows the test-retest consistency of classification for compensation-seeking patients. Among the compensation-seeking patients, 23 (85%) obtained the same classification at retest. Classifications of two patients changed from valid to questionable at retest; one changed from questionable to valid; and one changed from invalid to valid. The contingency coefficient for this data (.69; $p < .001$) indicates that patient classifications tended to remain stable over time. However, this effect is largely due to stability of classifications among patients obtaining scores in the valid range at initial testing; only 2 of 21 (10%) of these patients changed classification at retest. In contrast, off the 6 patients classified as questionable or invalid at initial testing, 2 (33%) changed category at retest. Although the current sample is small, results suggest that a substantial minority of patients who score in the questionable or invalid range are likely to change category of classification at retest.

Table 22. Test-retest classification consistency

		Retest		
		Valid	Questionable	Malingering
Test	Valid	19	2	0
	Questionable	1	3	0
	Malingering	1	0	1

Divergent Validity

Divergent construct validity is demonstrated by small correlations between scores on tests which are designed to measure dissimilar constructs. This is particularly important for tests of symptom validity, which should be insensitive to actual level of cognitive function. Spearman rank correlations between VSVT scores and selected neuropsychological test scores for all patients (compensation- and non-compensation-seeking) who obtained VSVT scores in the valid range ($n = 208$) are presented in Table 23.[6] MMPI-2 variables were not used from protocols with raw VRIN scores > 12 (suggesting random responding) or more than 10 blank responses (Butcher, et al., 1989). Spearman correlations were chosen because scores for the VSVT were significantly positively skewed, and were therefore not suitable for Pearson correlational analyses.

Table 23. Divergent validity: Spearman correlations with tests of memory and other cognitive functions

Test	Easy Correct	Hard Correct	Easy RT	Hard RT
FSIQ	-.09	.09	-.27	-.18
NAART	-.01	-.03	***.34***	***.32***
Digits Forward	.04	.18	-.28	***-.32***
Peterson Trigrams	-.02	.10	-.12	-.20
Logical Memory I	-.08	.12	-.08	-.07
Logical Memory II	-.07	-.02	.17	.10
RAVLT Total	-.04	.08	-.09	-.08
RAVLT Recognition	-.18	-.01	.03	.08
RCFT Copy	-.04	.05	-.04	-.03
RCFT Recall	-.01	.07	-.04	.03
Stroop Colors	-.22	-.24	.15	.16
Stroop Word	-.29	-.21	.26	***.30***
Stroop Color-Word	-.23	-.27	.19	.26
Trails A	-.14	-.12	***.31***	***.30***
Trails B	-.05	-.16	.27	.24
WCST Categories	-.07	.05	-.09	-.02
WCST Pers. Errors	.05	-.04	.05	.04
MMPI-2: DEP	-.17	.05	.24	.28
MMPI-2: ANX	-.14	.20	.14	.18
MMPI-2: HEA	-.01	.08	.18	.17
MMPI-2: BIZ	.06	.05	.02	.05
MMPI-2: TRT	.01	.21	.15	.12

Note. each subject received individually tailored assessment, therefore *n's* are not the same for all correlations in the table (n range = 47 - 130); see page 39 for a list of tests and abbreviations.

Small correlations (.29 or less; Cohen 1988) were considered evidence of divergent validity (i.e., low sensitivity to real level of cognitive function). Generally, easy and hard item scores from the VSVT showed excellent divergent validity, although borderline

relationships was observed with scores from the Stroop Test. No memory test shared more than five percent of its variance with easy or hard items scores from the VSVT. Response latencies to easy and hard items showed less divergent validity however, as they were moderately correlated with digit span and measures with heavy processing speed components.

Convergent Validity

Convergent validity was evaluated by examining correspondence with MMPI-2 validity scales. MMPI-2 variables were not used from protocols with raw VRIN scores > 12 (suggesting random responding) or more than 10 blank responses (Butcher, et al., 1989). MMPI-2 scores meeting the criteria for inclusion were available for the following numbers of participants: Control=20, Feigning=18, Non-Comp=6, and Comp=87. Spearman Rank correlations of selected MMPI-2 validity scales with VSVT scores are presented in Table 24. Only correlations of medium to large size (.30 or greater; Cohen 1988) were considered evidence of a meaningful relationship. Correlations of this magnitude were obtained between VSVT item correct sores and F, F-K, and Lees-Haley Fake Bad scales (FBS) from the MMPI-2. Correlations with the F(p) and Obvious-Subtle scales from the MMPI-2 were all below threshold. With the exception of correlations with the Fake-Bad scale, relationships between response latencies and MMPI-2 validity scales were negligible.

Table 24. Convergent validity: Spearman correlations between VSVT scores and MMPI-2 validity scale scores

	Easy Correct	Hard Correct	Easy RT	Hard RT
F	***-.31***	***-.33***	.26	.25
F-K	***-.30***	-.24	.21	.19
F(p)	-.21	-.15	.24	.21
FBS	***-.37***	***-.32***	***.42***	***.42***
O-S	-.05	.06	-.01	-.01

Note. RT = response time

Moderate to large correlations support conclusions of convergence between measures, but for clinical purposes, additional, more practical evaluations of convergence are often more useful. This is because estimates of linear relationships between variables may only

weakly correspond with actual convergence of test-based diagnoses (e.g., impaired vs. unimpaired). Correlations may in fact grossly overestimate the level of diagnostic convergence of tests. The diagnostic convergence of the VSVT and MMPI-2 was therefore directly evaluated using the subsample of participants who had MMPI-2 scores as described above. MMPI-2 profiles were labeled malingered if any one or more of the following cutoffs were exceeded:

F: T > 99 (Butcher & Williams, 1992).
F-K: Raw F-K > 11 (Butcher & Williams, 1992).
F(p): T > 100 (Arbisi & Ben-Porath, 1995).
O-S: Total Obvious T minus Subtle T > 199 (Greene, 1991).
FBS: Raw scores > 24 (men) or 26 (women) (Lees-Haley, 1992).

None of the controls produced a protocol with scores over the cutoffs. Thirteen of 18 (65%) feigning participants, 1 of the 6 (17%) non-compensation seeking patients, and 33/87 (38%) compensation-seeking patients produced protocols with at least one elevation over the cutoff.

Table 25 shows the overlap of classifications of compensation-seeking patients obtained from the VSVT and MMPI-2. A low rate of agreement was obtained, with 30/87 (34%) compensation-seeking patients receiving opposing classifications from the two tests. This finding strongly suggests that the VSVT and MMPI-2 validity scales are not measuring the same constructs.

Table 25. Classification of compensation-seeking patients by VSVT scores vs. MMPI-2 scores

		MMPI-2	
		Valid	Malingering
VSVT	Valid	47	27
	Questionable	4	5
	Malingering	3	1

Discussion

This research extends the knowledge base for assessing the validity of neuropsychological complaints. Symptom validity tests are capable of providing unambiguous evidence of biased responding, but only for relatively extreme scores. However, the acquisition of normative data enhances the efficacy with which the VSVT can be used to detect biased responding. The proposed three-level classification system demonstrated superior efficiency in comparison to the traditional, single cutoff classification systems. Adding the questionable category greatly enhanced sensitivity by drawing attention to an additional 43 percent of known feigning participants and 12 percent of compensation -seeking patients. At the same time, specificity remained adequate for screening or corroborative purposes; only 25% of the compensation-seeking patients classified as questionable were judged to be non-malingerers, and none of the non-compensation-seeking patients were misclassified as questionable, despite the fact that most of these patients had extensive neurological histories and objective evidence of memory impairment, severe in some cases. Additionally, the discriminant validity analysis demonstrated that VSVT accuracy scores from valid protocols were not related to a variety of standard cognitive measures in general, and tests of memory in particular, indicating that performance on the VSVT is largely unaffected by level of cognitive function. This data supports the contention that scores in the questionable range, especially those at the low end, are likely to reflect some degree of exaggeration.

One criticism of the current study is that the non-compensation-seeking patients were not head injury cases. Thus, the low false-positive rate obtained with this group may not be representative of that found with compensation-seeking patients, most of whom have known or suspected head-injuries. However, Berry and his colleagues (personal communication, July, 1995) also found a low false-positive rate (0% falsely classified invalid; 3% falsely classified questionable) when the VSVT was administered to a group of moderately to severely closed-head injury patients. Furthermore, scores in this sample did not decline as retention interval increased. Thus, Berry et al.'s data support the contention that scores in the questionable range are uncommon in legitimately head-injured patients.

Scores in the questionable range should thus raise serious suspicion about the possibility of symptom exaggeration or poor effort due to other factors (e.g., depression, fatigue, etc.). Given a non-zero false-positive rate, however, caution is warranted when making clinical

determinations about the status of patients who perform within the questionable range, especially where scores are near the high end of the confidence interval. At the very least, scores within the questionable range should be treated as tentative indicators of symptom exaggeration to be confirmed or disconfirmed by other clinical evidence (e.g., additional symptom validity testing or retesting, other indices of effort, behavioral observations and collateral information). Clearly, legitimate cognitive deficits, such as severe attentional or memory problems should always be carefully considered as a first possibility whenever performance is found to be at chance level. When other possibilities have been ruled out, scores in the questionable range, by virtue of their low *normative* likelihood may be used as corroboratory evidence when other information also suggests dissimulation.

Baysean diagnostic probability tables derived from the normative data also showed the utility of a normative scoring system for the VSVT. These tables demonstrate the increase in positive predictive power provided by normative data. Even with low estimated base rates, moderately elevated scores relative to chance were found to be highly likely to result from malingering.

As expected, VSVT scores showed only moderate convergent validity with most of the validity scales and indices from the MMPI-2. However, the finding of only moderate convergent validity correlations and limited classification overlap with MMPI-2 validity scales is not surprising, as the tests differ considerably in task (self-report vs. actual performance), and domain (memory vs. psychological adjustment). To successfully feign CHI symptoms on the MMPI-2 requires knowledge or intuition about the type and extent of psychiatric and neurological symptoms that are likely, as well as those that are not. Successful feigning on the VSVT requires knowledge or intuition about the type and extent of memory deficits that might be expected following CHI, as well as self-monitoring of actual test performance. The most sophisticated feigning participants were able to produce valid protocols on both the MMPI-2 and VSVT, while the least sophisticated participants produced suspect protocols on both tests. In between these two extremes of feigning ability are the two groups of participants made up of people who were only able to "fool" either the MMPI-2 or the VSVT, but not both.

Consistent with findings from other studies (Beetar & Williams, 1994; Holden, 1995; Holden & Kroner, 1992; Rose, Hall, & Szalda-Petree, 1995;Strauss, Spellacy, Hunter, & Berry, 1995), response latencies from the VSVT also showed promise for helping

differentiate feigned from real impairment, although modest convergent and divergent validity findings, and correlations with education caution against over-interpretation of response latencies at this point. Further studies will likely increase the confidence with which response time data may be factored into decisions about patient motivation.

Despite generally positive findings, conclusions drawn from the current findings and clinical application of the VSVT must be tempered by several factors: (1) The age, education, and socio-economic characteristics of the samples (especially the analog samples) limits generalizability. Larger, more representative samples are a requirement for the construction of fully adequate norms. For example, having samples of legitimately impaired patients attempt to exaggerate deficits on retesting would provide very valuable data for the construction of classification confidence matrices. Cross validation with larger and more representative samples would greatly increase the confidence with which such systems may be applied clinically. Until such time as this type of data are available, clinicians who use the VSVT will need to exercise caution when using the existing data for determining whether questionable scores are the result of severe impairments or malingering.

(2) Feigning participants were not provided with incentives for successful malingering. Although a variety of strategies may be available to participants, some strategies (missing increasing proportions of items as difficulty increases) require more effort directed to online monitoring of performance and score-keeping than others (e.g., missing all items, missing every other item, or random responding). Unless incentives are provided, participants may be more likely to choose strategies that require the least effort – the same strategies that tend to be easiest to detect. Although research on the effects of incentives has produced mixed results (Bernard, 1990; Frederick et al., 1994, Martin et al., 1993), the sensitivity of the VSVT observed in the current study may be overestimated.

(3) The level of neuropsychological sophistication of feigning participants was not measured or systematically manipulated. One must assume that "naïve" participants come to malingering studies with varying degrees of applicable background knowledge obtained through occupational or educational experience, or contact with head-injured individuals. A relatively brief instrument, preferably containing some multiple-choice items with distracters and some short answer items, could have been administered to feigning participants prior to testing to ascertain their level of knowledge about the cognitive sequelae of head injury.

Besides allowing for a better characterization of the sample as a whole, this procedure would have allowed for direct analysis of the relationship between sophistication and malingering ability. Such an instrument could also be given to broader samples of the public at large and to targeted samples, such as compensation seeking patients. This normative data would be invaluable in making judgments about the generalizability of data from analog studies. Although manipulating participant sophistication is also an option for future studies, the ethical drawbacks to such research make this method difficult to justify, particularly given that other researchers have already demonstrated the effectiveness of coaching. Further research with the VSVT is planned to address the above listed shortcomings in experimental data collected to date, and to increase the size and breadth of the normative sample.

Conclusions

This dissertation describes the genesis of the Victoria Symptom Validity Test, a new measure of response bias for neuropsychology. The VSVT was designed to provide neuropsychologists with an efficient tool for screening for malingering or biased response due to other causes. Pilot and follow-up normative studies demonstrated that the design goals were largely met. Given that adequate interpretive precautions are taken, the existing normative based-scoring system can be used clinically.

Scores below chance are unequivocal evidence of biased responding. Scores in the questionable range or with moderate Bayesian probabilities of malingering direct the clinician to evaluate patients more extensively through additional symptom validity testing or retesting, interviewing, and background checking. When all other data was also considered, 55% of patients who obtained VSVT scores in the questionable range were judged by the clinicians involved to be malingering. Thus, scores in the questionable range can not presently be considered conclusive indicators of malingering. However, by virtue of their low normative likelihood, questionable range scores may properly be used as corroboratory evidence when other information also suggests dissimulation.

This project is one of very few to take to heart the recommendations of Meehl and others who have for over 50 years called for application of Bayesian principles to test score interpretation. Using two target populations (feigning and non-malingering head injured), a Bayesian diagnostic probability matrix was produced. When based on adequate normative data and used appropriately, this type of table provides the astute clinician with a much better estimate of the confidence with which test scores can be interpreted. Unfortunately, despite the obvious advantages of such methods, they have failed to catch on. This state of affairs is not unexpected however, as Bayesian methods require more work from clinicians (base rates must be accurately estimated either through literature search or development of local databases), and they often reveal the appallingly limited utility of many clinical measures. Nevertheless, it is time to do away altogether with the simplistic cutoff scores applied to many of our measures, and replace them with more appropriate Baysean matrices.

If one wishes to use the classification probability table, then decisions about base-rates must be made. The most economical approach is to start with an estimate of the population

base rate for patients coming to the practice, and determine whether certain subgroups of patients have higher estimated base-rates than others. Considerable differences have been found in the literature, with some investigators reporting base rates as low as 8% (Trueblood, 1993), and others reporting base rates as high as 60% (Greiffenstein et al., 1994). Clearly, appropriateness of samples and methods of classification enter into decisions about which reported base rates are appropriate for individual practices or patients. Perhaps the best method is for clinicians to develop their own local base-rates. For evaluating individual cases, the sequential screening process (Derogatis and DellaPietra, 1994) could be used. For example, patients meeting the criteria proposed by Geissenger et al. for overt malingering might be considered to have a very high prior probability of malingering. When such high prior probabilities can be assumed, confidence in classifications of malingering based on VSVT scores will be increased. One must always keep in mind however, that the validity of any classification confidence matrix varies with the extent that reference populations used for constructing it are representative of the populations with which it will be used in day-to-day clinical practice.

Clinicians clearly have many things to consider when making determinations about level of effort and motivation. These challenges are no harder or easier than any other diagnostic decisions. Like any other diagnostic endeavor, judgments about malingering are always associated with some type of error rate. Measures of effort and response bias are a necessary part of assessment whenever biased responding or poor effort may be suspected or anticipated. The routine use of symptom validity tests is strongly recommended in all cases where patients are seeking compensation, or have other motivation to exaggerate dysfunction. Given the current state of the art, the use of multiple measures (e.g., VSVT, RMT, PDRT, Rey 15 Item) is recommended for clarifying whether poor performance is intentional. Iverson and Franzen (1996) found an improvement in classification accuracy when multiple measures of malingering were used with a deficient performance on any one test indicating malingering. Most malingering participants in this study had deficient scores on more than one test. Where warranted, tests of psychosocial adjustment which include validity scales should also be used to verify patient claims of emotional distress.

One must also keep in mind that verified CNS damage or dysfunction does not rule out a diagnosis of malingering. It is possible for exaggeration or feigning to occur in cases where verified CNS pathology exists. Palmer, Boone, Allman, and Castro (1995) describe a

case study of a patient with a premorbid history of exaggerated physical and emotional dysfunction who subsequently developed radiographicaly verified cerebral infarcts. Results of neuropsychological testing, including a variety of measures of effort or response bias, showed that the patient was fabricating cognitive dysfunction. Thus, the extent of any real cognitive deficits could not be established.

Once biased responding has been identified, neuropsychologists must decide how to report it. Rogers (1988) provides very useful suggestions on how to categorize and report the severity of malingering. Binder and Thompson (1994) suggest that in some cases of where malingering is suspected "it is appropriate merely to comment upon the invalidity of testing and make no diagnosis" (p. 40).

Along with studies aimed at addressing limitations of the current normative data base for the VSVT, several other directions for research are available for those interested in malingering. Other methods for detecting biased responding need to be developed. For example, Richert, Hiscock, and Caroselli (1996) report on a new type of method for assessing compliance with test instructions that shows excellent promise for inclusion in medico-legal test batteries. In this method, one of two stimuli (P or Q) are repeatedly presented at subthreshold exposure durations, and participants are told to report which was observed. Although the likelihood of either stimulus being shown is 0.5, feedback is given suggesting a higher proportion of presentations of one of the stimuli. Participants who were instructed to do their best generally biased their guesses toward the direction suggested by the false feedback, producing a higher error rate. Participants who were instructed to malinger generally stayed at a base rate of guessing (0.5), producing a lower error rate. Thus, although the opposite applies to instruments such as the VSVT, cooperative patients may someday be identified with similar procedures by substantially *higher* error rates than uncooperative patients. Other approaches, using physiological measures, are also showing promise for detecting malingering (Rosenfeld, Ellwanger, & Sweet, 1995). Clearly, there is considerable work left to be done on the psychometric evaluation of malingering, and clinicians should soon have a variety of useful new measures at their disposal.

More needs to be done than simply developing and refining methods for detecting malingering, however. For example, almost nothing has been published on the characteristics of verified malingerers. We know little about whether they form a homogenous group, or can be divided into identifiable subgroups based on characteristics

such as nature of motivation (e.g., sociopathy vs. other causes). There has been almost no research on behavioral markers of malingering during testing (e.g., is oppositional behavior actually a marker of malingering?). Lastly, almost all research on malingering to date has focused on detection. It is time for neuropsychologists to start thinking about how to *prevent* malingering. Many litigants are in the system for years waiting for compensation that may never come. Because of this, patients may feign symptoms that were once real and debilitating in order to receive compensation that they feel is owed. Patients may also begin to exaggerate symptoms out of frustration and feelings of not being believed. The validity of these assumptions was supported by findings that length of time since injury correlated negatively with numbers of easy and hard items correct among patients who obtained questionable or invalid scores on the VSVT. These findings also suggest that decreasing the time between injury and final settlement outcome may reduce the rate of malingering. Another factor worth studying is social pressure from third parties. In the only study of this effect to date, Binder and Johnson-Greene (1995) report on a case where a patient's performance was significantly affected by environmental factors. Using an A-B-A-B design, they were able to conclusively demonstrate that the patient exaggerated deficits when a family member was present. They suggest that excluding significant others and attorneys from the testing environment may help prevent malingering. In sum, systematic research on the causes of malingering is long overdue, and will, when it arrives, indicate avenues through which psychologists can work on reducing the incidence of malingering through either direct clinical intervention or by advocating for systemic changes.

Footnotes

1. The prospects of this procedure passing a research review committee are extremely dim.

2. The homogeneity of covariance across repeated measures requirement was met (Greenhouse-Geisser Epsilons > .95). To ensure that between-group differences in variance were not unduly affecting findings, equivalent non-parametric rank-score tests were conducted (Puri & Sen, 1971). Results did not differ from any of those reported for parametric tests.

3. Because half of all items at each retention interval were difficult, a subset of items contributes to both overall scores on difficult items and overall scores on 15-second retention items. A single DF analysis including both scores was therefore considered inappropriate because the aim was to evaluate clinical utility by how accurately the groups could be distinguished using easily derived measures of performance. A DF analysis that included both scores would have produced results associated with statistically definable but practically unattainable variables (i.e., performance on difficult items with retention interval partialed out, and performance at 15-second retention interval with item difficulty partialed out).

4. The control participants in the pilot study did not differ significantly from those in the follow-up study in demographic characteristics or performance in scores on the VSVT. Demographic characteristics and VSVT scores for the feigning participants were likewise equivalent between the two studies.

5. Ideally, clinicians should use base rates derived from a local database. Failing that, base rates could be obtained from the literature. The starting base rate would then be attenuated by factors specific to the current patient. For example, patients meeting Greiffenstein et al.'s (1995) criteria for probable malingering may be assumed to have a higher base rate of malingering (ideally the increase in malingering base rate expected in this situation would be not be guessed at but would be empirically derived in the same fashion as the initial base rate).

6. Because participants who performed below chance on the VSVT were clearly exaggerating or feigning deficits, other test scores from these participants were likely to be invalid. Additionally, some participants who obtained VSVT scores in the questionable range were probably also exaggerating deficits. Therefore only data from

participants who performed above chance were included in the divergent validity analysis.

References

Arbisi, P.A., & Ben-Porath, Y.S. (1993). Interpretation of F scales for inpatients: Moving from art to science. Paper presented at the 28th Annual Symposium on Recent Developments in MMPI (MMPI-2/MMPI-A) Research. March, 1993, St. Petersburg, FL.

Arbisi, P.A., & Ben-Porath, Y.S. (1995). An MMPI-2 infrequent response scale for use with psychopathological populations: The Infrequency-Psychopathology Scale, F(p). *Psychological Assessment, 7(4)*, 424-431.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

American Psychological Association. (1985). *Standards for Educational and Psychological Testing.* Washington, DC: Author.

American Psychological Association. (1986). *Guidelines for Computer-Based Tests and Interpretations.* Washington, DC: Author.

American Psychological Association. (1992). *Ethical Principals of Psychologists and Code of Conduct.* Washington, DC: Author.

Aubrey, J.C., Dobbs, A.R., & Rule, B.G. (1989). Laypersons' knowledge about the sequelae of minor head injury and whiplash. *Journal of Neurology, Neurosurgery, and Psychiatry, 52,* 842-846.

Baker, G.A., Hanley, J.R., Jackson, H.F., Kimmance, S., & Slade, P. (1993). Detecting the faking of amnesia: Performance differences between simulators and patients with memory impairment. *Journal of Clinical and Experimental Neuropsychology, 15(5),* 668-684.

Barton, P.W., Boone, K.B., Allman, L., and Castro, D.B. (1995). Co-occurrence of brain lesions and cognitive deficit exaggeration. *The Clinical Neuropsychologist, 9(1),* 68-73.

Beetar, J.T., & Williams, J.M. (1995). Malingering response styles on the Memory Assessment Scales and symptom validity tests. *Archives of Clinical Neuropsychology, 10(1),* 57-72.

Ben-Porath, Y.S. (1994). The ethical dilemma of coached malingering research. *Psychological Assessment, 6(1), 14-15.*

Bernard, L.C. (1990). The detection of faked deficits on the Rey Auditory Verbal Learning Test: The effect of serial position. *Archives of Clinical Neuropsychology, 12(5),* 715-728.

Bernard, L.C. (1991). Prospects for faking believable memory deficits on neuropsychological tests and the use of incentives in simulation research. *Journal of Clinical and Experimental Neuropsychology, 6(1-2),* 81-88.

Bernard, L.C., Houston, W., & Natoli, L. (1993). Malingering on neuropsychological memory tests: Potential objective indicators. *Journal of Clinical Psychology, 49(1),* 45-53.

Bernard, L.C., McGrath, M.J., & Houston, W. (1996). The differential effects of simulating malingering, closed-head injury, and other CNS pathology on the Wisconsin Card Sorting Test: Support for the "Patter of Performance" hypothesis. *Archives of Clinical Neuropsychology, 11(3),* 231-245.

Berry, D.T.R., Baer, R.A., & Harris, M.J. (1991). Detection of malingering on the MMPI: A meta analysis. *Clinical Psychology Review, 11,* 585-598.

Berry, D.T.R., Lamb, D.G., Wetter, M.W., Baer, R.A., & Widiger, T.A. (1994). Ethical considerations in the research on coached malingering. *Psychological Assessment, 6(1), 16-17.*

Bickart, W.T., Meyer, R.G., & Connell, D. (1991). The symptom validity technique as a measure of feigned short-term memory deficit. *American Journal of Forensic Psychology, 9(2),* 3-11

Binder, L.M. (1992). Forced-choice testing provides evidence of malingering. *Archives of physical medicine and Rehabilitation, 73,* 377-380.

Binder, L.M. (1993a). Portland Digit Recognition Test Manual (2nd ed.). Portland, Oregon: Author.

Binder, L.M. (1993b). Assessment of malingering after mild head trauma with the Portland Digit Recognition Test. *Journal of Clinical and Experimental Neuropsychology, 15(2),* 170-182.

Binder, L.M. (1993c). An abbreviated form of the Portland Digit Recognition Test. *The Clinical Neuropsychologist, 7(1),* 104-107.

Binder, L.M., & Johnson-Greene, D. (1995). Observer effects on neuropsychological performance: A case report. *The Clinical Neuropsychologist, 9(1),* 74-78.

Binder, L.M., & Pankratz, L. (1987). Neuropsychological evidence of factitious memory complaint. *Journal of Clinical and Experimental Neuropsychology, 9,* 167-171.

Binder, L.M., & Thompson, L.L. (1994). The ethics code and neuropsychological assessment practices. *Archives of Clinical Neuropsychology, 10(1),* 27-46.

Binder, L.M., & Willis, S.C. (1991). Assessment of motivation after financially compensable minor head trauma. *Psychological Assessment, 3(2),* 175-181.

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

Brandt, J. (1988). Malingered Amnesia. In R. Rogers (Ed.), *Clinical Assessment of Malingering and Deception.* (pp. 65-83). New York: The Guilford Press

Butcher, J., Dahlstrom, W. G., Graham, J., Tellegen, A., & Kaemmer, B. (1989). *Minnesota Multiphasic Personality Inventory (MMPI-2): Manual for administration and scoring.* Minneapolis, MN: University of Minnesota.

Butcher, J., & Williams, C. (1992). *Essentials of MMPI-2 and MMPI-A interpretation.* Minneapolis, MN: University of Minnesota.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences, 2nd Ed.* Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Dannenbaum, S.E., & Lanyon, R.I. (1993). The use of Subtle items in detecting deception. *Journal of Personality Assessment, 61(3),* 501-510.

Elwood, R.W. (1993). Clinical discrimination and neuropsychological tests: An appeal to Bayes' Theorem. *The Clinical Neuropsychologist, 7(2),* 224-233.

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

Franzen, M.D., & Iverson, G.L. (1995). Biased responding: The detection of Neuropsychological Malingering in a Hospital Setting. *Advances in Medical Psychotherapy, 8,* 47-58.

Fredrick, R.I., & Foster, H.G., Jr. (1991). Multiple measures of malingering on a forced-choice test of cognitive ability. *Psychological Assessment, 3(4),* 596-602.

Fredrick, R.I., Sarfaty, S.D., Johnston, D., & Powel, J. (1994). Validation of a detector of response bias on a forced-choice test of nonverbal ability. *Neuropsychology, 8(1),* 118-125.

Gough, H.G. (1950). The F minus K dissimulation index for the MMPI. *Journal of Consulting Psychology, 14,* 408-413.

Graham, J.R., Watts, D., & Timbrook, R.E. (1991). Detecting fake-good and fake-bad MMPI-2 profiles. *Journal of Personality Assessment, 57(2),* 264-277.

Greene, R. (1991). *The MMPI-2/MMPI-A: An interpretive manual.* Boston, MA: Allyn and Bacon, Inc.

Graves, R.E., & Bradley, R. (1992). Millisecond timing on the IBM PC/XT/AT PS/2: A review of the options and corrections for the Graves and Bradley Algorithm. *Behavior Research Methods, Instruments, & Computers, 23(3),* 377-379

Greiffenstein, Baker, W.J., & M.F., Gola, T. (1994). Validation of malingered amnesia measures with a large clinical sample. *Psychological Assessment, 6(3),* 218-224.

Greiffenstein, M.F., Gola, T., & Baker, W.J. (1995). MMPI-2 validity scales versus domain specific measures in detection of factitious traumatic brain injury. *The Clinical Neuropsychologist, 9(3),* 230-240.

Grosz, H.J., & Zimmerman, J. (1965). Experimental analysis of hysterical blindness. *Archives of General Psychiatry, 13,* 256-260.

Grouvier, W.D., Prestholdt, P., & Warner, M. (1988). A survey of common misconceptions about head injury and recovery. *Archives of Clinical Neuropsychology, 3,* 331-343.

Guilmette, T.J., Hart, K.J., & Giuliano, A.J. (1993). Malingering detection: The use of a forced-choice method in identifying organic versus simulated memory impairment. *The Clinical Neuropsychologist, 7(1),* 59-69.

Guilmette, T.J., Hart, K.J., Giuliano, A.J., & Leininger, B.E. (1994). Detecting simulated memory impairment: Comparison of the Rey Fifteen-Item Test and the Hiscock Forced-Choice Procedure. *The Clinical Neuropsychologist, 8(3),* 283-294.

Hathaway, S.R., & McKinley, J.C. (1983). *The Minnesota Multiphasic Personality Inventory Manual.* New York: Psychological Corporation.

Haughton, P.M., Lewesly, A., Wilson, M. & Williams, R.G. (1979). A forced-choice procedure to detect feigned or exaggerated hearing loss. *British Journal of Audiology, 13,* 135-138.

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

Helmes, E., & Reddon, J.R. (1993). A perspective on developments in assessing psychopathology: A critical review of the MMPI and MMPI-2. *Psychological Bulletin, 113(3),* 453-471.

Hiscock, C.K., Branham, J.D., & Hiscock, M. (1993). Detecting feigned cognitive impairment : The two-alternative forced-choice model compared with conventional tests. Paper presented at the annual meeting of the International Neuropsychological Society, Galveston, Texas.

Hiscock, M., & Hiscock, C.K. (1989). Refining the forced choice method for the detection of malingering. *Journal of Clinical and Experimental Neuropsychology, 11(6),* 967-974.

Holden, R.R. (1195). Response latency detection of fakers on personnel tests. *Canadian Journal of Behavioral Science, 27(3),* 343-355.

Holden, R.R., & Kroner, D.G. (1992). Relative efficacy of differential response latencies for detecting faking on a self-report measure of psychopathology. *Psychological Assessment, 4(2),* 170-173.

Iverson, G.L. (1993). *Detecting simulated memory deficits with the Recognition Memory Test: Evaluation of alternate scores.* Poster presented at the 21st annual meeting of the International Neuropsychological Society, Galveston, Texas, February.

Iverson, G.L. (1995). Qualitative aspects of malingered memory deficits. *Brain Injury, 9(1),* 35-40.

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

Iverson, G.L., & Franzen, M.D. (1996). Using multiple objective memory procedures to detect simulated malingering. *Journal of Clinical and Experimental Neuropsychology, 18(1),* 38-51.

Iverson, G.L., Franzen, M.D., & McCracken, L.M. (1991). Application of a forced-choice memory procedure designed to detect experimental malingering. *Archives of Clinical Neuropsychology, 15(6),* 667-676.

Iverson, G.L., Franzen, M.D., & McCracken, L.M. (1991). Evaluation of an objective technique for the detection of malingered memory deficits. *Law and Human Behavior, 15(6),* 667-676.

Knight, J.A., & Meyers, J.E. (1995). *Comparison of malingered and brain-injured productions on the Rey Osterrieth Complex Figure Test.* Poster presented at the 23th annual meeting of the International Neuropsychological Society, Seattle, Washington, 14-17, February.

Lamb, D.G., Berry, D.T.R., Wetter, M.W., & Baer, R.A. (1994). Effects of two types of information on malingering of closed-head injury on the MMPI-2. *Psychological Assessment, 6(1), 8-13.*

Lees Haley, P.R. (1991). MMPI-2 F and F-K scores of personal injury malingerers in vocational neuropsychological and emotional distress claims. *American Journal of Forensic Psychology, 9(3),* 5-14.

Lees Haley, P.R. (1992). Efficacy of MMPI-2 validity scales and MCMI-II modifier scales for detecting spurious PT*SD* claims: F, F-K, Fake Bad Scale, Ego Strength, Subtle-obvious subscales, DIS and DEB. *Journal of Clinical Psychology, 48(5),* 681-689.

Lees Haley, P.R., & Dunn, J.T. (1994). The ability of naïve subjects to report symptoms of mild brain injury, post-traumatic stress disorder, major depression, and generalized anxiety disorder. *Journal of Clinical Psychology, 50(2),* 252-256.

Lees-Haley, P.R., & Fox, D. (1990). MMPI Subtle-Obvious scales and malingering: Clinical versus simulated scores. *Psychological Reports, 66,* 907-911.

Lees-Haley, P.R., English, L. T., & Glenn, W. J. (1991) A fake bad scale on the MMPI-2 for personal injury claimants. Psychological Reports, 68, 203-210.

Lindeboom, J. (1989). Who needs cutting points? *Journal of Clinical Psychology, 45(4),* 679-683.

Martin, R.C., Bolter, J.F., Todd, M.E., Gouvier, W.D., & Niccolls, R. (1993). Effects of sophistication and motivation on the detection of malingered memory performance using a computerized forced-choice task. *Journal of Clinical and Experimental Neuropsychology, 15(6),* 867-880.

Miller, E. (1968). A note on the visual performance of a subject with unilateral functional blindness. *Behavior Research and Therapy, 6,* 115-116.

Miller, E. (1986). Detecting hysterical sensory symptoms: An elaboration of the forced choice technique. *British Journal of Clinical Psychology, 25,* 231-232.

Millis, S.R., & Kler, S. (1995). Limitations of the Rey Fifteen-Item Test in the detection of malingering. *The Clinical Neuropsychologist, 9(3),* 241-244.

Nies, K.J., & Sweet, J.L. (1994). Neuropsychological assessment and malingering: A critical review of past and present strategies. *Archives of Clinical Neuropsychology, 9(6),* 501-552.

Palmer, B.W., Boone, K.B., Allman, L., & Castro, D.B. (1995). Co-occurrence of brain lesions and cognitive deficit exaggeration. *The Clinical Neuropsychologist, 9(1),* 68-73.

Pankratz, L. (1979). Symptom validity testing and symptom retraining: Procedures for the assessment and treatment of functional sensory deficits. *Journal of Consulting and Clinical Psychology, 47,* 409-410.

Pankratz, L. (1983). A new technique for the assessment and modification of feigned memory deficit. *Perceptual and Motor Skills, 57,* 367-372.

Pankratz, L. (1988). Malingering on intellectual and neuropsychological measures. In R. Rogers (Ed.). *Clinical Assessment of Malingering and Deception,* (pp. 183-192). New York: The Guilford Press.

Pankratz, L., & Erickson, R.C. (1990). Two views of malingering. *The Clinical Neuropsychologist, 4(4),* 379-389.

Pankratz, L., Fausti, S.A., & Peed, S. (1975). A forced-choice technique to evaluate deafness in the hysterical malingering patient. *Journal of Consulting and Clinical Psychology, 43,* 421-422.

Portland Digit Recognition Test - Computerized: Measuring response latency improves the detection of malingering. *The Clinical Neuropsychologist, 9(2),* 124.134.

Prigatano, G.P., & Amin, K. (1993). Digit Memory Test: Unequivocal cerebral dysfunction and suspected malingering. *Journal of Clinical and Experimental Neuropsychology, 15(4),* 537-546

Puri, M.L., & Sen, P.K. (1971). *Nonparametric methods in multivariate analysis.* New York: Wiley.

Richert, H.S., Hiscock, M., & Caroselli, J.S. (1996). *The emperor's clothes phenomenon: Implications for detection of malingering.* Poster presented at the 24th annual meeting of the International Neuropsychological Society, Chicago, Illinois, 14-17, February.

Rogers, R. (1988a). Current status of clinical methods. In R. Rogers (Ed.). *Clinical Assessment of Malingering and Deception,* (pp. 285-307). New York: The Guilford Press.

Rogers, R. (1988b). Researching dissimulation. In R. Rogers (Ed.). *Clinical Assessment of Malingering and Deception,* (pp. 311-327). New York: The Guilford Press.

Roger, R. (1990). Development of a new classificatory model of malingering. *Bulletin of the American Academy of Psychiatry and Law, 18(3),* 323-333.

Rogers, R., & Cavanaugh, J.L. (1983). "Nothing but the truth"... a reexamination of malingering. *The Journal of Psychiatry and Law, Winter,* 443-459.

Rose, F.E., Hall, S., & Szalda-Pertree, A.D. (1995). Portland Digit Recognition Test – Computerized: Measuring response latency improves the detection of malingering. *The Clinical Neuropsychologist, 9(2),* 124-134.

Rosenfeld, P.J., Ellwanger, J., & Sweet, J. (1995). Detecting simulated amnesia with event-related brain potential. *International Journal of Psychophysiology, 19,* 1-11

Rothke, S.E., Friedman, A.F., Dahlstrom, W.G., Greene, R.L., Arredondo, R., & Mann, A.W. (1994). MMPI-2 normative data for the F-K index: Implications for clinical, neuropsychological, and forensic practice. *Assessment, 1(1),* 1-15.

Russell, M.L., Spector, J., & Kelly, M. (1993). Primacy and recency effects in the detection of malingering using the WMS-R Logical Memory Subtests. Poster presented at the 21st annual meeting of the International Neuropsychological Society. February, 1993, Galveston, TX.

Segalowitz, S.J., & Graves, R.E. (1990). Suitability of the IBM XT, AT, and PS/2 keyboard, mouse, and gameport as response devices in reaction time paradigms. *Behavior Research Methods, Instruments, & Computers, 22(3),* 283-289

Slick, D., Hopp, G., & Strauss, E. (1992). *The Victoria Revision of the Hiscock Digit Memory Test.* Victoria, British Columbia: Author.

Slick, D., Hopp, G., & Strauss, E. (1995). *The Victoria Symptom Validity Test.* Odessa, Florida: PAR.

Slick, D., Hopp, G., Strauss, E., Hunter, M., & Pinch, D. (1994). Detecting dissimulation: Profiles of simulated malingerers, traumatic brain-injury patients, and normal controls on a revised version of Hiscock and Hiscock's forced-choice memory test. *Journal of Clinical and Experimental Neuropsychology 16(3).*

Spreen, O., & Strauss, E. (1991). *A Compendium of neuropsychological tests: Administration norms and commentary.* New York: Oxford University Press.

Strauss, E., Spellacy, F., Hunter, M., & Berry, T. (1995). Assessing believable deficits on measures of attention and information processing capacity. *Archives of Clinical Neuropsychology.*

Tenhula, W.N., & Sweet, J.J. (1996). Double cross-validation of the Booklet Category Test in detecting malingered traumatic brain injury. *The Clinical Neuropsychologist, 10(1),* 104-116.

Theodor, L.H., & Mandelcorn, M.S. (1973). Hysterical blindness: A case report and study using a modern psychological technique. *Journal of Abnormal Psychology, 92,* 552-3.

Wiener, D.N. (1948). Subtle and obvious keys for the MMPI. *Journal of Consulting Psychology, 12,* 164-170.

Wiggins, E.C., & Brandt, J. (1988). The detection of simulated amnesia. *Law and Human Behavior, 12(1),* 57-78.

Youngjohn, J.R., Burrows, L., & Erdal, K. (1995). Brain damage or compensation neurosis? The controversial post-concussive syndrome. *The Clinical Neuropsychologist, 9(2),* 112-123.

Appendix I: Formulas

Binomial probability calculation

$$z = \frac{C - (N * p)}{\sqrt{N * p * q}} = \frac{C - (N * 0.5)}{\sqrt{N * 0.5 * 0.5}} = \frac{C - (N * 0.5)}{\sqrt{N * 0.25}}$$

Where:

N = total number of responses
C = total number of correct responses
p = random chance probability of success per item, or 0.5
q = 1 - p, or 0.5

Bayes' Theorem for determining the
probability of group membership (classification)

$$\hat{p}(A) = \frac{p' \cdot LR_x}{1 - p' + (p' \cdot LR_x)}$$

Where:

$\hat{p}$ = estimated confidence of classification as A (e.g., malingerer)

p = prior probability or base rate of A

LR_x = likelihood ratio of score X (ordinate for score X from a sample of persons known to be A divided by the ordinate for score X from a distribution of a sample known to be *not* A) – see Lindeboom (1989) for additional information

Appendix II: Binomial z values and one-tailed p values

N correct	Max. N = 48		Max. N = 24		Max. N = 16		Max. N = 8	
	z	*p*	*z*	*p*	*z*	*p*	*z*	*p*
0	-6.930	<.0000001	-4.900	.0000005	-4.000	.00003	-2.830	.002
1	-6.640	<.0000001	-4.490	.000004	-3.500	.0002	-2.120	.017
2	-6.350	<.0000001	-4.080	.00002	-3.000	.001	-1.410	.079
3	-6.060	<.0000001	-3.670	.0001	-2.500	.006	-.710	.239
4	-5.770	<.0000001	-3.270	.0005	-2.000	.023	.000	.500
5	-5.490	<.0000001	-2.860	.002	-1.500	.067		
6	-5.200	.0000001	-2.450	.007	-1.000	.159		
7	-4.910	.0000005	-2.040	.021	-.500	.309		
8	-4.620	.000002	-1.630	.052	.000	.500		
9	-4.330	.000007	-1.230	.110				
10	-4.040	.00003	-.820	.206				
11	-3.750	.00009	-.410	.341				
12	-3.460	.0003	.000	.500				
13	-3.180	.0007						
14	-2.870	.002						
15	-2.600	.005						
16	-2.310	.010						
17	-2.020	.022						
18	-1.730	.042						
19	-1.440	.075						
20	-1.160	.123						
21	-.870	.192						
22	-.580	.281						
23	-.290	.386						
24	.000	.500						

Note. *p* values in the table are one-tailed, providing significance level for testing for divergence from chance-level performance. For example, a patient gets a score of 10 out of 24 correct (or z = -.82) on hard items. The *p* value associated with this level of performance is .207, indicating that a randomly responding participant would be expected to score this low approximately 21 times if tested 100 times. Conversely, a randomly responding participant would be expected to score higher (i.e., get more than 10 out of 24 correct) 79 times if tested 100 times. The *p* values in the table are symmetric for positive values of z and a one-tailed test for performance significantly above chance level.

Appendix III: Item list (difficult items shaded)

	Block 1			Block 2			Block 3		
Item #	Target	Left	Right	Target	Left	Right	Target	Left	Right
1	58730	92149	58730	45238	45238	17096	42365	42635	42365
2	27491	27491	63585	23601	26301	23601	52347	52347	53247
3	96451	96451	94651	22877	22877	10935	89547	23021	89547
4	85952	31403	85952	76159	76519	76159	96902	96902	83581
5	61956	61956	80374	95122	63407	95122	82684	82864	82684
6	63070	60370	63070	29730	27930	29730	74316	74316	73416
7	91327	93127	91327	13980	24745	13980	25083	25083	19967
8	68462	53171	68462	61748	61748	61478	12302	12032	12302
9	45219	45219	42519	20249	20249	76113	22105	22105	22015
10	60082	60082	75454	71224	65093	71224	29722	29722	44305
11	67458	67548	67458	73879	73879	78379	99110	85227	99110
12	31472	31742	31472	34709	34709	37409	78924	78924	79824
13	61885	61885	32797	18891	18891	24603	36847	52109	36847
14	16089	52744	16089	55782	34609	55782	61429	64129	61429
15	20185	20185	20815	29341	29431	29341	84140	53226	84140
16	61397	61397	61937	90612	90612	90162	39293	39293	51764