Journal of Clinical and Experimental Neuropsychology
2001, Vol. 23, No. 6, pp. 809–828

1380-3395/01/2306-809$16.00

Assessment of Response Bias in Mild Head Injury: Beyond Malingering Tests

Scott R. Millis[1] and Chris T. Volinsky[2]
[1]Kessler Medical Rehabilitation Research and Education Corporation, University of Medicine and Dentistry of New Jersey, Medical School, West Orange, NJ, USA, and [2]Shannon Laboratory, AT & T Labs – Research, Florham Park, NJ, USA

ABSTRACT

The evaluation of response bias and malingering in the cases of mild head injury should not rely on a single test. Initial injury severity, typical neuropsychological test performance patterns, preexisting emotional stress or chronic social difficulties, history of previous neurological or psychiatric disorder, other system injuries sustained in the accident, preinjury alcohol abuse, and a propensity to attribute benign cognitive and somatic symptoms to a brain injury must be considered along with performances on specific measures of response bias. This article reviews empirically-supported tests and indices. Use of the likelihood ratio and other statistical indicators of diagnostic efficiency are demonstrated. Bayesian model averaging as a statistical technique to derive optimal prediction models is performed with a clinical data set.

Paralleling the increased interest in mild traumatic brain injury (TBI) and use of neuropsychological evidence in the courtroom, numerous comprehensive reviews of the assessment of response bias and malingering of neuropsychological impairment have appeared in the literature over the past decade (Etcoff & Kampfer, 1996; Iverson & Binder, 2000; Millis & Putnam, 1996; Nies & Sweet, 1994; Rogers, Harrell, & Liff, 1993). Taking the next step in integrating the quickly-expanding literature, Slick, Sherman, and Iverson (1999) recently presented diagnostic criteria for 'malingered neurocognitive dysfunction (MND)' that are relevant in the assessment of mild TBI. These diagnostic criteria represent a significant contribution to the field because they present a systematic and coherent set of diagnostic guidelines based on empirical findings. Slick et al. (1999) define MND as "the volitional exaggeration of cognitive dysfunction for the purpose of obtaining substantial material gain, or avoiding or escaping formal duty of responsibility" (p. 552).

Recognizing that there are various levels of diagnostic certainty, Slick et al. (1999) proposed separate criteria for 'definite,' 'probable,' and 'possible' malingered neurocognitive disorder. Except in rare cases, persons who are feigning cognitive impairment will not disclose this fact. Consequently, we do not have the idealized diagnostic 'gold standard' typically used to derive estimates of prevalence and other diagnostic parameters. However, this lack of a diagnostic gold standard is actually rather common in medicine and epidemiology, and not limited to neuropsychology. Joseph, Gyorkos, and Coupal (1995) note, "In fact, one may argue that this is virtually always the situation, since few tests are considered to be 100% accurate" (p. 262). One response to this diagnostic dilemma, used by Slick et al. (1999), is not to rely on 'malingering' tests alone. They recommend using multiple

Address correspondence to: Scott R. Millis, Kessler Medical Rehabilitation Research and Education Corporation, 1199 Pleasant Valley Way, West Orange, NJ 07052, USA. Tel.: +1-973-243-6976. Fax: +1-973-243-6990. E-mail: smillis@kmrrec.org
Accepted for publication: August 1, 2001.

sources of data when considering the diagnosis of MND. Assignment to 1 out of the 3 MND categories is based on 4 different sets of criteria: (a) presence of a substantial external incentive; (b) evidence from neuropsychological testing; (c) evidence from self-report; and (d) behaviors meeting necessary criteria from groups B or C that are not fully accounted for by psychiatric, neurological, or developmental factors. Another response to this diagnostic challenge that does not exclude the approach advocated by Slick et al. (1999) is to use mathematical methods to quantify aspects of uncertainty in prediction – an approach demonstrated in later sections of this paper.

Although MND criteria are the proposed guidelines, it is our opinion that they represent the current state-of-the-art. Our aim is to supplement these guidelines. We do not believe that a single neuropsychological test in isolation is capable of accurately diagnosing *any* condition, be it brain dysfunction or malingered neurocognitive disorder. As Iverson and Binder (2000) have emphasized, response bias on a test is not tantamount to malingering: "To diagnose malingering, the clinician must infer that the negative response bias was designed to achieve some identifiable incentive" (pp. 831–832). Contextual variables such as initial injury severity, time postinjury, premorbid and comorbid factors, and environmental contingencies are needed to interpret any neuropsychological test score meaningfully.

We begin with a brief discussion of these contextual factors. We then selectively highlight psychometric advances in the detection of malingering. Test scores provide evidence in support of a hypothesis (in this case, a diagnosis) when combined with prior information. We then demonstrate the explicit, quantitative use of prevalence rates, or prior probabilities, with tests as a way to evaluate the strength of neuropsychological evidence for a particular diagnosis. We conclude with a discussion and application of a newly developed statistical approach that shows considerable promise in improving prediction and diagnostic accuracy.

CONTEXTUAL FACTORS

A starting point in the assessment of response bias is to establish the patient's initial injury severity and time postinjury. Injury characteristics assist the clinician in placing a patient's test scores in proper context. Larrabee (1990) has termed the consistency between test performance and injury severity as 'severity indexing' or 'referencing.' There exists a dose-response relationship between the length of coma and the degree of cognitive impairment (Dikmen, Machamer, Winn, & Temkin, 1995), i.e., one expects greater cognitive impairment as the severity of brain injury increases. For example, one would not anticipate a patient with a Glasgow Coma Scale (GCS) score of 15 examined at 12 months postinjury to produce neuropsychological test scores similar to the findings from a patient with an initial GCS score of 5 examined at 2 months postinjury. Differential diagnosis takes on central importance whenever there is an inconsistency between initial injury severity and level of neuropsychological test performance. Several studies provide data on typical neuropsychological test performances of persons from different backgrounds with varying levels of TBI severity that can assist the neuropsychologist in determining whether an individual's neuropsychological test profile is atypical (Dikmen et al.,1995; Kreutzer, Gordon, Rosenthal, & Marwitz, 1993; Levin et al., 1987; Ponsford et al., 2000).

Although the issue of long-term neuropsychological impairment following mild TBI has generated controversy, the preponderance of empirical evidence at this time does not support an association between chronic, severe neuropsychological impairment and uncomplicated mild TBI. Studies of mild TBI that have included control groups have found that neuropsychological deficits attributed to brain injury generally resolve within 1–3 months postinjury (Dikmen, McLean, & Temkin, 1986; Gentilini et al., 1985; Levin et al., 1987; Ponsford et al., 2000). Studies with findings to the contrary have been hampered by a number of methodological flaws, including inconsistency or inaccuracy in the classification of brain injury severity, enrollment of participants on the basis of symptoms rather than history of brain injury, failure to control pre-existing conditions, or lack of appropriate control groups (Dikmen & Levin, 1993).

The study by Ponsford et al. (2000) is notable for its inclusion of a trauma control group and consecutive enrollment of subjects on the basis of injury. At 1 week postinjury, the mild TBI participants performed worse on complex attention tasks compared to trauma controls. Interestingly, the TBI group's mean performance was *superior* to the control group's mean on a memory test, the Rey Auditory Verbal Learning Test. At 3 months, there were no statistically significant differences between the groups on neuropsychological measures. In a separate study that investigated outcome at a longer interval, Dikmen et al. (1995) found that performance of their mild TBI group was indistinguishable from the performance of the trauma control subject group on the Halstead–Reitan Neuropsychological Test Battery and other procedures at 1 year postinjury.

In examining the relationship between mild TBI and cognitive impairment more broadly, Binder, Rohling, and Larrabee (1997) conducted a meta-analysis of prospective studies of mild TBI with a minimum of 3 months follow-up and an attrition rate of less than 50%. In the 11 samples that were located, Binder et al. (1997) included 314 mild TBI patients and 308 control subjects in their analysis. The overall effect size of mild TBI on neuropsychological test performance was small ($g = .07$ and $d = .12$). In other words, the mild TBI group's mean was shifted one-eighth of a standard deviation below the control group mean. Using the Wechsler Adult Intelligence Scale – Revised (WAIS – R) metric, this reflects a change of about 2 points.

Despite the favorable prognosis for a single, uncomplicated mild TBI, there is a subset of persons who report cognitive difficulties and somatic symptoms that extend beyond 90 days after the accident (Ruff, Camenzuli, & Mueller, 1996; Wrightson & Gronwall, 1981). The question of malingering is most often raised in this group because their outcome is atypical, the severity of their symptoms and claimed disability may be disproportionate to the initial injury severity, and civil litigation is frequently the context in which the neuropsychologist examines the patient. The role of litigation cannot be ignored because the association between financial incentives and neuropsychological test results appears significant. In a meta-analysis of 18 studies containing 2,353 subjects, Binder and Rohling (1996) found a moderate overall effect size of 0.47. This does not imply that all persons involved in civil litigation or disability proceedings are malingering. Yet, the presence of external incentives may be a risk factor for biased responding in the neuropsychological examination that certainly must be considered. The magnitude of the effect size of mild TBI on neuropsychological test performance is substantially *lower* than that associated with financial incentives (0.07 vs. 0.47). The neuropsychologist faces an obvious and formidable diagnostic challenge in the chronic, symptomatic mild TBI case, sometimes termed the persistent postconcussive syndrome (PPCS; Alexander, 1995). Although physiological processes may account for the acute symptoms of mild TBI, there does not appear to be a single cause for the protracted symptoms characterizing PPCS (Alexander, 1995).

When attempting to discern and disentangle the factors that cause and maintain the symptomotology in the PPCS case, it is important to consider not only the initial injury severity but also premorbid and comorbid factors and current environmental contingencies. It is unlikely that the diagnostic possibilities are limited to only malingering versus brain dysfunction. Preexisting emotional stress or chronic social difficulties (Fenton, McClelland, Montgomery, MacFlynn, & Rutherford, 1993; Klonoff & Lamb, 1998; Ponsford et al., 2000), learning disability (Dicker, 1992), history of previous neurological or psychiatric disorder (Ponsford et al., 2000), other system injuries sustained in the accident (Dikmen et al., 1986), preinjury alcohol abuse (Dikmen & Levin, 1993), and a propensity to attribute benign cognitive and somatic symptoms to a brain injury (Mittenberg, DiGiulio, Perrin, & Bass, 1992) are but a handful of factors among many potential conditions, along with malingering and injury-related brain dysfunction, that may be responsible for PPCS.

TESTS, PROCEDURES, AND INDICES

Along with the need to consider the above contextual factors in the differential diagnosis, there are a variety of tests and indices that may be used to detect response bias, some of which have greater empirical support than others. A selective review of the tests and procedures follows. We chose tests having a minimum of one cross-validated study in a second independent group of participants following the initial derivation study. Other tests are highlighted if they appeared to have diagnostic promise on the basis of innovation or sample size used in test development.

Finding a statistically significant difference (i.e., $p < .05$) between a TBI group and malingering group on a response bias measure does not necessarily mean that the measure is diagnostically useful. A statistically significant difference simply indicates that the difference in the group means is unlikely to be zero. This is essentially an uninteresting finding from a diagnostic standpoint. Other statistical indicators are needed, such as effect size, likelihood ratio, or signal detection theory variables. However, there is considerable variability in the diagnostic efficiency statistics that are reported in studies. We will focus on effect sizes and sensitivity / specificity rates in our review because they are relatively easy to calculate from the available summary data reported in most published studies. We also illustrate the use of graphical techniques of violin plots and receiver operating characteristic (ROC) curves in our section on performance patterns to detect response bias. A separate section is devoted to the specific use of the likelihood ratio.

The effect size can be defined as ''the difference between two population means expressed in units of the standard deviation'' (Chow, 1996, p. 133). A large effect size implies that the two populations' score distributions on a given test or measure are far apart. Thus, the test may be diagnostically powerful in differentiating individuals from the two different populations. The effect size is also useful for comparing tests because it is expressed as a standardized metric.

To assist the reader to integrate findings from the following test review, Table 1 presents the individual tests and indices grouped thematically along with their associated effect sizes. We calculated effect sizes from the data reported in the published studies. The studies we reviewed generally employed one of the two research designs. In one design, persons with unequivocal evidence of brain injury were compared with persons with mild injuries whose performance on neuropsychological measures was marked by poor effort as determined by incongruously low test scores or below chance performance on symptom validity tests ('clinical comparison group'). A second type of design compared persons with brain injuries with persons instructed to 'fake' or malinger neuropsychological impairment ('analog comparison group').

Table 1. Effect Sizes of Selected Response Bias Measures and Indices: Traumatic Brain Injury Versus Response Bias/Incomplete Effort.

Test or Index	Effect size (g)	Comparison group
Forced-Choice Tests		
Portland Digit Recognition	0.98–1.21	Clinical
Hiscock Forced-Choice Procedure	2.36	Analog
	5.44	Clinical
Test of Memory Malingering	1.87	Clinical
Victoria Symptom Validity Test	1.06	Clinical
Word Memory Test	0.38–0.42	Clinical
Recognition Memory Test	0.90–1.28	Clinical
	2.80–4.62	Analog
Seashore Rhythm Test	1.33	Clinical
	0.67–1.09	Analog
Speech-sounds Perception Test	1.59	Clinical
	0.66–1.61	Analog
Floor effect		
Digit Span	1.97	Clinical
	0.92–1.02	Analog
Reliable Digit Span	1.75	Clinical
Vocabulary minus Digit Span	1.48	Clinical
Performance patterns		
Wechsler Adult Intelligence Scale – Revised discriminant function	2.10	Clinical
California Verbal Learning Test Hits	1.26–2.59	Clinical

Forced-Choice Tests

Forced-choice tests (FCTs), also known as symptom validity tests (SVTs), are among the earliest developed (Pankratz, Fausti, & Peed, 1975) and most extensively evaluated tests for the detection of response bias and malingering. A stimulus item such as a word, number, photograph, or line drawing is presented and then followed by a two-choice recognition task with the original stimulus item paired with a distractor or foil. The individual is asked to choose the target item. An individual with no memory for the stimuli will perform at chance level. Response bias can thus be assessed by comparing an individual's performance on the two-alternative FCT with the binomial distribution to determine the probability of obtaining a particular score. As scores decline below 50% correct, it is increasingly likely that the individual was deliberately choosing wrong answers. As Pankratz has suggested, "'motivated wrong answering' is the smoking gun of intent" (Pankratz & Erickson, 1990). Performance on one or more FCTs that is below chance at a statistically significant level ($p < .05$) is termed 'definite negative response bias' by the proposed MND diagnostic guidelines (Slick et al., 1999) and is considered to be "closest to an evidentiary 'gold standard' for malingering" (p. 551), excluding confession by the examinee. If the below chance performance cannot be accounted for by psychiatric, neurological, or developmental factors and there is an external incentive, the examinee meets the criteria for 'Definite Malingered Neurocognitive Disorder' (Slick et al., 1999). According to these proposed criteria, other psychometric tests or indices may be used in the diagnosis of Probable or Possible MND, but only below chance performance on a FCT meets the criteria of definite response bias needed for the diagnosis of Definite MND.

Of the digit recognition FCTs, the Hiscock Forced-Choice Procedure (HFCP; Hiscock & Hiscock, 1989) and Portland Digit Recognition Test (PDRT; Binder & Willis, 1991) have received the greatest empirical support to date. Individuals with psychiatric disorders or brain dysfunction obtained mean scores that ranged from 84 to 99% correct on the PDRT or HFCP (Binder, 1993; Guilmette, Hart, & Giuliano, 1993; Guilmette, Hart, Giuliano, & Leininger, 1994; Prigatano & Amin, 1993). In contrast, clinical subjects suspected of malingering scored 56–74% correct on the PDRT or HFCP (Binder, 1993; Greiffenstein, Baker, & Gola, 1994; Prigatano & Amin, 1993). Analog malingering participants obtained mean scores that ranged from 53 to 60% correct on these measures (Binder & Willis, 1991; Guilmette et al., 1993).

Although a performance on a FCT that is below chance is persuasive evidence of response bias, the majority of malingerers will not perform below chance. In surveying analog studies in which subjects were instructed to malinger cognitive deficits, Hiscock, Branham, and Hiscock (1994) found that no greater than 34% of the cases demonstrated below chance performance. A strategy to address this shortcoming has been to select cut-off scores for FCTs that are above chance that have acceptable diagnostic efficiency. Often, the cut-off score is one that few if any, persons with brain dysfunction perform below. For example, scores less than 54–63% correct on the PDRT may indicate poor effort (Binder, 1993; Greiffenstein et al., 1994) while a cut-off of 90% or less has been used with the HFCP (Guilmette, Hart, Giuliano, & Leininger, 1994). Depending on the sample and cut-off score, the PDRT and HFCP have been reported to accurately classify 89–100% of persons with TBI and 75–90% of suspected clinical malingerers or analog simulators.

Other FCTs that show promise include the Computerized Assessment of Response Bias (CARB; Conder, Allen, & Cox, 1992), the Test of Memory Malingering (TOMM; Tombaugh, 1997), the Victoria Symptom Validity Test (VSVT; Slick, Hopp, Strauss, & Thompson, 1997), and the Word Memory Test (WMT; Green, Allen, & Astner, 1996). The CARB is a computerized digit recognition task that has been used to examine litigants with a wide range of brain injury severity. Green and Iverson (in press) found that litigants with mild TBI performed worse and had longer latencies on the CARB than litigants with moderate and severe TBI.

The TOMM is a visual recognition task that uses 50 drawings of common objects in 3 trials. Mean performances across groups of persons with a variety of neurological disorders ranged from 91.4 to 98.6% correct on Trial 2 (Tombaugh, 1997). To date, most validation studies of the TOMM have employed an analog design in which college students have been instructed to malinger (Rees, Tombaugh, Gansler, & Moczynski, 1998). However, in a small clinical study, a litigating mild TBI group obtained statistically significant lower scores on Trial 2 than the non-litigating TBI group. The effect size was very large ($g = 1.87$) (Rees et al., 1998).

The VSVT is a 48-item computerized version of the digit recognition FCT paradigm that uses both correct number and response latency to assess response bias. Although the initial normative sample was small, additional empirical support for the VSVT with larger samples has appeared (Doss, Chelune, & Naugle, 1999). Doss et al. (1999) found that their TBI compensation-seeking group had disproportionately more individuals scoring in the questionable / invalid range on the VSVT than a group of non-compensation-seeking patients. The effect size was large, $g = 1.06$.

In the WMT, the individual is presented with 20-word pairs, which are followed by immediate and delayed recognition trials (Green et al., 1996). The subject's task is to select the original words from target–foil pairs. Thus far, WMT has been applied primarily to litigated cases of mild to severe TBI (Green, Iverson, & Allen, 1999; Iverson, Green, & Gervais, 1999). In this series of litigated cases, the mild injury group showed a greater degree of biased responding as measured by the WMT with medium effect sizes ($g = .38$–$.43$). This may have been an especially rigorous test of the WMT because of the possibly low prevalence of malingering in the sample. Many studies have typically included equal numbers of participants in their groups or have had inclusion criteria increasing the likelihood of more extreme forms of malingering, and, thus, maximized the performance of the detection measure.

Standard neuropsychological measures with a forced-choice format have also been used to detect response bias. The Recognition Memory Test (RMT; Warrington, 1984) differentiated persons with moderate to severe TBI from litigating mild head injury claimants (Millis, 1992; Millis & Putnam, 1994) with overall correct classification ranging from 76 to 83% with associated large effect sizes ($g = 0.90$–1.28). Similar patterns with even larger effect sizes ($g = 2.80$–4.62) were found in an analog malingering study with the RMT by Iverson and Franzen (1994). Other standard measures having a forced-choice format that appear useful in detecting response bias include the Seashore Rhythm Test (SRT) and Speech-sounds Perception Test (SSPT) (Gfeller & Cradock, 1998; Goebel, 1983; Heaton, Smith, Lehman, & Vogt, 1978; Mittenberg, Rotholc, Russell, & Heilbronner, 1996; Millis, Putnam, & Adams, 1996; Trueblood & Schmidt, 1993). Based on these studies, errors in excess of 17 on the SSPT or 8 on the SRT in the litigated mild TBI case should raise the question of response bias.

Floor Effect

Some neuropsychological tasks are reasonably easy for persons with severe TBI. For example, forward digit span is often relatively intact in a variety of neurological disorders. In contrast, a strikingly poor forward span may signal response bias (Binder & Willis, 1991; Greiffenstein, Baker & Gola, 1994; Greiffenstein, Gola, & Baker, 1995; Iverson & Franzen, 1994; Meyers & Volbrecht, 1998; Mittenberg, Azrin, Millsaps, & Heilbronner, 1993; Mittenberg, Theroux-Fichera, Zielinski, & Heilbronner, 1995; Trueblood, 1994; Trueblood and Schmidt, 1993). The Digit Span (DS) subtest from the WAIS – R or Wechsler Memory Scale – Revised has been used most frequently in these studies. Large effect sizes were associated with response bias in analog designs, ($g = 0.92$–1.02) (Heaton et al., 1978; Mittenberg et al., 1995) and clinical designs, ($g = 1.97$) (Millis, Ross, & Ricker, 1998). In these studies, the response bias groups' mean performances on the DS were substantially lower than the mean performances generated by persons with brain injury.

A second method developed by Greiffenstein et al. (1994) uses the longest number of digits repeated accurately on both trials of the DS (forward plus backward) and is termed Reliable Digit Span (RDS). In comparing a group of persons with TBI with a group of probable clinical malingerers on the RDS, Greiffenstein et al. (1994) found a large effect size, ($g = 1.75$).

Using a score of seven or less to indicate response bias, Meyers and Volbrecht (1998), in a cross-validation study, found that 96% of their non-litigating TBI participants were correctly classified. Of their litigating mild TBI participants, who failed a separate FCT, 78% were correctly classified by the RDS.

A third approach has been to examine DS in relationship to the WAIS – R Vocabulary subtest. Mittenberg et al. (1995) reported that analog malingerers showed reductions on DS relative to Vocabulary, while persons with TBI showed similar levels of performance on both subtests. A Vocabulary-Digit Span discrepancy of two or more age-corrected scaled score points correctly classified 71% of their sample. In a study of probable clinical malingerers and persons with moderate to severe TBI, Millis et al. (1998) found that the same cut-off score yielded a correct classification rate of 79%.

Performance Patterns

Response bias may be a complex set of behaviors that is inadequately assessed by a single test. Attempts have been made to improve diagnosis by developing multivariable composites. An early investigation by Heaton et al. (1978) used discriminant analysis with the Halstead – Reitan Neuropsychological Test Battery (HRB) that contrasted the performance pattern of head-injured patients with control subjects instructed to feign neuropsychological impairment. The study, however, was hampered by statistically over-fitting the model with an inadequate subject-to-variable ratio.

In a later study of 80 participants with TBI and 80 normal volunteers instructed to simulate cognitive impairment, Mittenberg et al. (1996) used stepwise discriminant function analysis and derived a function containing 10 HRB variables that correctly classified 89% of the cases. When applied to a sample of VA patients with a history of TBI, the discriminant function correctly classified 78% of the sample (McKinzey & Russell, 1997).

In an earlier study, Mittenberg et al. (1995) reported that a 7-subtest WAIS – R discriminant function accurately classified 79% of head-injured patients and normal subjects instructed to malinger. Millis et al. (1998) cross-validated the function on a sample of 100 participants; 92% of persons with moderate to severe TBI were correctly classified, as were 88% of mild TBI litigants who had performed within chance on the RMT. Axelrod and Rawlings (1999) later applied the same function to patients with TBI who had received the WAIS – R 2– 4 times over the course of 1 year postinjury. Rates of correct classification ranged from 76 to 93% and were independent of the level of cognitive performance. No practice effect was found.

The diagnostic efficiency of this WAIS – R discriminant function can be evaluated from other perspectives as well. A large effect size ($g = 2.10$) was associated with the mean difference on this function between the TBI group and incomplete effort group in a study by Millis et al. (1998), which estimated that the group means were separated by about 2 SD (Table 1). This separation in the scores distributions on the discriminant function can be graphically portrayed by violin plots in Figure 1 (Hintze & Nelson, 1998). Violin plots combine the summary statistics of a box plot with the graphical information given by a local density estimator to show the distributional structure of the discriminant function scores in the different samples. Although there is some overlap in the groups' overall score distributions, there is complete separation in the interquartile ranges. The ROC curve in Figure 2 provides additional diagnostic information. If a diagnostic test has no utility, its Area under Curve (AUC) would be .50 and its plot would not depart from the 45° line. As a test's diagnostic utility increases, its AUC approaches 1.0. In the Millis et al. sample, the WAIS – R discriminant function has excellent diagnostic utility with an AUC that exceeds .90 (AUC = .94). The AUC also indicates that a randomly selected individual from the litigation group had a larger WAIS – R discriminant function value than a randomly chosen person from the TBI group, 94% of the time in the Millis et al. sample.

The California Verbal Learning Test (CVLT; Delis, Kramer, Kaplan, & Ober, 1987) has been used by numerous investigators to detect response bias. Millis, Putnam, Adams, and Ricker (1995a)

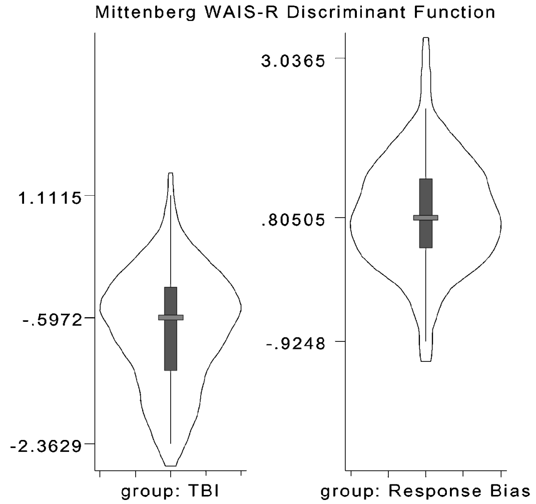

Fig. 1. *Violin Graphs of Mittenberg WAIS – R Discriminant Function* (Millis et al. 1998). The median is the short horizontal line, the interquartile range is the narrow shaded box, the *y*-axis is labeled at the minimum, median, and maximum scores for each group, and the data are surrounded by mirrored density traces.

found that a 3-variable CVLT discriminant function (Recognition Discriminability, Long Delay Cued Recall (LDCR), List A Trials 1–5 Recall) yielded an overall correct classification rate of 91% when applied to a group of patients with moderate to severe brain injuries and a group of probable clinical malingerers. Jackknife cross-validation also produced an overall hit rate of 91% with 91% (95% CI = 73.2–97.6) for sensitivity to response bias and 91% (95% CI = 73.2–97.6) for specificity to TBI. Millis et al. (1995a) also examined an individual CVLT variable, Recognition Hits, to attempt replication of an earlier study by Trueblood and Schmidt (1993). Sensitivity was 83% (95% CI = 63.9–93) and specificity was 96% (95% CI = 79–99.2). Subsequent studies (Baker, Donders, & Thompson, 2000; Coleman, Rapport, Millis, Ricker, & Farchione, 1998; Sweet et al., 2000) applied the original cut-off scores derived by Millis et al. (1995a) to other groups of persons with TBI and to groups of analog and probable clinical malingerers. Estimates of the diagnostic accuracy

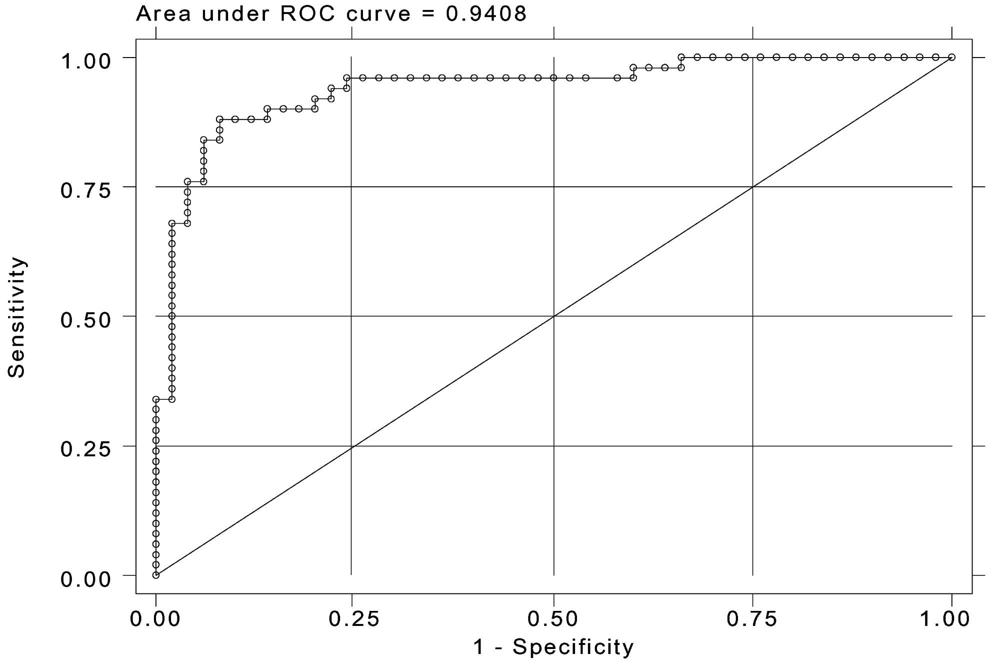

Fig. 2. *Receiver Operating Characteristic (ROC) Curve: Mittenberg WAIS – R Discriminant Function* (Millis et al., 1998).

of these cut-off scores have not been as high as those reported by Millis et al. (1995a), but they have tended to be within the original 95% confidence intervals. In general, sensitivity to response bias has been lower than the original estimates (63–80% for hits and 74% for the discriminant function), but specificity for TBI has remained high (e.g., 87–94% for hits and 83–93% for the function). Biased responding appears to be associated with disproportionate impairment on recognition tasks in an absolute sense and in relationship to free recall performance on list-learning tasks. A similar pattern has been found when using the Rey Auditory Verbal Learning Test (Bernard, 1991; Binder, Villaneuva, Howieson, & Moore, 1993; Suhr, Tranel, Wefel, & Barash, 1997).

Performance Curve Analysis

Frederick, Crosby, and Wynkoop (2000) have discussed the limitations of a dichotomous classification scheme in which the only categories are malingering versus non-malingering. They have proposed an alternative four-four scheme based on two dimensions: motivation and effort. Persons can vary from high to low on these dimensions and are characterized as compliant, careless, irrelevant, or malingered in their response style. Frederick (1997) developed the Validity Indicator Profile (VIP) based on this conceptualization. The VIP has verbal and non-verbal subtests, uses a dichotomous forced-choice format, and contains items with varying difficulty. Performance curves are generated from the patterns of items that are correctly answered, which then serve as the basis for classification. To date, the validation studies of the VIP have focused primarily on samples of criminal defendants referred for pretrial mental health evaluations (Frederick et al., 2000), and thus, the generalizability of the findings to mild TBI litigants remains to be established. Nonetheless, the VIP's performance curve approach may represent an important methodological advance that could be applied to other procedures. For example, Gudjonsson and Shakleton (1986) used a less complex form of performance curve analysis to examine the linear trend and item difficulty with Raven's Standard Progressive Matrices to detect response bias in an analog study. A second analog study by McKinzey, Podd, Krehbiel, and Raven (1999) found that the same linear trend formula correctly classified 95% of normal subjects taking the test under standard conditions and 74% of normal subjects instructed to fake.

Other Indices

High error rates on tactile finger recognition or localization tasks could signal response bias in the context of a litigated mild TBI case without peripheral injuries, particularly when errors are in excess of seven (Binder & Willis, 1991; Heaton et al., 1978; Mittenberg et al., 1996; Trueblood & Schmidt, 1993; Youngjohn, Burrows, & Erdal, 1995). Cognitive or physiological processes that may be difficult for persons to alter so as to appear genuinely impaired may have promise in detecting response bias, such as event-related potentials (Ellwanger, Tenhula, Rosenfeld, & Sweet, 1999; Rosenfeld, Sweet, Chuang, Ellwanger, & Song, 1996) and priming memory tasks (Davis et al., 1997; Hanley, Baker, & Ledson, 1999; McGuire & Shores, 1998).

Although a detailed discussion of the Minnesota Multiphasic Personality Inventory-2 (MMPI-2) is outside the scope of this paper, it is worth noting that the Fake Bad Scale (FBS), a rationally derived scale composed of 43 MMPI–2 items, may warrant particular attention in characterizing the manner in which neuropsychological malingerers respond. FBS elevations have been associated with litigating mild head-injured claimants who scored below chance on FCT (Larrabee, 1998; Millis, Putnam, & Adams, 1995b), with chronically symptomatic mild head injury patients whose WAIS–R FSIQ scores declined on retesting (Putnam, Kurtz, Millis, Adams, & O'Leary, 1995), and response time and the number of items correct on the VSVT (Slick, Hopp, Strauss, & Spellacy, 1996). Scores in excess of 22 on the FBS may suggest response bias in litigated mild TBI cases.

THE PROBLEM WITH TESTS

The core problem is that a test in isolation cannot 'prove' the diagnosis of malingering, brain dysfunction, or any disorder. Tests can only provide evidence in support of various diagnoses. Even then, the test result must be combined with prior information or knowledge before it can be interpreted meaningfully. Although high diagnostic sensitivity and specificity are desirable properties for a test to possess, these parameters alone cannot answer the fundamental diagnostic question, namely, given a positive test score, what is the probability that the patient has the disorder?

The only way to answer this question is to combine the test result with the pretest odds or prior probability, sometimes also referred to as a base rate or prevalence of the disorder. For the purposes of this calculation, a diagnostic test can be characterized in terms of a single number, known as the likelihood ratio (LR): sensitivity / (1 – specificity) (Sackett, Straus, Richardson, Rosenberg, & Haynes, 2000). The LR indicates how much more likely a positive test is to be found in a person with, as opposed to without, the disorder (Greenhalgh, 1997). The LR is then multiplied by the pretest odds to obtain the posttest odds (i.e., the probability that the person has the disorder given a positive test result).

For example, the prevalence of persisting neuropsychological impairment associated with mild TBI has been estimated to be 0.05, based on meta-analysis (Binder et al., 1997), yielding pretest odds of 0.05 / (1–0.05) = 0.053. The Average Impairment Rating (AIR) from the HRB has a sensitivity of 0.80 and a specificity of 0.88 (Heaton, Grant, & Matthews, 1991) when using a *T*-score cut-off of less than 40 to define impairment, yielding a LR of 6.7 (95% CI = 5.2–8.6). Thus, if a patient with a history of a mild TBI, in the chronic stage of recovery, obtained an impaired AIR, the posttest odds would be (0.053) × (6.7) = 0.36 (or 9:25) in favor of the diagnosis of brain injury. Converting odds to a probability, 0.36 / (1 + 0.36), there would be a 26% probability in support of a diagnosis of brain injury. On the basis of this one specific finding, there is insufficient evidence to support the diagnosis of TBI in this case.

In a second example, the prevalence of biased responding following mild TBI can be estimated to be 0.26, based on a weighted mean effect size from a meta-analysis of three studies by Binder and Rohling (1996) and the application of the *U*1 statistic (Cohen, 1988), yielding pretest odds of 0.26 / (1 – 0.26) = 0.35. Selecting the midpoints for sensitivity (82.5) and specificity (94.5) estimates from studies using the Hiscock and related digit recognition FCTs, a LR of 15 is obtained. Thus, if a hypothetical mild TBI litigant failed this FCT, the posttest odds would be (0.35) × (15) = 5.25, or 84% in favor of the diagnosis of biased responding. In this case, the evidence of biased responding appears to be compelling. Additional information would be helpful in determining whether a diagnosis of MND is warranted.

Similarly, likelihood ratios can be estimated for other response bias indices if one has reasonably good estimates of a measure's sensitivity and specificity (e.g., CVLT Hits = 9 and WAIS – R discriminant function = 11). An obvious issue in the use of the LR is the availability of reasonably accurate estimates of the prior probability or prevalence of the disorder. However, the accuracy of all diagnostic decisions depend on the estimates or assumptions that one makes about prevalence rates, even if one does not use the LR. The LR simply forces the diagnostician to be explicit about the assumptions. Increasingly, large sample studies of TBI (Dikmen et al., 1995) and meta-analyses (Binder et al., 1997) can assist in developing prevalence estimates. In addition, inferences about population prevalence can be estimated in the absence of a diagnostic gold standard with recently developed Bayesian methods using the Gibbs sampler, which is an iterative Markov-chain Monte Carlo technique (Joseph et al., 1995).

A separate issue involves the use of multiple diagnostic tests. If the tests were independent, one could simply multiply the running product by the likelihood ratio generated by each subsequent test. However, it is unlikely that most neuropsychological and response bias tests are unrelated. In fact, there may be a great deal of redundancy or multicollinearity among tests. A different approach is needed to combine several indices, to which we now turn.

BAYESIAN MODEL AVERAGING

From the foregoing review, it is apparent that neuropsychologists have at their disposal dozens of malingering tests and indices from which to choose. However, the optimal choice of tests may be far from obvious. Clinicians may opt to choose a subset from the battery of tests available to them. However, even with a moderate number of tests available, the space of available subsets is daunting; for example, if one considers 15 different tests, there are 32,768 potential subsets of tests to evaluate (i.e., 2^{15}). Of course, the clinician will want to combine the tests in a way that takes into consideration each test's diagnostically meaningful contribution. In addition, the redundancy of the tests must be considered. The finding that an examinee scores 'negative' on a dozen malingering tests may not be particularly enlightening if there is a high degree of collinearity among them.

Until recently, investigators have had limited methods to evaluate and select sets of tests. Choosing sets on the basis of theory offers little guidance because theoretical considerations often generate a large number of models and candidate variables. Stepwise regression methods (Efroymson, 1960) provide a standard methodology for finding optimal sets of tests and their weights for use in prediction. However, these methods have well documented flaws: they overemphasize the confidence in the model, they tend to include noise variables in the final model, and they are sensitive to small changes in the data (Freedman, 1983; Hocking, 1976). In addition, each potential test is dichotomized as 'significant' or 'not significant,' which oversimplifies the amount of diagnostically valuable information in each test.

By selecting a single collection of tests (i.e., a single model), stepwise regression and similar model selection techniques do not account for the substantial uncertainty in the model selection process. Recent research argues that averaging over many models can provide significantly better predictions by taking into account the model uncertainty. Bayesian model averaging (BMA, Hoeting, Madigan, Raftery, & Volinsky, 1999) provides an approach to hypothesis testing, model selection, and accounting for model uncertainty that overcomes the difficulties associated with conventional frequentist *p*-value significance tests and model selection procedures. BMA approaches the problem of model selection by finding a collection of the best models, and averaging over them in accordance with their posterior model probabilities. The different models and variables (in this case, the different tests for response bias) are incorporated into the predictions with weights proportional to the evidence we have for their utility. Madigan and Raftery (1994) have shown that averaging over many models in this manner provides superior out-of-sample predictive performance compared to the typical approach of evaluating a single model. BMA has been used successfully in several problem domains, including analysis of a Mayo Clinic study using a Cox proportional hazard model to investigate risk factors for stroke in the elderly (Volinsky, Madigan, Raftery, & Kronmal, 1997). The Appendix presents a brief discussion of the mathematical foundations of BMA. In the following section, we demonstrate an application of BMA in the development of response bias indices with the CVLT.

We used BMA to select and evaluate models composed of variables from the CVLT that sought to differentiate persons with documented moderate to severe TBI from persons with mild injuries in litigation who were not giving their best effort on neuropsychological testing (i.e., were responding in a biased manner). As previously discussed, several investigators have used single and multiple variables from the CVLT to detect response bias. Although there has been a reasonable degree of convergence in the findings from diverse studies, an unresolved question is whether investigators have yet identified optimal sets of the 17 possible CVLT variables to predict response bias. If we consider a specific subset of these 17 variables to be a model, these data generate $2^{17} = 131{,}072$ models to consider! The BMA techniques allowed us to identify the best models and average over them.

Participants

All participants were outpatients at a Midwest United States university-affiliated rehabilitation hospital. The litigation group (LG) was composed of 80 participants (age: $M = 39.1$, $SD = 11.3$; education: $M = 11.6$, $SD = 2.0$) with alleged TBI who were in litigation and claimed to be vocationally disabled. They had brief or no loss of consciousness (i.e., less than 5 min), normal CT or MRI brain scans, and no focal neurological deficits. Although none admitted to malingering, all subjects obtained scores within or below chance on one or both subtests of a dichotomous, forced-choice measure, the RMT (Warrington, 1984). These characteristics suggested that these persons were not giving their best effort to complete the neuropsychological tests, and thus represented a reasonable clinical approximation of malingering or response bias. Mean time postinjury was 21.0 months ($SD = 22.2$).

The TBI group was composed of 80 participants (age: $M = 37.1$, $SD = 9.2$; education: $M = 11.9$, $SD = 2.2$) who had sustained a moderate to severe TBI, as assessed by the GCS with scores ranging from 3 to 12. Mean time postinjury was 27.1 months ($SD = 40.2$).

Procedure

A logistic regression model was selected for fitting and evaluating because the objective of this investigation was the prediction of a binary outcome (i.e., TBI vs. Response Bias), on the basis of 17 CVLT variables (CVLT Total Recall, Trial 5 Recall, List B Recall, Short Delay Free Recall (SDFR), Short Delay Cued Recall (SDCR), Long Delay Free Recall (LDFR), Long Delay Cued Recall (LDCR), Semantic Clustering (SC), Primacy, Recency, Slope, Consistency, Perseverations, Intrusions, Hits, False Positives, and Bias). We chose two programs written in S-Plus (MathSoft, 1999) to perform BMA: BIC.GLM (Volinsky et al, 1997) was used to pare down the model space to a manageable set of candidate models, whereas GLIB (Raftery, 1996) calculated the marginal likelihoods, posterior model probabilities, and the averaged coefficients for each variable.

RESULTS

Using BIC.GLM to traverse the model space in search of the best models, we identified 13 models with support from the data over a predefined threshold. Table 2 presents these models in the order of decreasing estimated posterior probability. Only 7 out of the 17 possible CVLT variables appeared, indicating that the remaining 10 contained negligible added value over and above these variables. Also, SDFR and SDCR never appeared together in a model, which suggested marked collinearity. The most important variable was Hits, which appeared in all 13 models. The fact that no other variable appeared in all models indicated a high degree of model uncertainty.

GLIB was then used to refine the inferences by exact calculation of the 13 models' posterior probabilities. To perform the calculation, we provided a reference set of priors indexed by a

Table 2. Selection of Independent Variables.$^{+}$

Model	SDFR	SDCR	LDFR	Primacy	Hits	False Positives	Bias
1	X		X	X	X		X
2	X		X		X		X
3	X		X	X	X		
4		X	X		X		X
5		X			X		X
6	X		X		X		
7	X				X		X
8					X		X
9		X	X		X	X	
10				X	X		
11					X	X	
12		X	X		X		
13					X		

Note. $^{+}$SDFR = Short Delay Free Recall, SDCR = Short Delay Cued Recall, LDFR = Long Delay Free Recall.

Table 3. Posterior Model Probabilities for the Selected Models.$^{+}$

		Posterior probabilities (%)		
Model	Variables	$\phi = 1.00$	$\phi = 1.65$	$\phi = 5.00$
2	SDFR, LDFR, Hits, Bias	42	43	44
3	SDFR, LDFR, Primacy, Hits	25	24	24
4	SDCR, LDFR, Hits, Bias	25	26	26
9	SDCR, LDFR, Hits, False Positives	7	7	6

Note. $^{+}$SDFR = Short Delay Free Recall, LDFR = Long Delay Free Recall.

parameter ϕ (Raftery 1996). We then compared nested models and excluded models with more parameters if a simpler model had a higher posterior probability. Models 2–4, and 9 were retained for further consideration. Table 3 shows the normalized posterior probabilities of these four models under three different values of ϕ. Selection of ϕ may bias the results towards simple or complex models, but as Table 3 shows, ϕ did not make much of a difference in the posterior model probabilities. Under a wide range of values of ϕ, Model 2 had the most support from the data while the support for the other models showed moderate model uncertainty. To balance the complexity considerations, we focused on $\phi = 1.65$ in the subsequent analyses.

Using Table 4, we can derive the posterior effect probabilities for the individual CVLT variables. If we assume that a priori each variable has a 50% probability of having a non-zero parameter, Table 4 contains the posterior probabilities that the parameter is different from zero. For instance, the posterior distribution of the parameter for SDFR would have a point mass of 0.32 at the value zero, with the other 68% of the distribution centered around its averaged parameter estimate. Again, the probabilities were not significantly influenced by the selection of ϕ.

Table 4. Posterior Probabilities for Inclusion of Each Variable.$^{+}$

	Posterior probabilities (%)		
Variable	$\phi = 1.00$	$\phi = 1.65$	$\phi = 5.00$
SDFR	68	68	68
SDCR	32	32	32
LDFR	100	100	100
Primacy	25	24	24
Recognition Hits	100	100	100
False Positives	7	7	6
Response Bias	67	69	70

Note. $^{+}$ SDFR = Short Delay Free Recall, SDCR = Short Delay Cued Recall, LDFR = Long Delay Free Recall.

In summary, there was very strong evidence for the model that included Hits and LDFR in a CVLT model to predict malingering. In addition, moderately strong evidence was found for SDFR and Response Bias. The data appeared inconclusive regarding the roles of Primacy and SDCR and support was essentially nil for False Positives. For the remaining 10 CVLT variables, we found little evidence to include them in our models.

Table 5 presents the group means and standard deviations for the four variables that had the highest posterior support for predicting malingering. These four variables are contained in Model 2, which had the highest posterior probability of any single model. Overall, the litigation group's levels of performance on the free recall and

Table 5. Group Means and Standard Deviations for CVLT Variables Having Highest Posterior Probabilities.$^{+}$

	TBI		Litigating	
CVLT variables	*M*	(*SD*)	*M*	(*SD*)
SDFR	7.3	(3.2)	5.1	(2.9)
LDFR	8.0	(3.5)	4.5	(2.9)
Hits	14.0	(1.8)	9.2	(3.7)
Bias	0.01	(0.4)	−0.2	(0.4)

Note. $^{+}$SDFR = Short Delay Free Recall, LDFR = Long Delay Free Recall.

recognition measures (SDFR, LDFR, and Hits) were poorer than those of the TBI group. In addition, the litigation group showed disproportionate 'impairment' on the recognition task compared to their performance on free recall tasks. In contrast, the TBI group demonstrated comparable levels of performance across recall and recognition tasks relative to the CVLT norms, and a neutral response tendency on the recognition trial.

Predictive Performance

Ultimately the goal of any of these models is to predict whether a given individual is exhibiting response bias. Any single model can be used to calculate a probability that a given individual is in the TBI group. Similarly, the BMA analysis provides a model-averaged probability. One way to check the effectiveness of this new methodology is to see how well it is able to discriminate the TBI group from the litigating group in a hold-out sample. We compared the BMA analysis to a standard method that is commonly used, stepwise logistic regression. To accomplish this, we used 75% of the observations as a training sample to build all the models, and 25% as a test sample to assess the predictive performance. Our predictive measure was the rate of overall correct classification averaged over 100 iterations of this procedure. BMA correctly classified 78.2% of the test sample compared to 77.7% for the stepwise procedure.

DISCUSSION

The models selected with BMA in this investigation were consistent with findings from previous investigations. That is, one pattern of malingering is characterized by disproportionately poor performance on recognition measures in an absolute sense and relative to performances on free recall measures. In contrast, persons with TBI generally find recognition tasks easier than free recall or perform different memory tasks at comparable levels. Previous investigations (Millis et al., 1995a; Sweet et al., 2000; Trueblood & Schmidt, 1993) found Recognition Hits on the CVLT to be one of the most useful variables in detecting malingering. BMA corroborates this, as Hits was the only variable to show up in all models. However, by selecting a model that only includes Hits, the analyst is neglecting the handful of other variables that appeared to contribute incrementally significant information in the detection of malingering.

BMA also quantified the contribution of the other variables and compared them with the contribution of Hits via posterior effect probabilities. With the exception of Hits and LDFR and possibly Response Bias and SDFR, we unfortunately found little evidence to consider the remaining 13 CVLT variables. This was disappointing as we had hoped to detect and describe a qualitatively more complex response bias construct than is implied by a 'recognition versus free recall' pattern. On the other hand, response bias may be only partially captured by the CVLT. Alternatively, the construct of response bias may not be as psychometrically complex as previously thought.

To our knowledge, this is the first application of BMA in clinical neuropsychology and one that demonstrates the benefits of BMA. After narrowing down the model space to four models to average, the best model still only had 42% of the posterior probability. This indicated that there was too much model uncertainty to claim that any one single model could sufficiently fit these data. Conventional frequentist approaches tend to overstate the evidence for the effects of variables and have no method for quantifying model uncertainty. Moreover, standard approaches tend to reinforce the illusory notion that there is only one 'true' model that accurately describes phenomena. Accounting for this model uncertainty results in a better understanding of the variables, a better quantification of the models' effectiveness, and improved predictive performance. Admittedly, the difference in the predictive performance between BMA and stepwise selection was negligible in the current analysis. However, this finding was likely to be related to the study's small sample and the idiosyncrasies of this particular set of variables. That is, Recognition Hits and LDFR were powerful predictors, but the remaining variables were relatively weak. Almost any selection method would have minimal difficulty in choosing Hits and LDFR. We suspect that BMA's superiority would be more apparent with a larger sample and with a set of variables having a wider range of predictive capacity.

GENERAL SUMMARY

As humans, we often have a low tolerance for ambiguity, which impels us to impose meaning on experience. This tendency carries over into the diagnostic realm. If we rely on our clinical judgment alone, our diagnostic accuracy can be abysmal. Humans tend to ignore prevalence rates, assign non-optimal weights to predictor variables, disregard regression toward the mean, improperly assess covariation, and over-weigh vivid data (Grove, Zald, Lebow, Snitz, & Nelson, 2000). Meehl (1954) was among the first to alert psychologists to the superiority of statistical prediction compared to clinical judgment. Little has changed in this regard over the last 46 years. In a recent meta-analysis of 136 studies on health and human behavior, Grove et al. (2000) found that statistical prediction techniques were about 10% more accurate than clinical predictions. Statistical prediction significantly outperformed clinical prediction in 33–47% of the studies. Clinical prediction was more accurate in only 6–16% of the cases. The superiority of statistical prediction was consistent regardless of type of judgment or judge, judges' amount of experience, or type of data.

We are not advocating that statistical algorithms replace clinical judgment. However, we do believe that statistical algorithms deserve wider use in the detection of response bias. Detecting response bias will be likely to require more than a single test. We strongly recommend the use of at least one FCT to assess response bias. Below chance performance on a FCT is certainly compelling evidence of response bias, but only a minority of malingering cases will perform this poorly. Moreover, the simplicity of some FCTs may render their intent transparent so that malingerers may easily evade detection. In addition, the appearance and format of some FCTs are easily recognizable. Thus, an unremarkable FCT performance does not rule out response bias.

There is a clear need to supplement FCTs with other response bias indices such as those employing floor effects, multivariable composites, and/or performance curve analysis. Such indices may be less prone to coaching and can be easily integrated into the clinical examination. However, much work remains to be done to determine optimal sets of these indices and measures. Techniques like BMA appear to have considerable potential to assist us in finding parsimonious detection models without overly simplifying complex processes.

REFERENCES

Alexander, M.P. (1995). Mild traumatic brain injury: Pathophysiology, natural history, and clinical management. *Neurology*, *45*, 1253–1260.

Axelrod, B.N., & Rawlings, D.B. (1999). Clinical utility of incomplete effort WAIS – R Formulas: A longitudinal examination of individuals with traumatic brain injuries. *Journal of Forensic Neuropsychology*, *1*, 15–27.

Baker, R., Donders, J., & Thompson, E. (2000). Assessment of incomplete effort with the California Verbal Learning Test. *Applied Neuropsychology*, *7*, 111–114.

Bernard, L.C. (1991). The detection of faked deficits on the Rey Auditory Verbal Learning Test: The effect of serial position. *Archives of Clinical Neuropsychology*, *6*, 81–88.

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

Binder, L. M, & Rohling, M.L. (1996). Money matters: A meta-analytic review of the effect of financial incentives on recovery after closed head injury. *American Journal of Psychiatry*, *153*, 7–10.

Binder, L.M., Rohling, M.L., & Larrabee, G.J. (1997). A review of mild head trauma. Part I: Meta-analysis review of neuropsychological studies. *Journal of Clinical and Experimental Neuropsychology*, *19*, 421–431.

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

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

Chow, S.L. (1996). *Statistical significance: Rationale, validity and utility.* Thousand Oaks, CA: Sage Publications.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Coleman, R., Rapport, L., Millis, S., Ricker, J., & Farchione, T. (1998). Effects of coaching on the detection of malingering on the California Verbal Learning Test: An analog study of malingered head injury. *The Clinical Neuropsychologist*, *20*, 201–210.

Conder, R., Allen, L., & Cox, D. (1992). *Computerized Assessment of Response Bias Test Manual*. Durham, N.C.: Cognisyst.

Davis, H.P., King, J.H., Klebe, K.J., Bajszar, G., Bloodworth, M.R., & Wallick, S.L. (1997). The detection of simulated malingering using a computerized priming task. *Archives of Clinical Neuropsychology*, *12*, 145–153.

Delis, D.C., Kramer, J.H., Kaplan, E., & Ober, B.A. (1987). *California Verbal Learning Test: Adult version*. San Antonio, TX: The Psychological Corporation.

Dicker, B.G. (1992). Profile of those at risk for minor head injury. *Journal of Head Trauma Rehabilitation*, *7*, 83–91.

Dikmen, S.S., Machamer, J.E., Winn, H.R., & Temkin, N.R. (1995). Neuropsychological outcome at one-year post head injury. *Neuropsychology*, *9*, 80–90.

Dikmen, S.S., McLean, A., & Temkin, N. (1986). Neuropsychological and psychosocial consequences of minor head injury. *Journal of Neurology, Neurosurgery, and Psychiatry*, *49*, 1227–1232.

Dikmen, S.S., & Levin, H.S. (1993). Methodological issues in the study of mild head injury. *Journal of Head Trauma Rehabilitation*, *8*, 30–37.

Doss, R.C., Chelune, G.J., & Naugle, R.I. (1999). Victoria Symptom Validity Test: Compensation-seeking vs. non-compensation-seeking patients in a general clinical setting. *Journal of Forensic Neuropsychology*, *1*, 5–20.

Efroymson, M.A. (1960). Multiple regression analysis. In A. Ralston & H.S. Wilf (Eds.), *Mathematical methods for digital computers*. New York: Wiley.

Ellanger, J., Tenhula, W.N., Rosenfeld, J.P., & Sweet, J.J. (1999). Identifying simulators of cognitive deficit through the combined use of neuropsychological test performance and event-related potentials. *Journal of Clinical and Experimental Neuropsychology*, *21*, 866–879.

Etcoff, L.M., & Kampfer, K.M. (1996). Practical guidelines in the use of symptom validity and other psychological tests to measure malingering and symptom exaggeration in traumatic brain injury cases. *Neuropsychology Review*, *6*, 171–201.

Fenton, G., McClelland, R., Montgomery, A., MacFlynn, G., & Rutherford, W. (1993). The postconcussional syndrome: Social antecedents and psychological sequelae. *British Journal of Psychiatry*, *162*, 493–497.

Frederick, R.I. *Validity Indicator Profile manual*. Minnetonka, MN: NCS Assessments.

Frederick, R.I., Crosby, R.D., & Wynkoop, T.F. (2000). Performance curve classification of invalid responding on the Validity Indicator Profile. *Archives of Clinical Neuropsychology*, *15*, 281–300.

Freedman, D.A. (1983). A note on screening regression equations. *American Statistician*, *37*, 152–155.

Gentilini, M., Nichelli, P., Schoenhuber, R., Bortolotti, P., Tonelli, L., Falasca, A., & Merli, G. (1985). Neuropsychological evaluation of mild head injury. *Journal of Neurology, Neurosurgery and Psychiatry*, *48*, 137–140.

Gfeller, J.D., & Cradock, M.M. (1998). Detecting feigned neuropsychological impairment with the Seashore Rhythm Test. *Journal of Clinical Psychology*, *54*, 431–438.

Goebel R.A. (1983). Detection of faking on the Halstead – Reitan Neuropsychological Test Battery. *Journal of Clinical Psychology*, *39*, 731–742.

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

Green, P., & Iverson, G. (in press). Validation of the Computerized Assessment of Response Bias in ligitating patients with head injuries. *The Clinical Neuropsychologist*.

Green, P., Iverson, G.L., & Allen, L. (1999). Detecting malingering in head injury litigation with the Word Memory Test. *Brain Injury*, *13*, 813–819.

Greenhalgh, T. (1997). How to read a paper: Papers that report diagnostic or screening tests. *British Medical Journal*, *315*, 540–543.

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

Greiffenstein, M.F., Gola, T., & Baker, W.J. (1995). MMPI-2 validity scales versus domain specific measures in the detection of factitious traumatic brain injury. *The Clinical Neuropsychologist*, *9*, 218–224.

Grove, W.M., Zald, D.H., Lebow, B.S., Snitz, B.E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, *12*, 19–30.

Gudjonsson, G.H., & Shackleton, H. (1986). The pattern of scores on Raven's Matrices during 'faking bad' and 'non-faking' performance. *British Journal of Clinical Psychology*, *25*, 35–41.

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*, 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*, 283–294.

Hanley, J.R., Baker, G.A., & Ledson, S. (1999). Detecting the faking of amnesia: A comparison of the effectiveness of three different techniques for distinguishing simulators from patients with amnesia. *Journal of Clinical and Experimental Neuropsychology*, *21*, 59–69.

Heaton, R.K., Grant, I., & Matthews, C.G. (1991). *Comprehensive norms for an expanded Halstead–Reitan Battery*. Odessa, FL: Psychological Assessment Resources, Inc.

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

Hintze, J.L., & Nelson, R.D. (1998) Violin plots: A box plot-density trace synergism. *The American Statistician*, *52*, 181–184.

Hiscock, C.K., Branham, J.D., & Hiscock, M. (1994). Detection of feigned cognitive impairment: The two-alternative forced-choice method compared with selected conventional tests. *Journal of Psychopathologic Behavior*, *16*, 95–110.

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

Hocking, R.R. (1976). The analysis and selection of variables in linear regression. *Biometrics*, *32*, 1–49.

Hoeting, J.A., Madigan, D., Raftery, A.E., & Volinsky, C.T. (1999). Bayesian model averaging: A tutorial (with discussion). *Statistical Science*, *14*, 382–417. (a corrected version of the paper is available at http://www.stat.colostate.edu/~jah/documents/bma2.ps).

Iverson, G.L., & Binder, L.M. (2000). Detecting exaggeration and malingering in neuropsychological assessment. *Journal of Head Trauma Rehabilitation*, *15*, 829–858.

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*, 323–334.

Iverson, G., Green, P., & Gervais, R. (1999, March/April). Using the Word Memory Test to detect biased responding in head injury litigation. *The Journal of Cognitive Rehabilitation*, 4–8.

Joseph, L., Gyorkos, T.W., & Coupal, L. (1995). Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *American Journal of Epidemiology*, *141*, 263–272.

Kreutzer, J.S., Gordon, W.A., Rosenthal, M., & Marwitz, J. (1993). Neuropsychological characteristics of patients with brain injury: Preliminary findings from a multicenter investigation. *Journal of Head Trauma Rehabilitation*, *8*, 47–59.

Klonoff, P.S., & Lamb, D.G. (1998). Mild head injury, significant impairment on neuropsychological test scores, and psychiatric disability. *The Clinical Neuropsychologist*, *12*, 31–42.

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: An analog investigation. *Psychological Assessment*, *6*, 8–13.

Larrabee, G.J. (1990). Cautions in the use of neuropsychological evaluation in legal settings. *Neuropsychology*, *4*, 239–247.

Larrabee, G.J. (1998). Somatic malingering on the MMPI and MMPI-2 in personal injury litigants. *The Clinical Neuropsychologist*, *12*, 179–188.

Lees-Haley, P.R., English, L.T., & Glenn, W.J. (1991). A Fake Bad scale on the MMPI-2 for personal injury claimant. *Psychological Reports*, *68*, 203–210.

Levin, H.S., Mattis, S., Ruff, R.M., Eisenberg, H.M., Marshall, L.F., & Tabaddor, K. (1987). Neurobehavioral outcome following minor head injury: A 3-center study. *Journal of Neurosurgery*, *66*, 234–243.

Madigan, D., & Raftery, A.E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, *89*, 1535–1546.

MathSoft. (1999). *S-Plus 2000*. Seattle, WA: MathSoft.

McGuire, B.E., & Shores, E.A. (1998). Malingering of memory impairment on the Colorado Priming Test. *British Journal of Clinical Psychology*, *37*, 99–102.

McKinzey, R.K., Podd, M.H., Krehbiel, M.A., & Raven, J. (1999). Detection of malingering on Raven's Standard Progressive Matrices: A cross-validation. *British Journal of Clinical Psychology*, *38*, 435–439.

McKinzey, R.K., & Russell, E.W. (1997). Detection of malingering on the Halstead-Reitan Battery: A cross-validation. *Archives of Clinical Neuropsychology*, *12*, 585–589.

Meehl, P.E. (1954). *Clinical vs. statistical prediction: A theoretical analysis and review of the evidence*. Minneapolis, MN: University of Minnesota Press.

Meyers, J.E., & Volbrecht, M. (1998). Validation of reliable digits for detection of malingering. *Assessment*, *5*, 303–307.

Millis, S.R. (1992). The Recognition Memory Test in the detection of malingered and exaggerated memory deficits. *The Clinical Neuropsychologist*, *6*, 406–414.

Millis, S.R., & Putnam, S.H. (1994). The Recognition Memory Test in the assessment of memory impairment after financially compensable mild head injury: A replication. *Perceptual and Motor Skills*, *79*, 384–386.

Millis, S.R., & Putnam, S.H. (1996). Detection of malingering in postconcussive syndrome. In M.

Rizzo & D. Tranel (Eds.), *Head injury and postconcussive syndrome* (pp. 481–498). New York: Churchill Livingstone.

Millis, S.R., Putnam, S.H., & Adams, K.M. (1995b, March). *Neuropsychological malingering and the MMPI-2: Old and new indicators*. Paper presented at the 30th Annual Symposium on Recent Developments in the Use of the MMPI, MMPI-2, and MMPI-A, St. Petersburg Beach, FL.

Millis, S.R., Putnam, S.H., & Adams, K.M. (1996). Speech-sounds Perception Test and Seashore Rhythm Test as validity indicators in the neuropsychological evaluation of mild head injury [abstract]. *Archives of Clinical Neuropsychology*, *11*, 425.

Millis S.R., Putnam S.H., Adams, K.M., & Ricker, J.H. (1995a). The California Verbal Learning Test in the detection of incomplete effort in neuropsychological evaluation. *Psychological Assessment*, *7*, 463–471.

Millis, S.R., Ross, S.R., & Ricker, J.H. (1998). Detection of incomplete effort on the Wechsler Adult Intelligence Scale – Revised: A cross-validation. *Journal of Clinical and Experimental Neuropsychology*, *20*, 167–173.

Mittenberg, W., Azrin, R., Millsaps, C., & Heilbronner, R. (1993). Identification of malingered head injury on the Wechsler Memory Scale – Revised. *Psychological Assessment*, *5*, 34–40.

Mittenberg, W., DiGiulio, D.V., Perrin, S., & Bass, A.E. (1992). Symptoms following mild head injury: Expectation as etiology. *Journal of Neurology, Neurosurgery and Psychiatry*, *55*, 200–204.

Mittenberg W., Rotholc, A., Russell, E., & Heilbronner, R. (1996). Identification of malingered head injury on the Halstead – Reitan Battery. *Archives of Clinical Neuropsychology*, *11*, 271–281.

Mittenberg, W., Theroux-Fichera, S., Zielinski, R.E., & Heilbronner, R. (1995). Identification of malingered head injury on the Wechsler Adult Intelligence Scale – Revised. *Professional Psychology: Research and Practice*, *26*, 491–498.

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

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

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

Ponsford, J., Willmot, C., Rothwell, A., Cameron, P., Kelly, A, Nelms, R., Curran, C., & Ng, K. (2000). Factors influencing outcome following mild traumatic brain injury in adults. *Journal of the International Neuropsychological Society*, *6*, 568–570.

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

Putnam, S.H., Kurtz, J.E., Millis, S.R., Adams, K.M., & O'Leary, J.F. (1995). MMPI-2 correlates of unexpected cognitive deterioration in traumatic brain injury [Abstract]. *The Clinical Neuropsychologist*, *9*, 295.

Raftery, A.E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalized linear models. *Biometrika*, *83*, 251–266.

Rees, L.M., Tombaugh, T.N., Gansler, D.A., & Moczynski, N.P. (1998). Five validation experiments of memory malingering (TOMM). *Psychological Assessment*, *10*, 10–20.

Rogers, R., Harrell, E.H., & Liff, C.D. (1993). Feigning neuropsychological impairment: A critical review of methodological and clinical considerations. *Clinical Psychology Review*, *13*, 255–274.

Rosenfeld, J.P., Sweet, J.J., Chuang, J., Ellwanger, J., & Song, L. (1996). Detection of simulated malingering using forced choice recognition enhanced with event-related potential recording. *The Clinical Neuropsychologist*, *10*, 163–179.

Ruff, R., Camenzuli, L., & Mueller, J. (1996). Miserable minority: Emotional risk factors that influence the outcome of a mild traumatic brain injury. *Brain Injury*, *10*, 551–566.

Sackett, D.L., Straus, S.E., Richardson, W.S., Rosenberg, W., & Haynes, R.B. (2000). *Evidence-based medicine*. New York: Churchill Livingstone.

Slick, D.J., Hopp, G., Strauss, E., & Spellacy, F.J. (1996). Victoria Symptom Validity Test: Efficiency for detecting feigned memory impairment and relationship to neuropsychological tests and MMPI-2 validity scales. *Journal of Clinical and Experimental Neuropsychology*, *18*, 911–922.

Slick, D.J., Hopp, G., Strauss, E., & Thompson, G.B. (1997). *Victoria Symptom Validity Test: Professional manual*. Odessa, FL: Psychological Assessment Resources.

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

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

Sweet, J.J., Wolf, P., Sattlberger, E., Numan, B., Rosenfeld, J.P., Clingerman, S., & Nies, K.J. (2000). Further investigation of traumatic brain injury versus insufficient effort with the California Verbal Learning Test. *Archives of Clinical Neuropsychology*, *15*, 105–113.

Tombaugh, T.N. (1997). The Test of Memory Malingering (TOMM): Normative data from cognitively intact and cognitively impaired individuals. *Psychological Assessment*, *9*, 260–268.

Trueblood, W. (1994). Qualitative and quantitative characteristics of malingered and other invalid WAIS-R and clinical memory data. *Journal of Clinical and Experimental Neuropsychology*, *16*, 597–607.

Trueblood, W., & Schmidt, M. (1993). Malingering and other validity considerations in the neuropsychological evaluation of mild head injury. *Journal of Clinical and Experimental Neuropsychology*, *15*, 578–590.

Volinsky, C., Madigan, D., Raftery, A.E., & Kronmal, R.A. (1997). Bayesian model averaging in proportional hazard models: Predicting the risk of a stroke. *Applied Statistics*, *46*, 443–448.

Warrington, E.K. (1984). *Recognition memory test manual*. Berkshire, England: NFER-Nelson.

Wrightson, P., & Gronwall, D. (1981). Time off work and symptoms after minor head injury. *Injury*, *12*, 445–454.

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

APPENDIX

Following the notation of Hoeting et al. (1999), let Δ be the quantity of interest, such as the probability of response bias. We are interested in the posterior distribution of Δ given the data, D. We average over all possible models, M_k; $k = 1 \ldots$ K, where M_k indicates a unique subset of Response Bias tests:

$$pr\,(\Delta \mid D) = \sum_{k=1}^{K} pr(\Delta \mid M_k, D) pr(M_k \mid D). \tag{1}$$

This is a weighted average of the posterior distributions under each of the models considered, where the weights are the posterior model probabilities, $pr\,(M_k \mid D)$. These posterior probabilities are given by:

$$pr(Mk \mid D) = \frac{pr(D \mid M_k)\,pr\,(M_k)}{\sum_{i=1}^{I} pr(D \mid M_i)\,pr\,(M_i)'} \tag{2}$$

where $pr\,(M_i)$ is the prior for model M_i, and

$$pr\,(D|M_k) = \int pr(D|\theta_k, M_k)\,pr\,(\theta_k|, M_k)\,d\theta_k \tag{3}$$

is the marginal likelihood under model M_k (with parameters θ_k). Although Equation 1 suggests that we will be averaging over all of the models in our model space, Madigan and Raftery (1994) showed that Equation 1 is well approximated by averaging over a small group of parsimonious data-supported models, substantially reducing computational complexity. For further discussion of exhaustive summation and the computation of the integrals implied by the foregoing equations, see Hoeting et al. (1999).