Law and Human Behavior, Vol. 24, No. 1, 2000

Development and Validation of the Validity Indicator Profile

Richard I. Frederick[1,3] **and Ross D. Crosby**[2,4]

The Validity Indicator Profile (VIP; Frederick, 1997) is a two-alternative forced choice (2AFC) procedure intended to identify when the results of cognitive and neuropsychological testing may be invalid because of malingering or other problematic response styles. The test consists of 100 problems that assess nonverbal abstraction capacity and 78 word-definition problems. The VIP attempts to establish whether an individual's performance in an assessment battery should be considered representative of his or her true overall capacities (valid or invalid). Performances classified as valid are classified as "compliant" and reflect a high effort to respond correctly. Performances classified as invalid are subclassified as "careless" (low effort to respond correctly), "irrelevant" (low effort to respond incorrectly), or "malingering" (high effort to respond incorrectly). The VIP development sample included 944 nonclinical participants and 104 adults undergoing neuropsychological evaluation. The cross-validation sample consisted of 152 nonclinical participants, 61 brain-injured adults, 49 individuals considered to be at risk for malingering, and 100 randomly generated VIP protocols. The nonverbal subtest of the VIP demonstrated an overall classification rate of 79.8%, with 73.5% sensitivity and 85.7% specificity. The verbal subtest of the VIP demonstrated an overall classification rate of 75.5%, with 67.3% sensitivity and 83.1% specificity.

OVERVIEW OF THE VALIDITY INDICATOR PROFILE

The valid assessment of cognitive ability with psychological tests depends to a great extent on the cooperation of the test taker. Inferences made about the cognitive ability of an individual based on psychological testing assume that the demonstrated performances is a valid representation of the individual's maximal capacity to respond correctly. A valid assessment is possible only when an examinee

[1]Springfield, Missouri.
[2]Neuropsychiatric Research Institute, Fargo, North Dakota.
[3]Correspondence concerning this manuscript should be sent to Richard Frederick, Ph.D., Department of Psychology, U.S. Medical Center for Federal Prisoners, Springfield, Missouri 65807, or by electronic mail to rfrederi@ipa.net.
[4]Ross Crosby was at NCS Assessments, Minnetonka, Minnesota, during much of this project.

0147-7307/00/0200-0059$18.00/1

is properly motivated and exerts sufficient effort throughout the assessment. *Motivation* refers to the intended goal of the test taker (to perform well or poorly, to respond correctly or incorrectly). *Effort* is defined as the intensity of application to produce the intended goal of the test taker. Invalid assessments result from motivation to perform poorly or from poor effort when motivated to perform well.

NEED FOR BETTER ASSESSMENT OF TEST VALIDITY

Unlike many personality tests, standardized psychological and neuropsychological tests that measure cognitive ability do not routinely include an assessment of this sort of validity. Consequently, clinicians have traditionally been limited to subjective judgments about a test taker's level of cooperation. The ability of trained evaluators to detect cognitive malingering on the basis of subjective judgments has been consistently reported to be poor across a number of studies (Faust & Ackley, 1998; Faust, Hart, & Guilmette, 1988; Faust, Hart, Guilmette, & Arkes, 1988; Heaton, Smith, Lehman, & Vogt, 1978; Trueblood & Binder, 1997). This is no small problem. Perhaps as many as one-half of workers' compensation cases (Youngjohn, 1991), two-thirds of personal injury evaluations (Heaton et al., 1978), and one-fifth of Social Security disability claims (Griffin, Normington, May, & Glassmire, 1996) may be compromised by malingering. Greiffenstein, Baker, & Gola (1994) found that 41% of 106 consecutive referrals for neuropsychological evaluation of mild traumatic brain injury met two of four criteria for malingering. Rogers, Sewell, & Goldstein (1994) polled 320 forensic mental health specialists (96% of whom were psychologists) who had an average of 14 years experience and who had completed an average of more than 300 forensic evaluations. Their mean estimation of malingering in forensic examinations was 15.7%. A meta-analysis by Rohling, Binder, & Langhinrichsen-Rohling (1995) revealed that the potential for compensation resulted in increased reports of pain and decreased treatment effectiveness (an effect size of 0.60).

The relatively high prevalence of malingering in diverse forensic clinical examinations, coupled with the obvious limitations of subjective clinical judgment, point to the need for objective validity indicators in the assessment of cognitive functioning (Bigler, 1990; Faust & Guilmette, 1990). The recent spate of published articles concerning the detection of malingered cognitive impairments has much of its roots in the development of symptom validity testing (SVT) to evaluate suspicious complaints of sensory impairment (Pankratz, 1979; Pankratz, Fausti, & Peed, 1975). SVT involves the presentation of a large number of trials in which patients are asked to choose between two alternatives (e.g., sound or no sound). The inherent level of difficulty required to respond correctly across trials remains constant. Binomial probabilities are then derived to determine whether the number of incorrect responses exceeds what would be expected by chance alone. Performance that is unexpectedly lower than chance is considered evidence of malingering. This methodology has been adapted successfully to the assessment of malingered memory deficits (e.g., Binder, 1993; Frederick, Carter, & Powel, 1995; Hiscock & Hiscock, 1989; Iverson, Franzen, & McCracken, 1994; Slick, Hopp, Strauss, Hunter, & Pinch,

1994). Some recent adaptations have retained the SVT format, but have practically eliminated classification based on binomial probabilities (e.g., Tombaugh, 1997).

INITIAL DEVELOPMENT OF THE VALIDITY INDICATOR PROFILE

The Validity Indicator Profile (VIP) was originally designed as a detector of malingered cognitive impairment (Frederick & Foster, 1991). The VIP modified SVT by establishing a hierarchy of difficulty across trials. This means that a "performance curve" representing average correct responses by average item difficulty can be generated, demonstrating the average performance of the test taker across an increasingly difficult range of test items. Because the VIP retains a two-alternative forced choice (2AFC) format, the expected progression of performance is from 100% correct for easy items to 50% correct for more difficult items (see Fig. 1, top

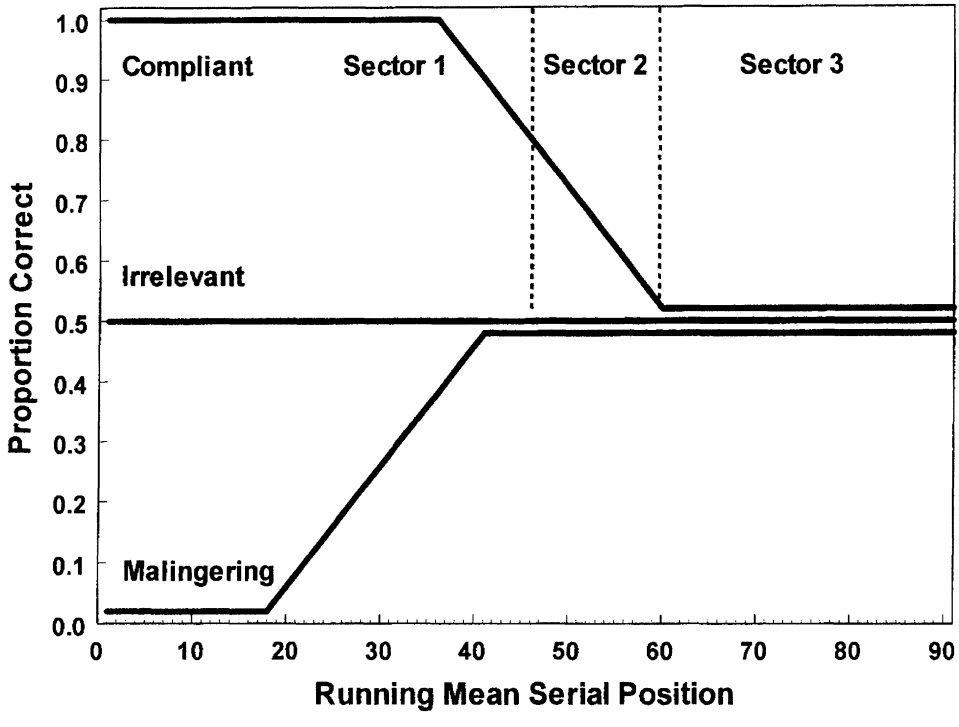

Fig. 1. Examples of performance curves representing three different response styles on the Validity Indicator Profile (VIP). The top line represents compliant responding; the test taker has evidenced consistently good effort to respond correctly. Performance begins at 100% for easy items and reduces to 50% at the test taker's ceiling of ability. The middle line represents irrelevant responding; the test taker's response are irrelevant to item content. This style of responding indicates a token effort to respond incorrectly. The performance curve consistently remains at 50%. The bottom line represents malingering; the test taker has demonstrated a strong effort to respond incorrectly. The test taker consistently incorrectly answered items that were solvable, but improved to 50% at the test taker's ceiling of ability. The shaded region represents the expected range of running mean values generated by irrelevant responding.

line). This results in a performance curve with a standard shape for compliant test takers across a broad range of capacity to answer correctly. Individuals with a relatively low capacity to solve the problems will approach 50% responding sooner than their counterparts with higher ability, but their curves will be similarly negatively sloped. Frederick and Foster (1991) found that a positive slope of the performance curve identified instances of motivation to perform poorly. Because the lower limit for response accuracy for items that could not be solved is 50%, individuals who intentionally chose incorrect answers for items that they could solve generated positively sloped performance curves (see Fig. 1, bottom line).

In addition, the hierarchy of difficulty allowed an examination of the consistency of responding. Frederick and Foster (1991) identified a means to compute the consistency of responding for items of equivalent difficulty while accounting for the random responding that inevitably occurs once an individual is no longer to respond correctly. A decision rule incorporating the product of the consistency measure and the performance curve slope provided an effective way to classify test takers as compliant or non-compliant. Frederick, Sarfaty, Johnston, & Powel (1994) suficcessfully cross-validated this classification rule (which has been substantially modified for the VIP) for cognitive malingering, but Rose, Hall, & Szalda-Petree (1998) reported the classification rate of the early rule was ineffective.

ITEM DEVELOPMENT

The VIP consists of two subtests; each can be administered and scored separately. The VIP nonverbal subtest (VIP-NV) was developed by modifying the items on the original Test of Nonverbal Intelligence (TONI; Brown, Sherbenou, & Johnsen, 1982). The TONI presents 100 picture-matrix problems on two equivalent forms that require simple matching, complex matching, analogous decision making, progression, addition, subtraction, and abstraction. Originally, the TONI presented four to six alternatives for selection for each trial in a standard format of increasingly difficult problems. The VIP-NV includes all 100 items from both forms of the TONI; however, the items were modified to include only two alternativesthe correct solution and one foil. Presentation was modified so that all items were required to be solved; presentation order was random with respect to item difficulty.

The VIP verbal subtest (VIP-V) consists of 78 word-definition problems. Test takers are presented with a stimulus word (e.g., *carpet*) and are asked to choose one of two possible answers that is more similar in meaning to the stimulus (e.g., *rug* or *shoe*). Like the VIP-NV, the items have a hierarchy of difficulty and are presented randomly with respect to item difficulty.

VIP INDICATORS OF RESPONSE VALIDITY

Once the VIP subtests have been administered, the items are scored and then reordered by difficulty. The derivation of item difficulty was based on prior normative samples and is described in Frederick (1997). Items that have been solved

correctly are assigned a "1"; items solved incorrectly are assigned a "0." The performance curve is derived and indices of consistency are computed from the reordered scored responses.

Performance Curve Measures

To derive the performance curve, *running means* are computed by averaging a set of 10 consecutive scored item responses. Responses to items 1 through 10 are averaged to yield the first running mean, representing the individual's average performance on the 10 easiest items. Responses 2 through 11 are averaged to compute the second running mean. This process continues until the last (most difficult) item has been included in a running mean. Table 1 shows an example of how running means are computed for 20 items that have been reordered by difficulty. The performance curve is obtained by plotting the value of the running mean (which ranges from 0.0 to 1.0) on the vertical axis against its serial position. To compute the slope of this performance curve, the best linear (i.e., straight-line) representation of the performance curve is derived. *Slope* is defined as the slope of the linear regression line fit to the performance curve. *Curvature* is a measure of the extent to which the performance curve changes after an initial stable (i.e., straight-line) pattern. Curvature is the value of the regression weight (b_2) in a nonlinear regression equation for predicting performance from item difficulty:

$$Y = a + b_1X + b_2X$$

where Y is the value of the running mean (performance) and X is the corresponding serial position of the running mean (item difficulty).

Consistency Measures

Consistency Ratio

The consistency ratio (CR) is an index of the extent to which an individual answers items of comparable difficulty correctly. That is, the CR assesses response consistency in terms of similar responding to pairs of items of comparable difficulty. Two measures are computed to derive the CR.

Equivalent item pairs. Equivalent item pairs (EIPs) are derived after the items are ordered by difficulty. Contiguous pairs, beginning with items 1 and 2, are considered equivalent in difficulty. There are 50 EIPs for the VIP-NV and 39 EIPs for the VIP-V. If an item pair contains two correct answers, that pair's EIP value is 1. It the item pair contains one or two wrong answers, the pair's EIP value is 0.

Adjusted Score. The adjusted score is an estimate of the number of items that could be *solved* by the test taker. On any 2AFC test, the total number of items *answered correctly* typically overrepresents the actual number of items the test taker was able to solve. That is, the total number of items correctly answered (total score) makes no distinction between items that the test taker answered correctly based on knowledge and items that he or she answered correctly by guessing. The adjusted score employs a standard correction for guessing described by Cronbach (1990,

Table 1. An Example of How Running Means Are Computed to Derive a Performance Curve

Item number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Item response weight	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	0	1	0	0
Running mean serial position										1	2	3	4	5	6	7	8	9	10	11
Running mean value										1	1	1	1	.9	.9	.8	.7	.7	.6	.5

Note. The item response weight is "1" for correct responses and "0" for incorrect responses. The item response weights for items 1 through 10 have a mean of 1.0. The running mean serial position for these ten items is "1" and its value is 1.0. The running mean value at serial position 11 has a value of 0.5. This reflects the contribution of 5 correct response and 5 incorrect responses for items 11 through 20.

p. 67). The correction is to subtract the number of wrong answers from the number of right answers. For example, consider an individual who earned a total score of 75 (number of right answers) on the VIP-NV. To derive the adjusted score, subtract 25 (number of wrong answers) from 75 to get 50. This adjusted score estimates that the person knew the answer to 50 items and earned a total score of 75 by correctly guessing on 25 of the remaining 50 items.

The CR is derived by dividing the subtest's EIP value by the maximum EIP value (EIP_{max}) possible given the test taker's adjusted score. EIP_{max} is computed as follows:

$$EIP_{max} = (\text{Adjusted Score}/2) + ([N - \text{Adjusted Score}]/4)$$

where N is the number of items on the subtest. Once EIP_{max} is computed, CR is computed:

$$CR = EIP/EIP_{max}$$

CR values can range from 0 to 1.0, with higher values indicating greater response consistency between comparable items.

Norm Conformity Index

The Norm Conformity Index (NCI; Tatsuoka & Tatsuoka, 1982, 1983) expresses the extent to which an individual's responses resemble a pattern where all correct responses precede all incorrect responses. To compute NCI, a vector is formed comprising the string of 0's and 1's representing the examinee's responses in their order of difficulty. The NCI is computed as follows:

$$\text{NCI} = (2U/V) - 1$$

where U is the sum of the number of 0's to the right of each 1 in the vector and V is the product of the number of 0's and 1's in the vector. NCI values range from −1.0 to 1.0, with higher values indicating greater response consistency.

Individual Consistency Index

As implemented in the VIP, the Individual Consistency Index (ICI; Tatsuoka & Tatsuoka, 1982, 1983) examines response consistency with respect to an *individual's* average response pattern across five parallel sets of items (the NCI examines response consistency with respect to a normative group). The first set of items on the VIP-NV contains responses to items 1, 6, 11, and so forth, through item 96 (when arranged by difficulty). The fifth set of items contains responses to items 5, 10, 15, and so forth, through item 100. The second set begins with item 2, the third set begins with item 3, and the fourth set begins with item 4, progressing to every fifth item thereafter. Each set contains the response (0 or 1) to 20 items. The VIP-V contains five sets for the first 75 items; to ensure that each set has the same number of items, items 76–78 are not used. The ICI is essentially the average NCI across the five sets of items. The ICI has been shown to identify examinees who

are not performing consistently over parallel sets of items. ICI values range from −1.0 to 1.0, with higher values indicating greater response consistency.

Interaction Scores

Interaction scores—Score by Correlation (SCC) and Slope by Consistency Ratio (SLCR)—combine performance characteristics from separate measures. Each interaction score is derived by the product of its constituent measures. The decision rule for SLCR (Frederick & Foster, 1991; Frederick et al., 1994) was modified in these validation studies.

DEVELOPMENT PHASE

Participants

The development sample included clinical and nonclinical groups. Participants in the nonclinical group (n = 944) were college students and employees of National Computer Systems, Inc. (NCS). Informed consent was obtained from all nonclinical participants. Clinical participants (n = 104) were adults undergoing neuropsychological evaluation. Some were involved in litigation; others were being evaluated as part of ongoing clinical care. Clinical participants were informed during the initial interview that they would receive a wide range of tests, including tests that measure motivation and effort. All clinical participants agreed to allow their results to be used for test development if their anonymity was maintained. All clinical participants were administered an individualized battery of tests in addition to the VIP. Demographic characteristics of the participants can be found in Table 2.

Instruments

One hundred thirty participants took both VIP subtests, 909 participants took the nonverbal subtest, and 269 participants took the verbal subtest. In addition, some participants were administered the Rey malingering tests. These tests are reviewed briefly (for a detailed review, see Frederick, 1997, or Frederick, Crosby, & Wynkoop, in press). Cutoff scores for the Rey malingering tests were based on a taxometric analysis (Meehl, 1995) reported by Frederick (1995).

Rey 15-Item Memory Test

The Rey 15-item Memory Test (RMT; originally described in Rey, 1958) comprises five rows of three sequential items (e.g., 1, 2, 3; circle, square, triangle). Examinees were told to remember all 15 items during a 10-second exposure. After the stimulus items were removed, examinees were told to reproduce the items on a blank sheet of paper. The score is the number of items recalled correctly. Scores range from 0 to 15. Low scores (i.e., < 9 items) are consistent with an intention to perform poorly.

Table 2. Demographic Characteristics of the Validity Indicator Profile Development Sample

	Nonverbal sample (n = 909)			Verbal sample (n = 269)	
	Nonclinical		Clinical	Nonclinical	
	Honest normals (n = 336)	Coached normals (n = 469)	Brain injured (n = 104)	Honest normals (n = 137)	Coached normals (n = 132)
Gender					
Male	137	148	62	55	46
Female	197	320	42	80	85
Not indicated	2	1	0	2	1
Age group					
15–17	0	2	2	0	0
18–25	102	189	14	89	92
26–45	47	25	58	34	29
46–65	6	7	28	8	7
66–71	0	0	2	0	0
Not indicated	181	246	0	6	4
Race					
White	295	408	96	122	116
Black	23	33	4	8	7
Hispanic	4	7	3	1	1
Asian	6	8	1	2	0
Native America	2	8	0	2	2
Other	6	4	0	1	5
Not indicated	0	1	0	1	1
Education					
Grade 8 or less	0	1	3	0	1
Grades 9–11	0	0	20	0	1
High school graduate	6	3	33	22	23
1–3 years of college/technical school	303	444	20	87	82
4 or more years of college	27	20	28	28	24
Not indicated	0	1	0	0	1

Word Recognition Test

The Word Recognition Test (WRT; Rey, 1941) comprises two word lists: one list of 15 words (stimulus list) and the other list of 30 words (memory test). The memory test contains the 15 stimulus words and 15 distractors. In our administration, the stimulus list was read to the examinee. The memory test was then read and the examinee was instructed to say "Yes" if a word was recognized as being on the stimulus list and "No" if it was not. The score was derived by adding the number of correctly recognized words to the number of correctly rejected words (i.e., total correct response). Scores range from 0 to 30. Low scores (i.e., < 18) indicate motivation to respond incorrectly.

Dot Counting Test

The Dot Counting Test (DCT; Rey, 1941) consists of 12 3 $\times$ 5 cards on which have been placed either random (ungrouped) or patterned (grouped) dots. The

cards were presented to examinees who are instructed to count the dots as quickly as possible without making mistakes. The absolute difference between the correct answer and the subject's response was summed over the 12 cards. The response time for each card was summed for grouped and ungrouped categories. The score used in this study (following Frederick, 1995) was computed by multiplying the absolute number of errors (plus 0.5 to avoid scores of zero) by the ratio of grouped time to ungrouped time. Scores can range from near zero to infinity. However, to control for outliers, the maximum score was set at 15. High scores (i.e., >8) indicate a motivation to perform poorly.

Procedure

Nonclinical participants were assigned to "compliant" or "noncompliant" groups. Instructional sets and detailed information about group assignment is presented in Frederick and Foster (1991) and Frederick et al. (1994). Compliant participants were instructed to give their best effort when completing tests. Noncompliant participants were instructed to fake believable impairment without being obvious. Some of the noncompliant participants (n = 139) were provided with strategies to avoid detection. Some of the noncompliant participants (n = 177) were offered $20 if they could feign believable impairment without being detected.

The task of assigning clinical patients to criterion groups was more problematic because no independent and objective measures existed for classifying patients as compliant or noncompliant. Instead, these classifications were made using (1) the individual's performance on the Rey malingering tests or (2) *a priori* clinician ratings regarding the likelihood that the patient would malinger during assessment. Based on these criteria, 65% of the clinical patients were assigned to the compliant group and 35% were assigned to the noncompliant group.

Rey Test Performance

Clinical patients who earned a malingering score on two or more of the Rey tests were included in the noncompliant group. Patients who earned acceptable scores on all three tests were included in the compliant group. Equivocal cases (one positive score, two negative scores) were not classified by Rey test performance as compliant or noncompliant.

Clinician Ratings

Clinician evaluations regarding the potential for malingering were based on subjective impressions from clinical interviews prior to testing. Patients were rated from 0% to 100% on the likelihood that they would feign cognitive impairment on neuropsychological tests. Patients whose ratings were 50% or higher were considered likely to be noncompliant test takers. Patients with ratings of 10% or lower were considered likely to be compliant. Patients with ratings from 11% to 49% were not classified by clinician rating as compliant or noncompliant.

Results And Discussion

Based on this scheme for classification of participants, decision rules to classify performance as valid or invalid were constructed for six VIP measures (CR, NCI, ICI, curvature, SCC, and SLCR; collectively known as *primary validity indicators*). Cut scores were set so that 90% of the compliant participants would be classified correctly by the VIP (i.e., a 10% false-positive rate). This is a distinctly different approach than some other malingering tests that minimize false positives at the cost of sensitivity (e.g., the Structured Interview of Reported Symptoms, SIRS; Rogers, Bagby, & Dickens, 1992). We considered false negatives as problematic as false positives and chose not to sacrifice sensitivity for gains in specificity. The resulting cutoff values and decision rules are reported in Table 3. Because the consistency measures (i.e., CR, NCI, and ICI) individually tended toward oversensitivity, we required two positive cutoff scores for these consistency measures to constitute one *rule* violation. Three positive consistency scores counted as two rule violations. A positive score on any of the other three measures constituted one rule violation. Two or more rule violations out of the five rules (positive cutoff values) resulted in classification as "invalid." This classification scheme was then cross-validated.

Table 3. Decision Rule Cutoff Scores for the Primary Validity Indicators

	Nonverbal subtest	
Primary validity indicator	Critical conditions and cutoff scores	Number of rule violations
Consistency Ratio (CR)	≤ 0.779	[a]
Norm Conformity Index (NCI)	Total score <75 and NCI ≤ 0.683	[a]
Individual Consistency Index (ICI)	Total score <75 and ICI ≤0.816	[a]
Score by correlation	≥−31.79	1
Slope by consistency ratio	Total score <75 and slope by CR ≥−0.00700	1
Curvature	Total score <75 and curvature ≥−0.00011	1
	Verbal subtest	
Primary validity indicator	Critical conditions and cutoff scores	Number of rule violations
Consistency Ratio (CR)	≤0.821	[a]
Norm Conformity Index (NCI)	Total score <59 and NCI ≤0.758	[a]
Individual Consistency Index (ICI)	Total score <59 and ICI ≤0.846	[a]
Score by correlation	≥−28.87	1
Slope by consistency ratio	Total score <59 and slope by ≥−0.00998	1
Curvature	Total score <59 and curvature ≥−0.00026	1

[a]The CR, NCI, and ICI are used in combination to determine the number of rule violations. When one or none of the measures exceed the specified cutoff score, no rules are considered to be violated. When two measures exceed the cutoffs, one rule is considered to be violated. When three measures exceed the cutoffs, two rules are considered to be violated.

CROSS-VALIDATION OF THE PRIMARY VALIDITY INDICATORS

Participants

The cross-validation sample consisted of nonclinical adults, patients with traumatic brain injury (TBI), computer-generated cases of random responding, and patients at risk for malingering cognitive deficits. Participants were assigned to the compliant or noncompliant criterion groups on an a priori basis. Demographics for this group are presented in Table 4.

Nonclinical participants were 152 community members who were between the ages of 18 and 69 and spoke English as a primary language. They did not have a history of TBI, learning disability, or severe substance abuse and were not currently in treatment for a mental disorder. Individuals were recruited to participate in the study as part of a school fundraiser; the school received $40 for each complete set of test data. One hundred served as compliant test takers; 52 served as noncompliant test takers. They completed the VIP, the Rey malingering tests, and the Portland Digit Recognition Test (PDRT; Binder, 1993).

TBI patients (n = 61) were recruited from group homes and private practice clinics. They were included if they were between the ages of 18 and 69; spoke English

Table 4. Demographic Characteristics of the Validity Indicator Profile Cross-Validation Samples

	Nonclinical (N = 152)		Clinical (N = 110)	
	Honest normals (n = 100)	Coached normals (n = 52)	Brain injured (n = 61)	Suspected malingerers (n = 49)
Gender				
Male	49	20	48	20
Female	51	32	13	29
Age group				
18–25	13	9	25	1
26–45	61	35	30	32
46–65	26	8	6	16
Race				
White	82	40	49	38
Black	8	6	3	5
Hispanic	6	5	4	4
Asian	1	1	0	0
Native American	2	0	4	1
Pacific Islander	1	0	1	1
Education				
Grade 8 or less	0	0	2	0
Grades 9–11	1	0	8	3
High school graduate	13	7	21	22
1-3 years of college/technical school	30	18	17	13
4 or more years of college	54	27	13	11
Not indicated	2	0	0	0

as a primary language; had sustained a closed head injury within the past 36 months; had their brain injury documented by computed tomography (CT), magnetic resonance imaging (MRI), or electroencephalogram (EEG), had a Glascow Coma Scale of less than 13 at admission; and had symptoms consistent with medical history. They were excluded if they had ongoing posttraumatic amnesia, a history of mental retardation prior to injury, a coincident dementing condition, active substance intoxication or active substance dependence causing cognitive impairment, or a history of a learning disability. TBI patients were paid $40 for their participation.

Suspected malingerer(n = 49) were tested as part of a clinical evaluation. As part of the standard informed consent process in each clinical setting, examinees agreed to allow their test scores to be used for test development or other clinical research. Clinical patients were considered at risk for malingering if they were between the ages of 18 and 69, claimed cognitive impairment, were seeking financial remuneration for injury, had normal findings on medical examination, and presented with improbable or exaggerated symptoms with respect to their medical history. They were excluded if they had an abnormal CT, MRI, or EEG (not including SPECT or QEEG); a history of loss of consciousness for 5 minutes or longer; any focal neurological disorder; a history of brain trauma or cerebral disease prior to the injury in question; or active substance intoxication.

Fifty VIP-NV and 50 VIP-V response sets were generated by computer. They were entered into analyses as noncompliant participants.

Procedure

Clinical and nonclinical participants were assigned to criterion groups prior to test administration. The nonclinical participants were randomly assigned to a criterion group and asked to complete the battery of tests according to specific instructions. Noncompliant subjects received their instructions by videotape with an appeal to their social conscience to help develop effective malingering-detection tests. After delivery of the instructions, participants were given the opportunity to review

Table 5. Analysis of Test Classification for the Nonverbal and Verbal Subtests of the Validity Indicator Profile Cross-Validation Sample

	Test classification							
	Nonverbal subtest				Verbal subtest			
	Valid		Invalid		Valid		Invalid	
Criterion group	*n*	%	*n*	%	*n*	%	*n*	%
Compliant group								
Honest normals	93	93	7	7	95	95	5	5
Brain injured	45	74	16	26	38	63	22	37
Noncompliant group								
Coached normals	13	25	39	75	14	27	38	73
Suspected malingerers	27	55	22	45	35	73	13	27
Random responding	0	0	50	100	0	0	50	100

at length pamphlets detailing the nature and symptoms of brain injury. They were also offered an additional $50 for the school fundraiser if they could fake believable impairment and not be detected by the tests they received. One TBI patient and one suspected malingerer did not provide a sufficient number of responses on the VIP-V for inclusion.

Results

Classification rates for the VIP-NV and VIP-V are reported in Table 5. The VIP-NV demonstrated an overall classification rate of 79.8%, with 73.5% sensitivity and 85.7% specificity. The VIP-V demonstrated an overall classification rate of 75.5%, with 67.3% sensitivity and 83.1% specificity. Table 6 reports the rate of agreement between the VIP-NV and VIP-V (n = 310, kappa = .70, $p < 0.001$). Breakdowns of the subclassification as careless, irrelevant, or malingering are given for the VIP-NV and VIP-V in Table 7. Table 8 reports the mean total score for each subtest based on the subclassification as careless, irrelevant, or malingering.

Table 9 compares the classification rates of the VIP subtests with the Rey malingering tests and the PDRT. Two hierarchical logistic regression analyses were performed to compare the incremental validity of the PDRT with the VIP. For both analyses, the Rey tests were entered on the first step (model chi-square = 22.7, df = 3, $p < 0.0001$) and correctly classified 66.7% of the cases. On the second step, either the PDRT or VIP (VIP-NV and VIP-V) classifications were entered. On the third step, the remaining test was entered. When the VIP classifications were entered as the third step, added to the PDRT, correct classifications improved from 69.8% to 77.3% (improvement chi-square = 41.0, df = 2, $p < 0.0001$). In contrast, when the PDRT classifications were entered as the third step, added to the VIP, correct classifications improved only from 76.5% to 77.3% (improvement chi-square = 7.7, df = 1, $p < 0.01$). Hierarchical logistic regression analyses in which the Rey tests were entered as the third step were nonsignificant and demonstrated no improvement over the PDRT and VIP classifications, which together correctly classified 77.3% of the cases.

Discussion

Classification rules for the primary validity indicators were defined in the validation phase so that the false-positive rate (equal to 1.0 minus specificity) would

Table 6. Cross-Tabulation of Validity Indicator Profile Verbal and Nonverbal Subset Classification Rates

	Verbal subtest	
Nonverbal subtest	Invalid	Valid
Invalid	157	21
Valid	25	107

Note. n = 310. Kappa = .70, $p < 0.001$.

Table 7. Analysis of Response Style Classification of Invalid Performances on the Validity Indicator Profile Cross-Validation Sample

	Response style classification											
	Nonverbal subtest						Verbal subtest					
	Careless		Irrelevant		Malingered		Careless		Irrelevant		Malingered	
Criterion group	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Compliant group												
Honest normals	7	7	0	0	0	0	5	5	0	0	0	0
Brain injured	14	23	2	3	0	0	19	32	3	5	0	0
Noncompliant group												
Coached normal	20	38	16	31	3	6	26	50	8	15	4	8
Suspected malingerers	18	37	4	8	0	0	11	23	2	4	0	0
Random responding	1	2	48	96	1	2	1	2	48	96	1	2

Table 8. Mean Validity Indicator Profile Total Score by Subclassifications of Response Style for the Nonverbal and Verbal Subtests

Subgroup	Nonverbal subtest			Verbal subtest		
	n	*M*	*SD*	*n*	*M*	*SD*
Valid cases						
Compliant criterion groups						
Normal controls (CV)	93	89.5	6.0	95	64.9	3.9
Normal controls (V)	205	87.7	5.7	95	62.6	3.9
Brain injured (CV)	45	84.4	6.0	38	62.1	3.2
Nonlitigating neuropsych (V)	26	81.9	7.4	—	—	—
Litigating neuropsych (V)	28	82.5	5.5	—	—	—
Noncompliant criterion groups						
Analog malingerers (CV)	13	84.7	4.9	14	63.0	4.8
Analog malingerers (V)	20	80.1	5.3	19	59.9	3.7
Suspected malingerers (CV)	27	85.2	6.8	35	63.0	4.8
Nonlitigating neuropsych (V)	11	89.4	8.2	—	—	—
Litigating neuropsych (V)	6	82.8	4.5	—	—	—
Invalid cases						
Careless cases						
Normal controls (CCG, CV)	7	61.9	13.8	5	56.6	1.1
Brain injured (CCG, CV)	13	62.8	8.0	19	53.8	3.2
Analog malingerers (NCG, CV)	20	62.4	8.3	26	49.1	6.4
Suspected malingerers (NCG, CV)	18	61.5	7.7	11	52.4	5.0
Irrelevant cases						
Brain injured (CCG, CV)	3	60.3	9.0	3	34.0	6.6
Analog malingerers (NCG, CV)	18	52.3	11.0	9	37.6	10.1
Suspected malingerers (NCG, CV)	4	56.8	6.4	2	39.9	4.2
Malingering cases						
Analog malingerers (NCG, CV)	1	23.0	—	3	25.3	3.5

Note. CCG = compliant criterion group; NCG = noncompliant criterion group; V = validation study; CV = cross-validation study.

be approximately 10%. This value increased slightly on cross-validation (as would be expected) to 14% for the VIP-NV and 17% for the VIP-V. The VIP demonstrated incremental validity over the Rey tests and the PDRT in this sample. The VIP subtests demonstrated greater sensitivity than the Rey malingering tests and the PDRT, although these retained higher specificity.

The lack of perfect agreement between the VIP-NV and VIP-V and the significantly different rates of agreement among all malingering tests suggests that some noncooperative individuals may not choose to perform poorly on all tests adminis-

Table 9. Comparison of Classification Rates Among Malingering Tests Within Cross-Validation Sample

Test	Sensitivity	Specificity	Overall classification rate
VIP Nonverbal Subtest	73.5	85.7	79.8
VIP Verbal Subtest	67.3	83.1	75.5
Portland Digit Recognition Test	17.0	99.4	67.7
Rey 15-Item Memory Test	4.9	97.5	61.4
Rey Word Recognition Test	8.8	100.0	64.6
Rey Dot Counting Test	11.8	97.5	64.1

tered. For example, some individuals may wish to be seen only as impaired in memory, but not in general cognitive capacity. Furthermore, attention and concentration may be adequate early in a test session, but may decrease later on. Consequently, a lack of perfect agreement among tests in a clinical situation warrants close scrutiny to determine the source of invalidity.

Most members of the compliant criterion groups (98.1% to 98.8%) were classified as "motivated to respond correctly" (compliant or careless) on the VIP-NV and VIP-V, respectively. About half of the noncompliant criterion group members (42.0% to 47.7%) were classified as "motivated to respond incorrectly" (irrelevant or malingering) on the VIP-V and VIP-NV, respectively. These values are very close to the specificity (99.5%) and sensitivity (48.5%) reported for the SIRS (Rogers et al., 1992, p. 24).

Differential Rates of Sensitivity and Specificity

Differential rates of classification for subgroups warrant comment. The overall reported rates of sensitivity and specificity are dependent on the rates of subgroup membership within the compliant and noncompliant groups. Within the compliant group, analog participants were classified more accurately than brain injured participants (93% to 95% specificity vs. 63% to 74% specificity). Within the noncompliant group, randomly generated performances were perfectly classified for both the VIP-NV and VIP-V (100% sensitivity). Analog noncompliant participants were classified more accurately than clinical participants at risk for malingering (73% to 75% sensitivity vs. 27% to 45% sensitivity). Consequently, had we not used roughly equivalent numbers for each of our subgroups, the reported overall values of sensitivity and specificity might have easily been swayed by one subgroup or another.

The explanation most favorable to the VIP for the differential rates of classification among subgroups is that these differences reflect the differential rates of purity in the criterion groups, especially for the noncompliant subgroups. This explanation assumes that the derived VIP classification rates accurately estimate the true ratios of compliance and noncompliance within the subgroups and that some noncompliant subgroups were not totally composed of invalid responders.

Within the noncompliant category, the randomly generated VIPs can be construed as the purest subgroup; all cases were known *a priori* to be invalid. The rate of classification as "invalid" was 100%. The coached normals were probably the group with next highest level of purity. They were volunteers who ostensibly agreed to try their best to feign believable impairment. They received extensive instructions on what their task was and were given the opportunity to prepare for their roles by reading literature on brain injuries. They were given the opportunity to generate individually a meaningful sum of money that would benefit a cause they had already shown interest in promoting. Nevertheless, it is commonly known that not all individuals who ostensibly agree to pretend to malinger will do so (e.g., Goebel, 1983). The rate of sensitivity for the VIP for this group was 73% to 75%.

The criterion group with the most potential contamination (least purity) comprised the clinical examinees who were considered at risk for malingering. No concurrent measure of malingering was used to establish this group. Membership was primarily determined by subjective judgments on the part of the evaluator

regarding inconsistencies between probable injury and reported symptoms. The rates of classification as "invalid" for this group were 27% (VIP-V) and 45% (VIP-NV).

The explanation least favorable to the VIP is that true malingerers (as represented by the "at-risk" group) had greater motivation to feign impairment without being detected; hence, they were better at hiding their malingering than were analog participants. A review of their mean total scores in Table 8 does not support this argument. Individuals suspected of being "at risk" for malingering, who nevertheless were classified as compliant by the VIP, earned total scores (VIP-NV, $M = 85.2$, $SD = 6.8$, $n = 27$; VIP-V, $M = 63.0$, $SD = 4.8$, $n = 35$) well above "at-risk" individuals classified as careless (VIP-NV, $M = 61.5$, $SD = 7.7$, $n = 18$; VIP-V, $M = 52.4$, $SD = 5.0$, $n = 11$) or irrelevant (VIP-NV, $M = 56.6$, $SD = 6.4$, $n = 4$; VIP-V, $M = 39.9$, $SD = 4.2$, $n = 2$). Suspected malingerers rated as "compliant" clearly generated strong effort to respond correctly on the VIP-NV and VIP-V. Consequently, the argument that this comparison group contained substantial contamination (many nonmalingering individuals) would seem to have merit.

Subclassifications of Invalid Responding

Dichotomous malingering classifications traditionally limit explanations of positive test scores to (1) test-taking noncompliance or (2) classification error. A classification scheme should be able to describe a performance as suboptimal (not representative of the individual's maximal capacity to perform well on cognitive tasks) without mechanically concluding that the invalidity was intentional (noncompliant). This is why positive scores on the VIP result in classification of "invalid" instead of "noncompliant." "Invalid" means only that evidence exists to indicate that testing does not likely represent the maximal capacity of the test taker to respond accurately. The VIP cross-classifies effort with intention to examine other factors besides intentionality that might explain poor performance (Fig. 2). Low effort (alternatively, token effort, compromised effort) to respond correctly is classified as "careless." High effort to respond incorrectly is categorized as "malingering." Token effort to respond incorrectly is classified as "irrelevant." Construct validation of these classifications has been reported in Frederick et al. (in press) and are based primarily on features of the performance curve.

Performance Curve Characteristics

Following are the categorizations with the "invalid" classification and the performance curve characteristics that are the basis for categorization.

Malingering: Slope

The slope of the performance curve is the incline of the line of best fit based on simple least squares linear regression (Cohen & Cohen, 1983) and reflects the rate of change in percentage correct as average item difficulty increases. Positively sloped performance curves indicate that the individual's performance improves as items become more difficult. This is completely counterintuitive and indicates that

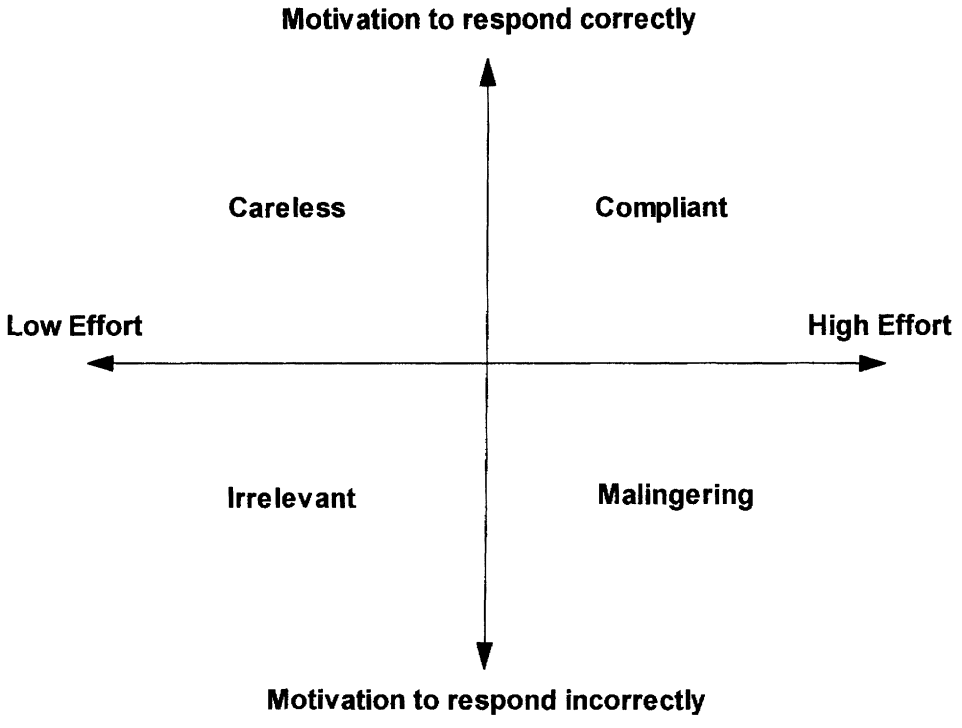

Fig. 2. A cross-classification of motivation and effort. Motivation refers to the intended goal of the test taker; to respond correctly or respond incorrectly. Effort refers to the intensity of application to produce the intended goal (low or high). High effort to respond correctly represents compliant responding. Low effort (or impaired effort) to respond correctly represents careless responding. High effort to respond incorrectly reflects malingering. Token effort to respond incorrectly represents irrelevant responding.

the individual has suppressed knowledge of the true answer for items that he or she can solve and has performed better (i.e., at about 50%) only when he or she can no longer has the capacity to choose the incorrect answer (i.e., at the ceiling of ability). A performance curve that represents malingering is represented by line C in Fig. 1. When the value of the slope is steeply positive (operationally, when the slope exceeds 0.0045), the performance is classified as "malingering."

Point of Entry

The first running mean is the beginning point of the curve and is referred to the *point of entry* (POE). The POE represents the average performance on the 10 easiest items and is used in the categorization of invalid performances. On the VIP, the proportion of correct responses for the 10 easiest items is compared to that expected by random responding. POEs ≥ 0.8 most likely represent a capacity to determine correctly the correct answers to the 10 easiest items because of the low cumulative probability ($p = 0.055$) of earning 8 ($p = 0.044$), 9 ($p = 0.010$), or 10 ($p = 0.001$) correct answers out of 10 items by chance (based on binomial expansion; Hays, 1973). Likewise, POEs of ≤ 0.2 are equally improbable for random responding

and more likely represent the capacity to correctly determine the correct answer to the 10 easiest items, despite an intention to respond incorrectly. POEs of 0.3–0.7 are consistent with random responding (i.e., responding irrelevant to item content). Line B of Fig. 1 represents irrelevant performance. If a curve is classified as invalid but not as malingering, the curve is classified as irrelevant if the POE is ≤0.7. Otherwise, invalid performances are classified as "careless."

Careless Performance Curves

The features of the performance curve that support classification of carelessness are based in an analysis of different segments of the performance curve.

Sector Analysis. Performance curves that have POEs that exceed 0.7 (not classified as malingering or irrelevant) are divided into three *sectors* (see Fig. 3). The first performance curve sector (sector 1) begins at the POE and continues to the first instance of 0.7, which is the first indication that the test taker is approaching the ceiling of his or her ability. These running mean values represent sustained

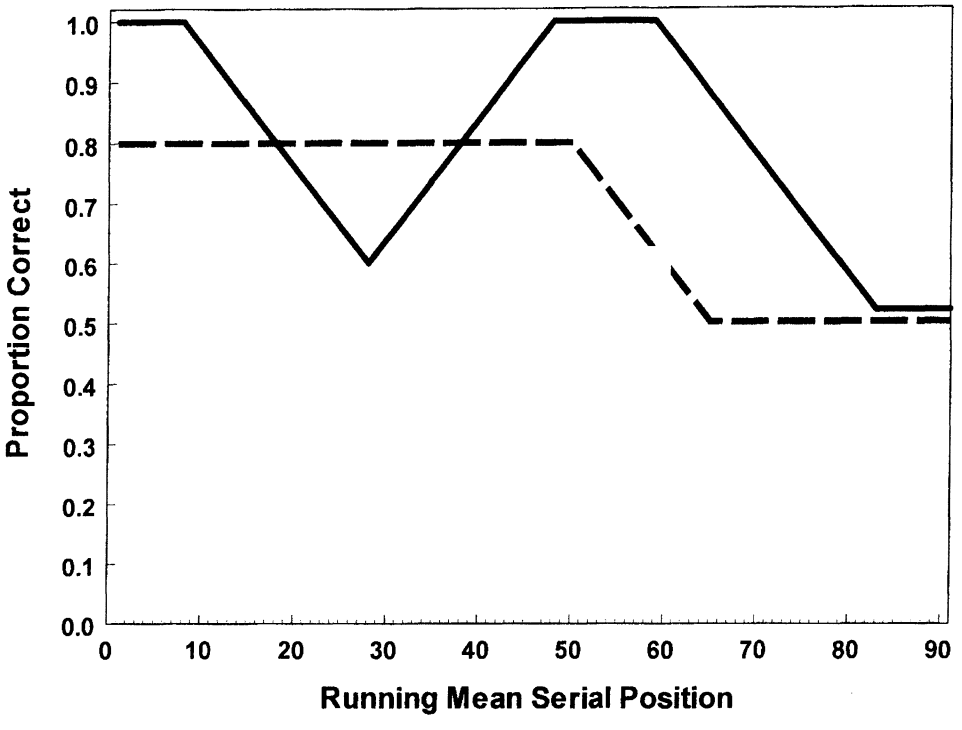

Fig. 3. Examples of performance curves representing careless responding on the VIP. In solid line A, the test taker appears to have responded perfectly to many items that were easy and difficult. Nevertheless, some items of intermediate difficulty were answered incorrectly. Given the extent of perfect responding for many difficult items, the relatively poorer performance for items of intermediate difficulty probably represents an instance of poor attention or less than optimal effort. In dashed line B, there is no occasion of perfect responding. Given that the test taker consistently demonstrates imperfect, but better than chance, performance for relatively difficult items, the less than perfect responding for much easier items probably reflects a token effort to respond correctly for those items.

"better-than-chance" performance. The greater the distance of sector 1 (i.e., the more running means in sector 1), the greater the ability of the test taker is assumed to be. When compliant test takers reach the ceiling of their ability, their performance begins to approach random responding, first decreasing to a running mean of 0.7 and then continuing downward to running means averaging 0.5, with a range primarily of 0.3–0.7.

The portion of the curve that transitions from sector 1 to random responding is referred to as sector 2. *Sector 2* distance is the length of the curve from the first occurrence of 0.7 until the first occurrence of 0.5. *Sector 3* distance is the length of the curve from the first occurrence of 0.5 to the last running mean. Random responding is expected within sector 3.

Comparison of Sector Distances. The relationship between sector 1 distance (which reflects ability) and sector 2 distance (which reflects the transition to random responding) is used to distinguish between careless and compliant responding on the VIP. When sector 1 is longer than sector 2, there is clear evidence of an effort to answer at least some items correctly. For individuals with high ability, sector 1 should typically be much longer than sector 2. For individuals with lower ability, the relative length of sector 1 to sector 2 will be reduced. When sector 2 distance exceeds that of sector 1 (Fig. 2, line A), there is good reason to believe that the individual did not properly attend to items within his or her range of ability and, consequently, sector 2 was prematurely initiated.

Sector 1 Residual. A quantitative analysis of errors is provided by the Sector 1 residual. Sector 1 residual refers to the extent to which running means in Sector 1 deviate from perfect performance (perfect performance is a running mean of 1.0). Errors within Sector 1 (except at the point of transition to Sector 2) are unexpected because, by dint of the extent of ability *already demonstrated by the examinee,* the examinee should be capable of answering the missed items correctly. For example, if the examinee correctly answered items 11 through 20 (when ordered by difficulty), it would be hard to explain incorrect responses to items 3 and 4, which are much easier to solve. To compute the sector 1 residual, each deviation from 1.0 is assigned a weight based on the difficulty of the item (deviations for easier items are weighted more heavily than deviation on difficult items). Weighted deviations are then averaged to yield the sector 1 residual. When the value of residual error is >0.045, the VIP classifies performance as "careless."

CLASSIFICATION OF INVALID RESPONDING IN THE CROSS-VALIDATION SAMPLE

The four-fold classification scheme is inherently less risky for clinicians and less condemning of test takers. Most members of the compliant criterion groups were classified as "motivated to perform well." Compliant performance incorrectly classified as invalid will almost certainly be classified as careless and not as motivated to perform poorly. That is, most individuals who intend to respond correctly and exert consistent effort will not miss more than two of the easiest items. Their POE will be ≥ 0.8 and they will be classified as careless.

Furthermore, other potential *false-positive* classifications are mitigated by the four-fold classification scheme. Truly careless performance will result in a classification as irrelevant, but will not result in a classification of malingering. Such misclassification will accurately reflect the compromised effort that went into generating correct answers, but will incorrectly raise the hypothesis of motivation to perform poorly. Finally, false positive classification of irrelevant responding will result in categorization as malingering about 2% of the time (see Table 7). When irrelevant responding is incorrectly classified as malingered, the clinician still can confidently assume there was no intention to answer items correctly (absent historically demonstrable, bona fide, severe cognitive impairment; Frederick, 1997).

About half of the noncompliant criterion group members were classified as "motivated to perform poorly." Potential problems of *false-negative* classifications are mitigated by the four-fold classification scheme as well. Truly irrelevant performances will be incorrectly categorized as careless only about 2% of the time (see Table 7). Incorrect classifications of malingering, in which the test taker is making a concerted effort to answer incorrectly, will be classified as irrelevant, which maintains the hypothesis that the test taker did not intend to answer correctly.

The most potentially problematic misclassification is "careless" responding. Uncommon fluctuations in the performance curve in sector 3 (response accuracy well above or below that expected by chance) can potentially unduly influence the consistency measures, curvature, or interaction scores to produce an "invalid" categorization of an compliant performance. All instances of "invalid" categorization that are not evidently the result of irrelevant responding or malingering are given a default classification as "careless." Instances of VIP performance that is categorized as careless for individuals who otherwise appear to be giving good effort to respond correctly on cognitive tests can be addressed by reviewing the performance curve. In this circumstance, if the sector 1 residual is within normal limits and sector 1 is longer than sector 2, the categorization as careless should be considered weak and potentially the result of a false-positive classification.

Tables 5 and 7 indicate that some malingerers can successfully develop a strategy of correctly answering some easier items so as to generate a classification as compliant or careless, false-negative classifications. Their consistent above-chance responding within sector 1 clearly demonstrates that they intended to respond correctly to at least some items. Consequently, the categorization as motivated to respond correctly is technically correct, although clearly insufficient. When malingerers answer some items correctly, however, they run the risk of demonstrating too much ability to generate the gain that impairment would produce. Because the VIP can usually generate an estimate of cognitive capacity for careless and compliant performance curves (Crosby & Frederick, 1997; Frederick, 1997), the clinician can compare that estimate with other estimates obtained during the examination. Such a comparison may shed light on the intentions and effort of the test taker. Future research should attempt to further distinguish between careless categorizations that result from sophisticated malingering and those that result from true carelessness, including those that result from clinical conditions that inhibit concentration.

ACKNOWLEDGMENTS

Stephen Sarfaty, Jeffrey Powel, Dennis Johnston, Daniel Brockett, Beth Karassik Curry, Harold Maphet, Victoria Rivamonte, James Schraa, James Youngjohn, and the Jarrett Middle School Parent, Teacher, and Student Association of Springfield, Missouri, assisted in data collection. Kathy Gailluca, Sue Steinkamp, and Deb Ringwelski made important contributions to design and data analysis. Gary Kay, Jack Spector, and Victoria Starbuck helped develop inclusion and exclusion rules for cross-validation criterion groups. These contributions are greatly appreciated. This paper includes some data published in the Validity Indicator Profile manual, © 1997 National Computer Systems, Inc. All rights reserved. Reproduced with permission.

REFERENCES

Bigler, E. D. (1990). Neuropsychology and malingering: Comment on Faust, Hart, and Guilmette (1988). *Journal of Consulting and Clinical Psychology, 58,* 244–247.

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.

Brown, L., Sherbenou, R. J., & Johnsen, S. K. (1982). *Test of Nonverbal Intelligence: A language-free measure of cognitive ability.* Austin, TX: Pro-Ed.

Cohen, J. & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). New York: Harper Collings.

Crosby, R., & Frederick, R. I. (1997). Predicting IQ ranges from performance curve characteristics. Paper presented at the Annual Convention of the American Psychological Association, Chicago.

Faust, D. & Ackley, M. A. (1998). Did you think it was going to be easy? Some methodological suggestions for the investigation and development of malingering detection techniques. In C. R. Reynolds (Ed.) (pp. 1–54). *Detection of malingering during head injury litigation.* New York: Plenum.

Faust, D., & Guilmette, T. J. (1990). To say it's not so doesn't prove that it isn't: Research on the detection of malingering (reply to Bigler). *Journal of Consulting and Clinical Psychology, 58,* 248–250.

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,* 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,* 508–515.

Frederick, R. I. (1995, August). Taxometric analysis of malingering. Paper presented at the Annual Convention of the American Psychological Association, New York.

Frederick, R. I. (1997). *Validity Indicator Profile manual.* Minnetonka, MN: NCS Assessments.

Frederick, R. I., Carter, M., & Powel, J. (1995). Adapting symptom validity testing to evaluate suspicious complaints of amnesia in medicolegal evaluations. *Bulletin of the American Academy of Psychiatry and the Law, 23,* 231–237.

Frederick, R. I., Crosby, R., & Wynkoop, T. F. (in press). Performance curve classification of invalid responding on the Validity Indicator Profile. *Archives of Clinical Neuropsychology.*

Frederick, R. I., & Crosby, R. (1997, August). Performance curve analysis of feigned cognitive impairment and response invalidity. Paper presented at the Annual Convention of the American Psychological Association, Chicago.

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

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

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

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

Griffin, G. A. E., Normington, J., May, R., & Glassmire, D. (1996). Assessing dissimulation among Social Security disability income claimants. *Journal of Consulting and Clinical Psychology, 64,* 1425–1430.

Hays, W. L. (1973). *Statistics for the social sciences* (2nd ed.). New York: Holt, Rinehart, and Winston.

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

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.

Iverson, G. L., Franzen, M. D., & McCracken, L. M. (1994). Application of a forced-choice memory procedure designed to detect experimental malingering. *Archives of Clinical Neuropsychology, 9,* 437–450.

Meehl, P. E. (1995). Bootstraps taxometrics: Solving the classification problem in psychopathology. *American Psychologist, 50,* 266–275.

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

Rey, A. (1941). L'examen psychologie dans les cas d'encephalopathie traumatique. *Archives de Psychologie, 28,* 286–340.

Rey, A. (1958). *L'examen clinique de psychologie.* Paris: Presses Universitaires de France.

Rogers, R., Bagby, R. M., & Dickens, S. E. (1992). *Structured Interview of Reported Symptoms (SIRS) and professional manual.* Odessa, FL: Psychological Assessment Resources.

Rogers, R., Sewell, K. W., & Goldstein, A. (1994). Explanatory models of malingering: A prototypical analysis. *Law and Human Behavior, 18,* 543–552.

Rohling, M. L., Binder, L. M., & Langhinrichsen-Rohling, J. (1995). Money matters: A meta-analytic review of the association between financial compensation and the experience and treatment of chronic pain. *Health Psychology, 14,* 537–547.

Rose, F. E., Hall, S., & Szalda-Petree, A. D. (1998). A comparison of four tests of malingering and the effects of coaching. *Archives of Clinical Neuropsychology, 13,* 349–363.

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,* 472–481.

Tatsuoka, K. M., & Tatsuoka, M. M. (1982). Detection of aberrant response patterns and their effect on dimensionality. *Journal of Educational Statistics, 7,* 215–231.

Tatsuoka, K. M., & Tatsuoka, M. M. (1983). Spotting erroneous rules of operation by the Individual Consistency Index. *Journal of Educational Measurement, 20,* 221–230.

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., & Binder, L. M. (1997). Psychologists' accuracy in identifying neuropsychological test protocols of clinical malingerers. *Archives of Clinical Neuropsychology, 12,* 13–27.

Youngjohn, J. R. (1991). Malingering of neuropsychological impairment: An assessment strategy. *A Journal for the Expert Witness, the Trial Attorney, and the Trial Judge, 4,* 29–32.