Pergamon

Archives of Clinical Neuropsychology, Vol. 15, No. 4, pp. 281–300, 2000
Printed in the USA. All rights reserved
0887-6177/00 $–see front matter

PII S0887-6177(99)00002-5

Performance Curve Classification of Invalid Responding on the Validity Indicator Profile

Richard I. Frederick

U.S. Medical Center for Federal Prisoners

Ross D. Crosby

Neuropsychiatric Research Institute

Timothy F. Wynkoop

St. Francis Health Care Center

The purpose of this article is to provide evidence for the validity of performance curve classification on the nonverbal subtest of the Validity Indicator Profile (VIP-NV). A four-fold classification scheme of performance on cognitive testing is proposed. This scheme combines effort and motivation to generate four response classifications: compliant, careless, irrelevant, and malingering. Data are presented across six studies from cognitive and personality testing for 737 male pretrial criminal defendants. Additionally, computer-generated VIP-NV performances were subjected to four levels of randomization to investigate VIP-NV carelessness indicators. The findings support the validity of the four-fold classification scheme and support the classification of response on the basis of motivation and effort. Published by Elsevier Science Ltd

Keywords: malingering, response validity, performance curve analysis

Rogers, Harrell, and Liff (1993) identified six strategies that can be incorporated into neuropsychological assessments to detect feigned impairment, including *symptom validity testing*, *the floor effect*, *atypical performance*, *magnitude of error*, and *performance curve analysis.* The sixth strategy, analysis of demonstrated *psychological sequelae*, concerns evaluation with personality assessment instruments and is not discussed in this article. This article reviews how the nonverbal subtest of the Validity Indicator Profile (VIP-NV; Frederick, 1997) incorporates the first five detection strategies in categorizing the validity of cognitive assessments.

Symptom validity testing (SVT; Pankratz, 1979) involves the use of a two-alternative forced-choice test. The examinee's performance is compared to that expected by chance

The conclusions and opinions reported in this article do not necessarily represent the views of the Federal Bureau of Prisons or the U.S. Department of Justice.

Address correspondence to Richard I. Frederick, PhD, Department of Psychology, U.S. Medical Center for Federal Prisoners, Springfield, MO; E-mail: rfrederi@ipa.net

(i.e., random responding). Performances that are significantly worse than chance are considered to provide evidence of feigned or exaggerated impairment.

Floor effect involves observing performance on tasks or problems that concern overlearned material (e.g., stating one's identity or age, reciting the alphabet) or are typically easily accomplished by most individuals, including those with genuine impairment. The most commonly known floor effect test is the Rey 15-Item Memory Test (Rey, 1958), which requires the memorization and recall of easily retained information. SVT-like tests (e.g., the Portland Digit Recognition Test [PDRT; Binder, 1990]; and the Test of Memory Malingering [TOMM; Tombaugh, 1997]) are primarily floor effect tests. Although SVT-like tests use a two-alternative forced-choice format and can incorporate an SVT when performance is exceptionally poor, they typically rely upon a comparison of performance that is not worse than chance (or is better than chance) to the performance of impaired individuals.

Atypical performance concerns consistency of performance. Some approaches evaluate consistency by comparing test results with those of genuine patients or normal individuals. Mittenberg, Theroux-Fichera, Zielinski, and Heilbronner (1995), Mittenberg, Rotholc, Russell, and Heilbronner (1996), and McKinzey and Russell (1997) have proposed regression formulas to evaluate atypical performance on cognitive and neuropsychological test batteries. Another approach is to measure consistency of performance across two administrations of the same tests (Reitan & Wolfson, 1996, 1997, 1998). The VIP-NV assesses atypical performance by evaluating consistency of performance on a single administration of one test.

Magnitude of error involves a quantitative and qualitative analysis of performance failures. "Near miss" responses (e.g., 2 + 2 = 5) have historically been identified as predictive for dissimulation (Resnick, 1997). Martin, Franzen, and Orey (1996) evaluated which wrong answers were selected on a recognition memory task adapted from the Logical Memory and Visual Reproduction tests of the Wechsler Memory Scale-Revised. They determined that this approach was as effective as the 21-Item test (Iverson, Franzen, & McCracken, 1994) or the 16-Item test (Iverson & Franzen, 1996).

Performance curve analysis consists of an analysis of performance on test items across a broad range of difficulty. Essentially, the examinee's average performance on test items is compared against average item difficulty with the expectation that response accuracy will decrease as item difficulty increases. Frederick and Foster (1991) and Frederick, Sarfaty, Johnston, and Powel (1994) presented large scale studies demonstrating the effectiveness of a performance curve strategy to identify invalid responding.

DICHOTOMOUS CLASSIFICATIONS OF TEST VALIDITY

Most classification schemes for cognitive/memory malingering tests result in labels of "malingered" and "not malingered" or "compliant" and "noncompliant" (Binder, 1993; Tombaugh, 1997) This dichotomous classification scheme is problematic for at least four reasons.

First, the dichotomous classification puts the clinician in danger of making costly classification errors. False positive classification errors of "malingering" for nonmalingering examinees are pejorative and unfairly malign performances that were not intentional efforts to suppress ability. Without respect to the insulting and embarrassing nature of calling truly impaired individuals "malingerers," false-positive classification can result in undue denial of social resources, including remedial education and training, medical treatment, or disability income. In criminal proceedings, false-positive classification of "malingering" may lead to prosecution of the incompetent or more severe punishment for the convicted. False negative classifications (failing to identify malingerers as malingerers) are less costly to the examinee and are less risky for the clinician, but are costly to society. Scarce resources are inappropriately allocated from the truly deserving to the undeserving. Justice is delayed, if not prevented. Individuals who warrant prosecution may find their charges dismissed or reduced. Individuals whose crimes warrant incarceration as punishment may be hospitalized for "treatment" instead.

Second, most tests designed to detect malingering that use dichotomous classifications provide no demonstrable support for the classification beyond its atypicality. SVT initially depended upon a clear demonstration of performance worse than chance responding. More recent developments in SVT, however, have liberalized the criteria for identifying motivation to perform poorly. For example, malingering is suspected on the TOMM (Tombaugh, 1997) when the examinee correctly answers as many as 44 out of 50 items. The PDRT (Binder, 1993) has a cut score for noncompliance that is within the range of scores expected by chance. These modifications are a concession to the insensitivity of SVT. Floor effect tests, like the Rey 15-Item Memory Test, are not designed to demonstrate cooperation. When examinees earn compliant scores on floor effect tests, we cannot conclude that they displayed good effort, but only that they did not display bad effort (Faust & Ackley, 1998).

Third, dichotomous malingering classifications limit explanations of positive test scores to noncompliance or 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. A conclusion that a positive score on a dichotomous classification test was a classification error suggests that there was no point in administering the test in the first place. In other words, if one were capable of ruling out bad intentions independently of the cut score, one had no compelling reason to administer the malingering test. Consequently, positive findings on prudently administered malingering tests should typically result in a determination of noncompliance. If, however, a test could evaluate effort as well as test-taking intention, then by cross-classifications of effort with intention, one could potentially identify other factors besides intentionality that would explain poor performance. These might include fatigue, the distractibility caused by some brain injuries, the disrupted thinking processes prevalent in some mental disorders, or failure to engage in the assessment process (Frederick et al., 1994).

Fourth, the dichotomous classification scheme makes research into malingering difficult. It has proven difficult, if not nearly impossible, to develop "pure" groups of definite malingerers and fully compliant test-takers, even within analog settings (Arbisi & Ben-Porath, 1995; Frederick et al., 1994; Greiffenstein, Baker, & Gola, 1994, 1996; Greiffenstein, Gola, & Baker, 1995). Much clinical malingering research attempts to identify participants who are at risk for malingering, but assumes that control comparison groups are wholly compliant. In analog studies, token monetary awards are often offered to entice subjects to malinger, but little attention is paid to the incentives of participants who fill the role of compliers. Consequently, it is questionable whether criterion groups in much clinical research comprise fully compliant participants and bona fide malingerers. This is problematic because sensitivity and specificity are underestimated when validation is based on impure criterion groups (Dawes & Meehl, 1966).

A Four-Fold Classification Scheme

Frederick (1997) postulated that the potential validity of performance on cognitive testing is a function of two test-taking characteristics: *motivation* and *effort.* Motivation refers to the intention of the test-taker to perform well or poorly. Effort refers to the intensity of application of true ability to perform well or poorly (e.g., low effort or high effort). In this scheme, effort and motivation are independent constructs.

This cross-classification of motivation and effort results in four response styles: compliant, careless, malingered, and irrelevant (see Figure 1). *Compliant* responding is characterized by *high effort* and *motivation to perform well* (MPW; alternatively, an *intention to respond correctly*). Compliant test-takers are cooperative with testing procedures, and their performance accurately represents their ability. *Careless* test-taking is also characterized by the *motivation to perform well.* However, it differs from compliant responding in that there is *incomplete effort* to respond correctly. Careless test-taking may result from inattention, distraction, or fatigue. *Malingering* is characterized by *high effort when motivated to perform poorly* (MPP), in which the test-taker strives to feign cognitive deficits in a convincing manner. Finally, *irrelevant* responding is characterized by *token effort when motivated to perform poorly.* Irrelevant test-takers may be disengaged from the task of responding correctly, perhaps not caring about the outcome of the assessment. Random responding is included in this category.

We considered only the compliant response style to be valid. We considered careless, malingered, and irrelevant performances to be invalid response styles, given that each of these styles results in an underestimation of ability.

The Validity Indicator Profile (VIP). The VIP (Frederick, 1997) is a measure of response validity incorporating a four-fold classification scheme that is intended to be ad-

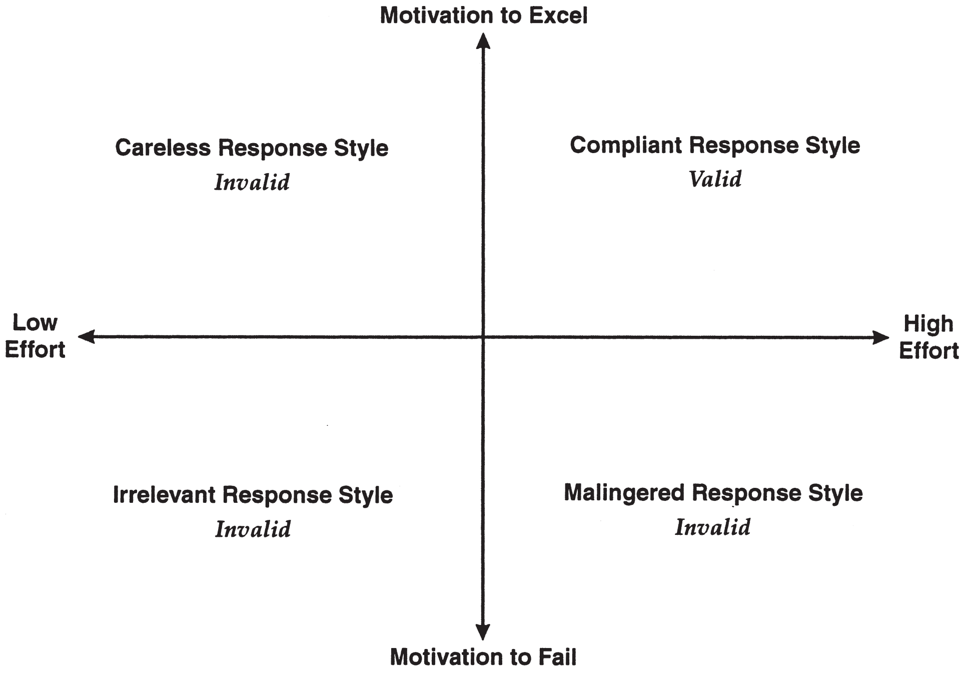

FIGURE 1. Test-taking response styles identified by a cross-classification of motivation and effort. Compliance is characterized by high effort to perform well. Carelessness is produced by low effort to perform well. Irrelevant responding represents a token effort to respond incorrectly. Malingering is characterized by a high effort to perform poorly.

ministered concurrently within a battery of cognitive tests. The VIP consists of two subtests, nonverbal and verbal. The nonverbal subtest (VIP-NV) is a 100-item picture matrix test derived from the Test of Nonverbal Intelligence (TONI; Brown, Sherbenou, & Johnsen, 1982). The TONI comprises a series of increasingly difficult matrices, which examinees solve by selecting the correct answer from among either four or six alternatives, depending on the particular item. For the VIP-NV, the TONI items were modified to have only two choices (one correct, and the other a distractor). Furthermore, presentation order was modified so that item difficulty was randomized, precluding a strategy to answer items correctly to a certain level of difficulty and to then respond randomly. The VIP verbal subtest (VIP-V) is composed of 78 two-alternative word knowledge test items. Because we had a much larger sample of the VIP-NV performances for our clinical sample than for the VIP-V, and because the two subtests share most psychometric qualities, only the VIP-NV is discussed in this article.

The VIP-NV has the same content, format, and administration as the Forced Choice Test of Nonverbal Ability (FCTNV; Frederick & Foster, 1991; Frederick et al., 1994; Rose, Hall, & Szalda-Petree, 1998). Although a few scores on the VIP-NV are computed in the same manner as on the FCTNV (e.g., total score and slope), most other scoring and cut-offs have been modified to increase diagnostic efficiency (Frederick, 1997).

VIP Performance Curve Analysis

As its fundamental analysis of approach to test-taking, the VIP generates a performance curve demonstrating the average performance of the test-taker across an increasingly difficult range of test items. Within each subtest of the VIP, all items are presented in a randomized order of difficulty. Upon completion of testing, items are reordered by difficulty and scored as “1” (correct) or “0” (incorrect). Once items are reordered to derive performance curves, a wide range of scores incorporating the five strategies of Rogers et al. (1993) are calculated (for computations, see Frederick, 1997). Performance curves are generated by plotting the average performance of the test-taker against the ordered difficulty of the test items.

Performance curves representing compliance. For high effort MPW test-takers (i.e., *compliant responding*), the two-alternative forced-choice format should result in near-perfect or perfect performance within the test-taker’s range of capacity to answer items correctly and random responding once the test-taker has reached his/her ceiling of ability. This means that performance curves for compliant test-takers should be fairly similar in shape regardless of differences in ability. The curves will start and remain at about 100% correct performance, go through a period of transition at the test-taker’s ceiling of ability, and then remain at about 50% correct performance (i.e., random responding) through the remainder of the curve (see Figure 2). Differences in cognitive capacity should result in differences in the length of the initial and ending segments of the curve, but the shape should remain similar among compliant performers.

Performance curves that are unexpected for compliant responding. Significant deviations from this expected curve have meaning and allow for some reasonable conclusions about the response style of the individual. For example, an individual who performs at 80% throughout most of the test (see Figure 3, Line A) may intend to respond correctly but is probably expending insufficient effort (*careless responding*). It is likely that the test-taker could have performed perfectly on much easier items, given that he/she correctly solved 80% of the moderately difficult items. Another deviation from the ex-

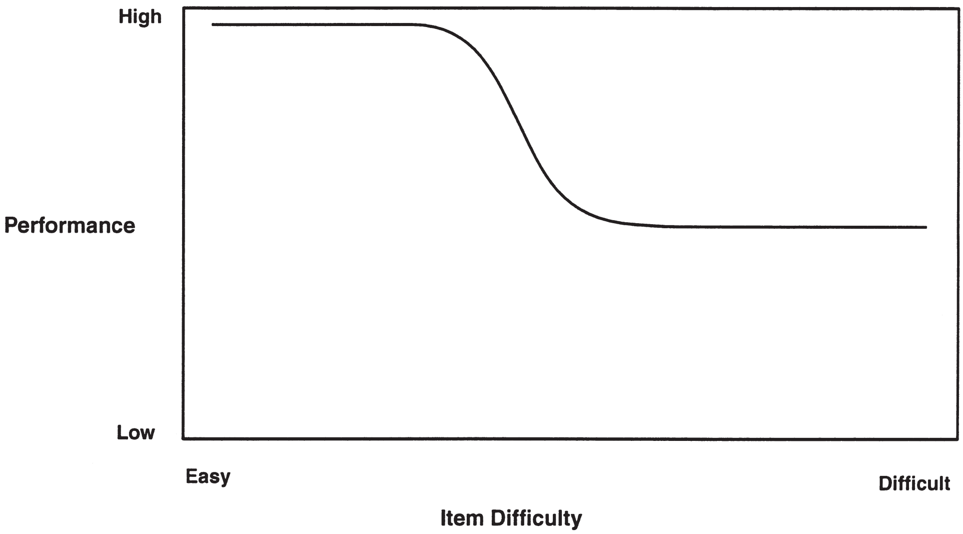

FIGURE 2. Expected performance curve for compliant performance on a two-alternative forced-choice test comprising a hierarchy of difficulty. The Point of Entry reveals perfect performance for the 10 easiest items. Perfect or near-perfect performance for more difficult items is demonstrated by the continuing straight line after the Point of Entry. As item difficulty precludes correct responding (reaching one's ceiling of ability), performance accuracy drops off, but will remain at about 50%, because guessing on two-alternative items will result in about 50% correct responses.

pected curve might be for an individual who performs at about 50% throughout the entire test (Figure 3, Line B). Such an individual is probably marking answers without regard to item content (*irrelevant responding*). As a final example, consider an individual who demonstrates a consistent increase in correct responding as the test items become more difficult (Figure 3, Line C). Such an individual is most likely choosing the wrong answer intentionally, despite knowing the correct solution (*malingering*), and responding randomly only when the correct answer is not known.

Components of Performance Curve Analysis

Running mean. The points used to plot an individual's performance curve are derived by a process of shifting averages and are referred to as running means. For the VIP-NV, there are 91 running means derived from the 100 test items. Each running mean represents the average test performance for consecutive runs of 10 test items when the test items are ordered by difficulty. The first running mean is the proportion of correct answers (represented by 1) earned for the 10 easiest items. The second running mean is derived by dropping the score for the easiest item and adding the score for the next most difficult item. This process continues until the most difficult item is included in computation of the last running mean. To derive performance curves, running means are plotted against the serial position of item difficulty (comprising the numbers 1 through 91). Once the running means are plotted, performance curve indicators are computed.

Performance curve analysis: Slope. The essential element of performance curve analysis is a comparison of average performance on easy items as compared to difficult items. This is accomplished by analysis of the *slope* of the regression line that best fits the plot

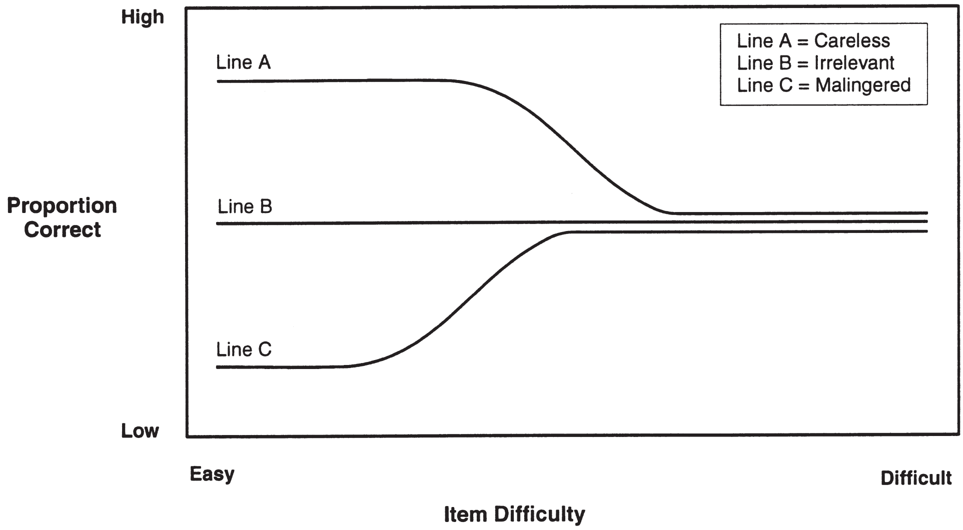

FIGURE 3. Unexpected performance curves for compliant performance on a two-alternative forced-choice test comprising a hierarchy of difficulty. Line A represents careless responding. The individual's imperfect responding at the Point of Entry and just beyond is made suspect by equivalent performance on more difficulty items. Line B represents irrelevant responding. Consistent performance accuracy at 50% indicates that responses were irrelevant to item content. Line C represents malingering. The improved performance as items become more difficult is best explained by intentional incorrect responding within the range of one's capacity to answer correctly and *improved* performance to 50% when one is forced to guess (once one reaches the ceiling of one's ability).

of running means. 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. Frederick and Foster (1991) and Frederick et al. (1994) demonstrated that an increase in performance accuracy in the face of increasing item difficulty (i.e., a positive performance curve slope) indicates motivation to perform poorly.

SVT and floor effect: Point of entry. The first running mean is the beginning point of the curve, and is referred to as 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.

SVT. 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 correctly determine the correct answers to the 10 easiest items because of the low cumulative probability (p = .055) of earning 8 (p = .044), 9 (p = .010), or 10 (p = .001) correct answers out of 10 items by chance (based on binomial expansion; Hays, 1973). Likewise, POEs ≤ 0.2 are equally improbable and most 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 to 0.7 are consistent with random responding.

Floor effect. The 10 items comprising the POE are extremely easy and should be correctly answered in the absence of severe impairment. Sixty-five percent of 40 individuals with mild to moderate mental retardation earned POEs ≥0.8 on the VIP-NV (Frederick, 1997). Scores below 0.8 are not expected when even minimal reasoning capacity is present, and thus represent a floor of performance for almost all individuals.

Atypical performance: Performance curve sector analysis. Inconsistent performance is evaluated by comparing Sector 1 and Sector 2 distance.

Sector distances. Performance curves that have POEs >0.7 are divided into three sectors. 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 "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 to 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, 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.

Magnitude of error: 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, since 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 reaches a sufficient magnitude (calculated at >.045 by the VIP-NV validation sample; Frederick, 1997), the VIP classifies performance as careless.

Expected performance curves. Some discussion in this article involves an analysis of *expected performance curves.* Expected performance curves are the curves that would result if the examinee responded in a perfectly reliable manner and the test items were themselves perfectly reliable. Expected performance curves are derived from the examinee's *adjusted score*, which is an estimate of the actual number of items for which the examinee can derive the answer based on knowledge alone.

Adjusted score. The number of correct answers on the VIP constitutes the total score. The total score does not typically reflect the total number of items for which the test-taker has knowledge. Instead, the total score is the sum of the number of correctly answered items based on knowledge added to the number of correctly chosen items based on guessing. Total score can be expressed as follows:

Total score = knowledge + correct guessing (chance)

Guessing on items that are beyond the ability of the individual test-taker poses a particular problem in the calculation of a total score on two-alternative forced-choice tests. This is because about 50% of the time the test-taker will guess right, which artificially inflates the total score. Cronbach (1990, p. 67) provided a solution to the artificial inflation of test scores by random chance (when two alternatives are available) by subtracting errors from right answers:

Adjusted score = total score – incorrect responses

For example, a total score of 75 (out of 100 for the VIP-NV) results in the following equation:

Adjusted score = 75 – 25 = 50

The adjusted score is the best estimate of the number of items for which the examinee knew the correct answer. *Total scores* for random responding should have a mean of 50. The average *adjusted score* for random responding is 0 (50 − 50), indicating no knowledge of correct answers. Total score deviations below 50 represent as much capacity to answer correctly as equal magnitude deviations above 50. Consequently, the adjusted score is calculated as the absolute value of the difference between the number of correct and incorrect responses. Figure 4 represents the relationship between total score and adjusted score.

Generating the expected performance curve. Expected performance curves are generated from the adjusted score. To generate an expected curve from an adjusted score of 50, we assume that the 50 *easiest* items could be answered correctly by knowledge and the examinee guessed answers on the remaining more difficult items. Furthermore, we assume that the pattern of guessing on the remaining 50 difficult items can best be represented by alternating correct and incorrect responses. Thus, the expected curve for an adjusted score of 50 is generated from running means computed on a series of 50 ones followed by an alternating pattern of ones and zeros.

Construct Validation of Categories Established by Detection Strategies

The purpose of this article is to evaluate the construct validity of the categorizations that result from VIP performance curve decision rules over six studies. Comparison groups were formed by VIP performance curve characteristics and the effect of group membership was considered for performance on other measures. Within several studies, we examined the effect of group membership on performance on each of three malingering tests devised by Rey (1941, 1958). We expected to see meaningful effects for comparisons of groups that presumably differed in test-taking intent (MPP vs. MPW), but not for groups that presumably differed only on levels of test-taking effort.

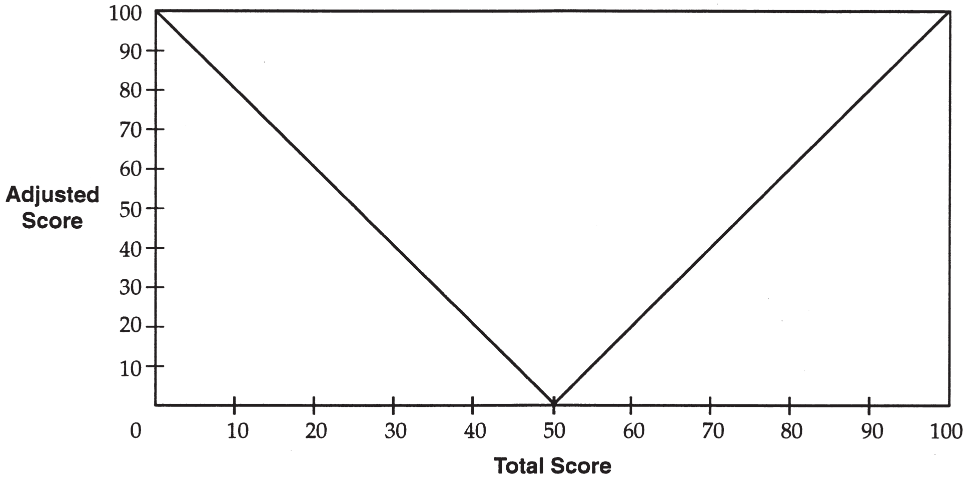

FIGURE 4. Comparison of total score and adjusted score for the Validity Indicator Profile nonverbal subtest. At a total score of 50 for a 100-item two-alternative forced-choice test, no ability is demonstrated; the adjusted score is 0. As performance deviates from a total score of 50, in either direction, there exists better evidence of a capacity to answer correctly. Consequently, total scores of 30 and 70 both reflect that the test-taker probably knew the answer to about 40 items. The difference in total score reflects a difference in test-taking intention. The total score of 30 represents motivation to perform poorly; the total score of 70 represents motivation to perform well.

First, we compared a group whose members generated negatively-sloped performance curves (MPW) with a group with positively sloped performance curves (MPP). Second, we compared a group classified as irrelevant responders (low-effort MPP) with a group classified as careless responders (low-effort MPW). Third, we compared a group classified as careless responders (low-effort MPW) with a group classified as compliant responders (high-effort MPW). Fourth, we compared careless and compliant participants on Minnesota Multiphasic Personality Inventory-2 (MMPI-2) indicators of carelessness. Fifth, we evaluated the VIP construct of carelessness by generating a large number of expected curves by computer at various adjusted scores to evaluate the effect of randomization of response on VIP carelessness indicators. Finally, we evaluated the effect of psychotic illness and affective disorders on VIP carelessness indicators.

METHOD

Participants

Criminal defendants. Data were obtained from evaluations of 737 male pretrial defendants referred to the mental health evaluation service of the U.S. Medical Center for Federal Prisoners (mean age = 36.6, *SD* = 10.9; mean years of education = 10.7, *SD* = 3.3). Examinees were referred consecutively for testing unless they refused or were too impaired for testing. Individuals examined included 414 White (56.2%), 199 Black (27.0%), 84 Hispanic (11.4%), 26 Native American (3.5%), 7 Asian (0.9%), and 7 (0.9%) from other categories. Reasons for referral often overlapped, but included 522 competency-to-stand-trial examinations, 325 criminal responsibility examinations, 13 risk assessments, 64 general psychological evaluations, and 134 commitments for restoration to competency to stand trial. These were real-world participants with meaningful potential long-term gains for believable impairment. Examinees completed the VIP-NV and the Rey malingering tests. Additionally, many (n = 527) completed the MMPI-2 (Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989).

Classifications of performance on the VIP-NV resulted in 415 classifications as compliant (56.3%), 223 as careless (30.3%), 91 as irrelevant (12.3%), and 8 as malingering (1.1%). This does not mean that only 1.1% of our sample was malingering. Classification of malingering on the VIP is intentionally conservative and is designed to categorize as "malingering" only those performances that have no competing explanations for their genesis.

Instruments

Participants were routinely administered the Rey malingering tests (Rey, 1941, 1958). These tests are easily and quickly administered, have no cost, and were available throughout the time period of clinical testing (1993 through 1998). These tests are generally considered less than ideal for establishing criterion groups for test validation because of their reportedly poor diagnostic efficiencies. Nevertheless, the tests have consistently yielded relatively large effect sizes (with noted exceptions) when subjected to malingering. For our first three studies, we chose to establish criterion groups based on VIP-NV classifications of compliant, careless, irrelevant, MPP, or MPW and evaluated the effect of those classifications on performance on the Rey malingering tests.

Rey 15-Item Memory Test. The Rey 15-Item Memory Test (RMT; Lezak, 1995; Rey, 1958) consists of a card with five rows of three items that appear in a familiar logical sequence (e.g., 1,2,3 or circle, square, triangle). Examinees were told to remember all 15 items during a 10-second exposure. After the stimulus items were removed, and a 10-second delay was interpolated, examinees were told to reproduce the items in the correct order on a blank sheet of paper. The score consistently reported in the literature is the number of correctly recalled items. Scores range from 0 to 15; lower scores are consistent with an intention to perform poorly. The lowest effect size[1] for malingering was reported by Arnett, Hammeke, and Schwartz (1995) at 0.4. An extremely large effect size (2.0) was reported by DiCarlo, Gfeller, and Drury (1996). Most other researchers have consistently found large to very large effect sizes for malingering on the RMT total score: 0.7 (Bernard & Fowler, 1990), 0.8 (Griffin, Normington, & Glassmire, 1996; Greiffenstein et al., 1995; Greiffenstein, Baker, & Gola, 1996; Guilmette, Hart, Giuliano, & Leininger, 1994), and 0.9 to 1.2 (Greiffenstein et al., 1994, 1995).

Word Recognition Test. The Word Recognition Test (WRT; Lezak, 1995; Rey, 1941) is composed of two word lists, one of 15 words (stimulus list) and the other of 30 words (memory test). The memory test contains the 15 stimulus words and 15 distractors. In our administration, we read the stimulus list the examinee. We then read the memory test 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 subtracting the number of misrecognized words from the number of correctly recognized words. Lower scores indicate motivation to respond incorrectly. Greiffenstein et al. (1995, 1996) reported that malingering produced an effect size of 1.1 to 1.7 on this score.

[1] By "effect size," we mean Cohen's d statistic, the standardized mean difference between two groups of interest (Cohen, 1988).

Dot Counting Test. The Dot Counting Test (DCT; Rey, 1941) consists of twelve 3 × 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. We followed the scoring alternative described by Binks, Gouvier, and Waters (1997), who found the most discriminating score was the number of cards for which an incorrect sum was reported. Their comparisons generated effect sizes of about 2.2 to 2.3 for this score. Rose et al. (1998) found an effect size of about 1.5 for number of cards for which a correct sum was reported. The number of occasions of miscounts range from 0 to 12; higher scores indicate a motivation to perform poorly.

COMPARATIVE ANALYSES OF MPP AND MPW CLASSIFICATIONS ON THE VIP

Slope as an Indicator of Motivation

Procedure. Forty-five persons (6.1%) from our sample of criminal defendants generated performance curves with positive slopes ($M_{\text{pos slope}} = .0027$, $SD = .002$), indicating an intention to perform poorly. Their performance curves were individually matched with the performance curves of 45 other defendants on the basis of equivalent magnitude of slope, but with a negative valence, indicating an intention to do well on the test (compliant; $M_{\text{neg slope}} = -.0027$, $SD = .002$). The two groups were compared on mean score differences for the Rey malingering tests. Correlations between performance curve slope and Rey malingering test scores were computed. RMT and WRT scores were expected to result in substantial negative correlations with slope values (lower RMT and WRT scores reflect motivation to perform poorly). DCT scores were expected to result in substantial positive correlations with slope values (higher DCT scores reflect motivation to perform poorly).

TABLE 1
Rey Malingering Test Performance as a Function of Validity Indicator Profile (VIP) Categorization

	VIP Adjusted Score		Rey 15-Item Memory Test					Word Recognition Test					Dot Counting Test				
	M	*SD*	*n*	*M*	*SD*	*d*	*t*	*n*	*M*	*SD*	*d*	*t*	*n*	*M*	*SD*	*d*	*t*
Positive slope	10.7	13.0	45	7.8	3.6			44	3.2	5.4			44	5.9	3.3		
Negative slope	43.3	33.3	45	12.4	3.2	1.36	−6.4*	44	7.8	3.8	1.02	−4.7*	44	2.9	2.6	1.04	4.9*
Irrelevant	11.8	10.6	74	8.7	3.3			73	4.6	4.8			74	4.9	3.4		
Careless	11.8	10.6	72	11.2	3.2	.77	−3.5*	73	7.1	3.7	.58	−3.5*	74	3.0	2.1	.68	4.1*
Careless	31.4	13.5	137	12.1	2.8			137	8.1	3.2			136	2.4	1.8		
Compliant	31.4	13.5	136	12.3	2.5	.08	−.8	136	8.4	3.4	.08	−.7	137	2.5	1.7	.06	−.4

Note. Participants were criminal defendants. Groups comprising positive and negative slope comparisons were matched on magnitude of Validity Indicator Profile (Nonverbal) slope values; valences are opposite ($n = 45$ each group). Irrelevant vs. Careless ($n = 75$ each group) and Careless vs. complaint ($n = 138$ each group) comparison groups were matched on Validity Indicator Profile adjusted score. Rey 15-Item Memory Test scores represent the number of items correctly reproduced. Scores for the Word Recognition Test scores were computed by subtracting the number of words incorrectly "recognized" words (false positives) from the number of correctly recognized words (true positives). The score for the Dot Counting Test is the number of trials (out of 12) that an incorrect total was generated.
*t-scores significant at $p < .05$.

Results. Large to very large effect sizes (1.0–1.4) were observed for comparisons of slope valence for each Rey malingering test (see Table 1, section 1). Correlations between performance curve slope and Rey malingering test scores were substantial and in the expected directions ($r_{RMT} = -.59, n = 90, p < .01; r_{WRT} = -.51, n = 89, p < .01; r_{DCT} = .48, n = 88, p < .01$).

Discussion. The effect sizes are consistent with the previously cited research showing large to very large effect sizes for individuals motivated to perform poorly. These results support the hypothesis that valence of performance curve slope on the VIP-NV is related to the intention to perform well or to perform poorly. As performance curve slopes increased in positive magnitude, scores on the Rey malingering tests indicated a greater likelihood of intention to do poorly. Not all participants in our sample who were classified as motivated to perform poorly manifested such clear evidence of intention to do poorly as those with positively sloped performance curves. Therefore, we also wanted to classification of MPP as generated by POE.

POE as an Indicator of Motivation

Procedure. We compared performances on the Rey malingering tests of a group of persons whose VIP-NV performance was classified as irrelevant (*low effort* to perform poorly; POE $< .8$; $n = 75$) with an equally sized group of persons whose VIP-NV performance was classified as careless (*low effort* to perform well; POE $> .8$). We wanted comparisons of performance to be independent of cognitive capacity. Consequently, we matched participants on their VIP-NV adjusted scores, which is the best estimate of how many items were actually known by the examinee. Irrelevant responders, because of their presumed motivation to perform poorly, were expected to have lower mean Rey test scores than those classified as careless. But, we suspected that low effort (which is presumed for both groups) would result in smaller differences between mean Rey test scores than that observed previously.

Results. Because examinees were matched on VIP-NV adjusted scores, no difference existed between mean VIP-NV adjusted scores (for each group, $M = 11.8$, $SD = 10.6$, $n = 75$). Differences in mean Rey test scores were in the expected directions (see Table 1, section 2). Effect sizes observed for irrelevant-careless comparisons were smaller than those observed for differences based on valence of performance curve slope, but remained sizable, in the medium to large range.

Discussion. These first two studies have demonstrated that classification on the VIP-NV as MPP results in significantly lower mean Rey test scores than for individuals classified as MPW. In the following analysis, we compared careless responders (low effort MPW) with compliant responders (high effort MPW).

Comparison of Careless and Compliant Responders on the Rey Malingering Tests

Procedure. We compared two groups of careless ($n = 138$) and compliant ($n = 138$) examinees to determine if they differed in test-taking intention. Participants were classified as careless if their Sector 2 distance was equal or greater than their Sector 1 distance, or if their Sector 1 residual was greater than .045. Groups were again matched on VIP-NV adjusted score to preclude differences based on ability to respond correctly.

Because both groups were believed to have intended to respond correctly, despite differences in effort, no significant differences were expected in their performances on the Rey malingering tests.

Results. Mean Rey test scores are presented in Table 1, section 3. There were no meaningful differences between the groups on measures of motivation to perform poorly.

Discussion. The lack of meaningful differences between the compliant and careless groups on the Rey malingering tests supports our hypothesis that these groups do not differ on the construct of motivation (MPP vs. MPW). These results do not speak to the effort expended by these two groups to perform well, which we hypothesize to be higher in a compliant group. The next analysis evaluated the validity of our conceptualization of effort by examining the effect of carelessness on responding on the MMPI-2.

Careless and Compliant Responders and the MMPI-2

Procedure. If the construct of effort is valid, then performance on other concurrently administered measures of effort and consistency should be different between a group of careless responders and compliant responders. As a self-report two-alternative test, the MMPI-2 should be subject to the same influences that could affect performance on the VIP-NV. Numerous indicators have been developed to evaluate careless and inconsistent responding on the MMPI-2 (Greene, 1991). Cramer (1995) evaluated such indices on different levels of randomized MMPI-2 performances (F, Fb, VRIN, |F − Fb|, VRIN + |F − Fb|, and F + Fb + |F − Fb|). Of the six indices evaluated by Cramer, only |F − Fb|; failed to demonstrate utility in differentiating authentic from random profiles. In this analysis, we compared the same two groups of careless and compliant test-takers used in the last analysis on these MMPI-2 measures of inconsistency.

Participants. Two hundred defendants were selected for this analysis. They had completed the MMPI-2 and VIP, were classified as compliant or careless responders (n = 100 for each group), and were matched on VIP adjusted score.

TABLE 2
Comparison of Minnesota Multiphasic Personality Inventory - 2 Carelessness Indicators Based on Validity Indicator Profile (VIP) Categorization

	Compliant		Careless			
	M	*SD*	*M*	*SD*	*d*	*t*
F	14.2	9.4	18.2	9.3	−.43	3.01*
Fb	10.1	8.0	13.3	8.1	−.39	2.76*
VRIN	6.7	3.2	7.9	3.8	−.34	2.41*
|F − Fb|	6.1	4.4	5.6	4.1	.14	1.01
F + Fb + |F − Fb|	29.8	18.7	37.6	18.3	−.42	2.95*
VRIN + |F − Fb|	12.2	5.8	14.0	6.3	−.30	2.12*
VIP Adjusted Score	32.6	13.1	32.6	13.1		

Note. Particiipants were criminal defendants. $N = 200$; $n = 100$ for each group. Groups were matched on VIP adjusted score.
d = Cohen's d.
*$p < .05$.

Results and discussion. The small to medium effect sizes (Table 2) for all indices except |F − Fb|; were statistically significant. The small effect size for |F − Fb|; (.14; with greater mean scores for compliant responders) is in keeping with the findings of Cramer (1995) that partial randomization of MMPI-2 responses had little effect on this indicator of response consistency. The effect sizes (−.30–−.43) for the other MMPI-2 indices of consistency provide support for the conceptualization of effort which underlies the VIP-NV distinction between "compliant" and "careless."

Computer Simulation of Careless Responding on Expected Curves

In our next analysis, we evaluated the VIP categorization of carelessness by simulating carelessness through computer simulation. We construed true carelessness to be a limited randomizing effect on otherwise compliant performance. By "limited randomizing effect," we mean there is no reason to expect carelessness to automatically result in wrong answers, but rather will produce a subset of responses that are irrelevant to item content. Responses that are potentially relevant become irrelevant through inattention, lack of concentration, or marking errors. On the average, about half of those irrelevant responses should be wrong responses. We planned our analysis to mimic this hypothesized limited randomizing effect.

We examined the effect of different levels of carelessness on expected curves. As noted earlier, expected curves are the curves that would be expected for perfectly consistent responding were the VIP a perfectly reliable test. Expected curves can be generated for any value of adjusted score, but larger adjusted scores result in larger Sector 1 distances and shorter Sector 3 distances. Expected curves have predictable features. For expected curves generated from adjusted scores of 10 or more on the VIP, the POE is always 1.0, the Sector 1 (ability sector) distance is always equal to the adjusted score minus 5, and the Sector 2 (transition sector) distance is always 4. Just as Sector 1 distance depends on the extent of true ability, the Sector 1 residual for expected curves is dependent on the length of Sector 1. For expected curves, only a few points at the end of Sector 1 (where ability transitions to guessing) contribute to the Sector 1 residual because all other points are 1.0. As the adjusted score increases for expected curves, the Sector 1 distance increases. Consequently, the Sector 1 residual decreases as those few points are a decreasingly small proportion of Sector 1. For example, the Sector 1 residual for an expected curve corresponding to an adjusted score of 10 is .010, but the Sector 1 residual for an expected curve for an adjusted score of 90 is .001.

Hypotheses. When examinees answer carelessly, some items they could answer correctly are answered incorrectly. Several consequences are predicted when a compliant performance is contaminated by carelessness in this way. First, measured ability (i.e., Sector 1 distance) should decrease. Second, Sector 1 should show elements of inconsistency with additional instances of running means less than 1.0. Inconsistency in Sector 1 should result in an increase in Sector 1 residual or premature initiation of a Sector 2. Third, Sector 2 distance should increase because of premature termination of Sector 1 or because of increased variability within Sector 2 (which could postpone the termination of Sector 2). Consequently, the fourth result should be an increase in the difference between Sector 2 distance and Sector 1 distance, resulting in more classifications of carelessness as Sector 2 exceeds Sector 1.

Procedure. We generated (by computer) 4,000 expected curves for adjusted scores from 10 to 90 in increments of 5 points (i.e., 10, 15, 20, etc). To simulate careless responding, we subjected the 4000 curves at each adjusted score to randomization by replacing 5, 10, 15, or 20 item responses (1000 curves at each level of randomization). Because carelessness should not always result in an incorrect item response (i.e., a "0"), we did not necessarily change each of the selected responses; rather, we replaced each one with a randomly selected 0 or 1. After the changes in item responses were completed, curves were regenerated from the new item responses. We computed Sector 1 distance, Sector 2 distance, their difference, and Sector 1 residual.

Results. Figure 5 demonstrates the effect of randomization on mean Sector 1 distances, mean Sector 2 distances, mean Sector 1 residuals, and the proportion of Sector 1 residuals which exceeded the VIP-NV cut-off of .045. at each total score level and each level of randomization.

Discussion. Sector 1 distances decreased and Sector 2 distances increased as randomization increased. The proportion of occasions on which Sector 2 distance exceeded Sector 1 distance increased. Sector 1 residuals increased as randomization increased. These changes were in the predicted directions. Consequently, when these features occur in a performance curve, it is reasonable to conclude that careless responding is a potential contributing factor.

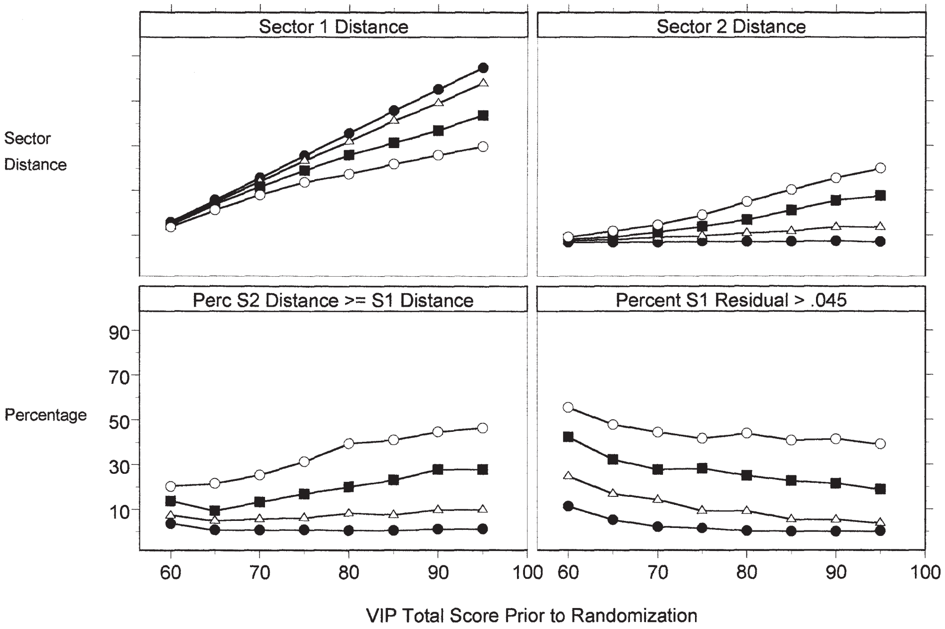

FIGURE 5. Effect of limited randomization on performance curve sector characteristics. Expected curves representing various total scores in increments of 5 points from a total score of 60 to a total score of 95 were subjected to limited randomization. Five, 10, 15, or 20% of test items were replaced with randomly selected right or wrong responses. ● = 5% randomization of item responses; △ = 10% randomization; ■ = 15% randomization; ○ = 20% randomization. As randomization increases and as initial total scores increase, performance curve measurements associated with careless responding increase proportionally.

Effects of Psychosis and Affective Disorders on VIP Performance

In conclusion, we examined clinical conditions that may induce careless responding. Psychosis and mood disorders are noted for their disruption in thought processes. Psychotic disorders are often characterized by perceptual distortions and disturbances in thinking. Affective disorders often are characterized by elements of distractibility, malaise, and inattention. The purpose of the next analysis is to consider the effect of these clinical conditions on VIP responding.

Procedure. Diagnoses from forensic reports were available for 523 of the criminal defendants. Three pools of participants were identified based on diagnostic categories. The first category was for individuals for whom no diagnosis was assigned (n = 108). The second and third categories comprised individuals who were assigned a diagnosis of psychotic disorder only (n = 84) or a diagnosis of affective disorder only (n = 44). These diagnoses were generated by the primary clinician and were not validated by an objective diagnostic tool. From within the pools of persons with no diagnosis and persons diagnosed as psychotic, 120 individuals (n = 60 for each group) were matched on adjusted score. Participants diagnosed with a psychotic disorder included 42 diagnosed with schizophrenia, 4 schizoaffective disorder, 6 delusional disorder, 6 psychotic disorder not otherwise specified, and 2 reactive psychosis. From within the pools of persons with no diagnosis and persons diagnosed with an affective disorder, 66 individuals (n = 33 for each group) were matched on adjusted score. Participants diagnosed with an affective disorder included 10 diagnosed with bipolar disorder, 8 major depression, 2 depressive disorder not otherwise specified, 8 posttraumatic stress disorder, and 5 adjustment disorder. Rates of VIP classification were summarized and characteristics of Sectors 1 and 2 were evaluated.

Results. Table 3 shows the rates of VIP-NV classifications for the psychosis and affective disorder groups as compared to matching groups of individuals with no diagnosed mental disorder. The rates of VIP-NV classification were not different between psychosis/no diagnosis groups ($\chi^2 = 2.09$, $df = 3$, $N = 120$, $p > .05$) or between affective disorder/no diagnosis groups ($\chi^2 = 2.07$, $df = 2$, $N = 66$, $p > .05$). Because not all individuals in the control or diagnostic groups produced careless or compliant responding, the number of individuals was reduced for groups comparisons of Sector 1 and Sector 2 characteristics (see Table 4). No significant differences existed between comparison groups in Sector 1 distance, Sector 2 distance, Sector 2 distance minus Sector 1 distance, or Sector 1 residual.

TABLE 3
Rates of Validity Indicator Profile Categorization Based on Assigned Diagnosis

	Validity Indicator Profile Categorization			
Assigned Diagnosis	Malingering	Irrelevant	Careless	Compliant
Psychotic disorder	1	8	18	33
No diagnosis	1	5	25	29
Affective disorder	0	4	7	22
No diagnosis	0	1	9	23

Note. In both comparisons, groups were matched on Validity Indicator Profile adjusted score; for comparisons with psychotic disorder, M = 38.0, SD = 2.17, n = 120. For comparisons with affective disorder, M = 46.6, SD = 23.0, n = 66.

Table 4
Validity Indicator Profile Careless Indicators Based on Assigned Diagnosis

Assigned		S1 Distance				S2 Distance				S1 Distance– S2 Distance				S1 Residual			
Diagnosis	*n*	*M*	*SD*	*d*	*t*	*M*	*SD*	*d*	*t*	*M*	*SD*	*d*	*t*	*M*	*SD*	*d*	*t*
Psychotic disorder	54	38.9	20.9			19.9	13.7			−20.5	28.6			.0035	.0027		
No diagnosis	51	35.1	22.9	−.17	.89	18.4	14.5	.11	−.54	−15.2	30.9	.18	−.91	.0035	.0029	−.01	0.00
Affective disorder	32	44.1	20.8			19.6	15.2			−31.0	28.8			.0029	.0024		
No diagnosis	29	49.0	21.4	−.24	.91	18.1	11.9	.11	−.43	−24.5	31.7	.21	−.83	.0027	.0022	.07	−0.29

Note. In both comparisons, groups were matched on Validity Indicator Profile adjusted score; for comparisons with psychotic disorder, $M = 38.0$, $SD = 21.7$, $n = 120$. For comparisons with affective disorder, $M = 46.6$, $SD = 23.0$, $n = 66$.
S1 = Sector 1; S2 = Sector 2; d = Cohen's d.

Discussion. It is noteworthy that one half of individuals with psychotic disorders and one third of individuals with affective disorders generated performances that were classified as invalid by the VIP. But these rates were not different than those for individuals who were not diagnosed with any disorder and were also pretrial defendants undergoing forensic examination. This analysis does not identify the reason for rates of invalidity for any of the groups. However, these results do suggest that one should not expect a differential rate of VIP categorization as careless primarily based on manifestation of a psychotic disorder or affective disorder. This is a surprising finding, and may reflect the lack of rigor involved in forming comparison groups only from a single clinical judge. More rigorous studies may shed light on which clinical conditions (e.g., attention deficit disorder) routinely result in careless responding.

CONCLUSION

In summary, classifications of response style by VIP performance curve characteristics were supported by an analysis of concurrently administered malingering tests. Large to very large effect sizes on Rey test performance were seen for differences in motivation to perform well versus motivation to perform poorly as measured by the VIP-NV performance curve slope. Moderate to large effect sizes on Rey test performance were seen for differences in motivation even when the effort of subjects was presumed to be low. There were zero-order effects on Rey test performance when compliant test-takers were compared with those classified as careless. But, examinees classified as careless had significantly higher scores on MMPI-2 carelessness indicators when compared to examinees classified as compliant. We demonstrated a predictable effect of randomizing a subset of responses in an attempt to understand the effect of careless responding. These results support the formulations of motivation and effort as different constructs, with meaningful categorizations of performance resulting from the cross-classification of effort and motivation.

REFERENCES

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

Arnett, P. A., Hammeke, T. A., & Schwartz, L. (1995). Quantitative and qualitative performance on Rey's 15-Item Test in neurological patients and dissimulators. *The Clinical Neuropsychologist*, **9**, 17–26.

Bernard, L. C., & Fowler, W. (1990). Assessing the validity of memory complaints: Performance of brain-damaged and normal individuals on Rey's task to detect malingering. *Journal of Clinical Psychology*, **46**, 432–436.

Binder, L. M. (1990). Malingering following minor head trauma. *The Clinical Neuropsychologist*, **4**, 25–36.

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.

Binks, P. G., Gouvier, W. D., & Waters, W. F. (1997). Malingering detection with the Dot Counting Test. *Archives of Clinical Neuropsychology*, **12**, 41–46.

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

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

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

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

Cramer, K. M. (1995). Comparing three new MMPI-2 randomness indices in a novel procedure for random profile derivation. *Journal of Personality Assessment*, **65**, 514–520.

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

Dawes, R. M., & Meehl, P. E. (1966). Mixed group validation: A method for determining the validity of diagnostic signs without using criterion groups. *Psychological Bulletin*, **66**, 63–67.

DiCarlo, M. A., Gfeller, J. D., & Drury, J. A. (1996). *Assessing feigned cognitive impairment with Rey's 15-Item Memory Test: Is this task easy or what?* Poster presented at the Annual Meeting of the National Academy of Neuropsychology, New Orleans.

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.). *Detection of malingering during head injury litigation* (pp. 1–54). New York: Plenum Press.

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

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.

Greene, R. L. (1991). *The MMPI-2/MMPI: An interpretive manual.* Boston: Allyn & Bacon.

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

Greiffenstein, M. F., Baker, W. J., & Gola, T. (1996). Comparison of multiple scoring methods for Rey's malingered amnesia measures. *Archives of Clinical Neuropsychology*, **11**, 283–293.

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

Griffin, G. A., Normington, J., & Glassmire, D. (1996). Qualitative dimensions in scoring the Rey Visual Memory Test of malingering. *Psychological Assessment*, **8**, 383–387.

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.

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

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

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.

Lezak, M. D. (1995). Neuropsychological assessment (3rd ed.). New York: Oxford.

Martin, R. C., Franzen, M. D., & Orey, S. (1996). *Comparative utility of the magnitude of error strategy in identifying feigned brain injury.* Poster presented at the 16th Annual meeting of the National Academy of Neuropsychology, New Orleans.

McKinzey, R. K., & Russell, E. W. (1997). A partial cross-validation of a Halstead-Reitan battery malingering formula. *Journal of Clinical and Experimental Neuropsychology*, **19**, 484–488.

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. L. (1995). Identification of malingered head injury on the Wechsler Adult Intelligence Scale-Revised. *Professional Psychology*, **26**, 491–498.

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.

Reitan, R., & Wolfson, D. (1996). The question of validity of neuropsychological test scoes among head-injured litigants: Development of a Dissimulation Index. *Archives of Clinical Neuropsychology*, **11**, 573–580.

Reitan, R., & Wolfson, D. (1997). Consistency of neuropsychological test scores of head-injured subjects involved in litigation compared with head-injured subjects not involved in litigation: Development of the Retest Consistency Index. *The Clinical Psychologist*, **11**, 69–76.

Reitan, R., & Wolfson, D. (1998). Detection of malingeringand invalid test results using the Halstead-Reitan Battery. In C. R. Reynolds (Ed.), *Detection of malingering during head injury litigation* (pp. 163–208). New York: Plenum Press.

Resnick, P. J. (1997). Malingered psychosis. In R. Rogers (Ed.), *Clinical Assessment of malingering and deception* (2nd ed.). New York: Guilford Press.

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

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.

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