

---

---

## Do Tests of Malingering Concur? Concordance Among Malingering Measures

Melanie R. Farkas, M.A.,<sup>†</sup> Barry Rosenfeld, Ph.D.,<sup>†\*</sup> Reuben Robbins, M.A.<sup>†</sup> and Wilfred van Gorp, Ph.D.<sup>‡</sup>

---

**Malingering test accuracy is increasingly a major issue in psychology and law. Integrating results across measures might offset limitations of a single test, but the practical benefits of using several tests depend on the extent to which they misclassify the same individuals. Data from 66 evaluatees were used to assess the degree of overlap and consistency of classification among several commonly used malingering instruments. Although correlative data indicated that measures were highly redundant even across symptom domains, classification accuracy analyses revealed that findings based on conjunctions of these scales may not overlap to the degree that the correlations might suggest. Copyright © 2006 John Wiley & Sons, Ltd.**

In the past half-century, expert testimony has played an increasingly important role in the American system of civil law. A significant portion of such testimony concerns defendant claims of mental illness and neuropsychological deficits. Psychologists are frequently called upon to conduct forensic evaluations of defendants who report psychological symptoms. Within this process, the clinician is required to make a judgment as to whether or not—or to what degree—the patient's reported symptoms are genuine. Individuals may be motivated to simulate mental illness or cognitive deficits at several points in the legal process, most notably in cases of insurance claims and in personal injury litigation, when financial compensation can provide a powerful impetus for symptom exaggeration.

Assessment of symptom exaggeration, also termed malingering or dissimulation, has generally relied upon both clinical interviews and formal psychological testing. Although little controlled research has investigated the utility of the clinical interview to detect malingering (Ziskin, 1984), psychological testing has been extensively studied in this context. Indeed, in most cases, decisions about the accuracy of a

---

\*Correspondence to: B. Rosenfeld, Ph.D., Department of Psychology, Fordham University, 441 East Fordham Road, Bronx, NY 10458, U.S.A. E-mail: rosenfeld@fordham.edu

<sup>†</sup>Fordham University.

<sup>‡</sup>Columbia University College of Physicians and Surgeons.

patient's report are aided by results from one or more tests designed specifically for, or adapted to, the detection of malingering. There is no single test of malingering that acts as the "gold standard" in the detection of symptom distortion; instead, there are several available measures, which vary widely in time required for administration, technique, format and theoretical approach. However, there is no means to independently "prove" the existence or extent of mental illness or neuropsychological impairment. Because the information derived from psychological assessment of malingering cannot be verified through comparison to any external standard, the accuracy with which tests can detect symptom distortion has become a major issue in science and the law.

The precision with which an instrument predicts the presence or absence of malingering can be described by its sensitivity, specificity, positive predictive accuracy (PPA), and negative predictive accuracy (NPA). The relative importance ascribed to measures of predictive accuracy varies across contexts, but in general tests seek to maximize all four indices. Sensitivity refers to the proportion of respondents with a given condition who have been identified by a test or procedure. When a test of malingering is highly sensitive, respondents who exaggerate symptoms are likely to obtain scores that exceed cutoffs so are likely to be detected. Specificity refers to the proportion of honest respondents without a given condition for whom the condition is ruled out by the test or procedure. When a test of malingering is highly specific, honest respondents are unlikely to obtain scores that exceed cutoffs and, accordingly, are unlikely to be misclassified as malingering.

Sensitivity and specificity are not affected by changes in base rate of the target condition in the population. The term "base rate" refers to a probability statement, usually expressed as a percentage, regarding the likelihood that a member of a specified population will have a certain characteristic, such as malingering (Kamphuis & Finn, 2002). However, base rate information does not allow the clinician to gauge whether any particular individual is malingering. In contrast to sensitivity and specificity, other measures of predictive accuracy are heavily influenced by the base rate of the target condition (i.e. malingering) in the population under investigation.

Positive predictive accuracy (PPA) reflects the probability that a respondent displays the target condition given a positive test result, or the proportion of respondents predicted to be malingering who actually are malingering. Negative predictive accuracy (NPA) reflects the probability that a respondent does not display the target condition given a negative test result, or the proportion of respondents predicted to be honest who actually are honest. PPA declines as the base rate decreases, so that when malingerers make up only a small percentage of the population, a significant proportion of positive test results will represent "false positives" (i.e. test results indicative of malingering when the respondent has actually answered honestly). Thus, a malingering instrument can exhibit very different rates of diagnostic accuracy depending on the base rate of malingering in the population in which the instrument is to be applied.

A number of studies have offered estimates of the base rate of malingering in specific contexts. Existing literature suggests that the base rate of malingering in general clinical practice settings is, on average, approximately 7% (Rogers, Salekin, Sewell, Goldstein, & Leonard, 1998; Rogers, Sewell, & Goldstein, 1994). Because clients assessed by forensic specialists are typically motivated to distort their symptoms, the average rate of malingering in forensic settings is estimated to be

somewhat higher, at approximately 15–17% (Rogers et al., 1994, 1998). This figure may represent an underestimation, since “successful” malingerers, by definition, are not detected by clinicians and thus would not be included in their subjective base rate estimates (Berry, Baer, Rinaldo, & Wetter, 2002). However, other research suggests that average base rates of malingering are unlikely to exceed 45% even in forensic contexts (Norris & May, 1998).

Clinicians working in forensic settings are apt to be aware of at least a proportion of the voluminous data available regarding the accuracy of various tests and indices of malingering. Because any one test of exaggeration might yield incorrect results (i.e. misclassify an honest respondent as malingering, or the reverse), some researchers recommend the use of multiple measures to provide converging evidence of malingering (e.g. Arnett, Hammeke, & Schwartz, 1995; Berry et al., 2002), but this practice is most helpful when the results of each test are independent of one another—that is, when an incorrect classification based on the results of one test does not increase the likelihood of incorrect classification on another (Rosenfeld, Sands, & van Gorp, 2000). Integrating results from multiple measures can, in theory, offset limitations of a single test, but the practical benefits of using several tests depends on two factors: the extent to which the tests tend to misclassify the same individuals, and whether the examiner requires performance that exceeds cutoffs on only one test or on multiple tests before rendering a determination of malingering.

However, few efforts have been made to investigate the degree of concordance or overlap between instruments commonly used for malingering assessment. Given that the results of two or more malingering measures administered in conjunction are often proffered in legal contexts, such research is particularly important in light of the U.S. Supreme Court’s ruling in *Daubert v. Merrell Dow Pharmaceuticals* (1993), which requires examination of test reliability, validity, and accuracy in order to determine admissibility of scientific evidence. The topic has been addressed only by one recent study (Nelson et al., 2003), which focused solely on measures of neuropsychological effort. Nelson et al. (2003) described correlations between multiple measures, but did not address the issue of classification accuracy. Because these authors examined test scores drawn only from suspected malingerers (without including honest respondents), likely restricting the variance in their data, their failure to detect significant correlations between tests is perhaps not surprising. Further, Nelson et al. (2003) did not include newer, more sophisticated malingering tests (such as the Test of Memory Malingering, TOMM; Tombaugh, 1996; or the Validity Indicator Profile, VIP; Frederick, 1997) in their analyses. The present study sought to assess the inter-test correlations and consistency of malingering classifications among several commonly used malingering measures instruments, including measures of both general psychopathology and neuropsychological deficit, using ecologically valid data drawn from real-world evaluations.

## METHOD

### Participants

Cases were drawn from the medical records of forensic psychologists and were included in this analysis if they had at least two different malingering tests

administered during the course of the evaluation. The sample consisted of 66 individuals who were referred for psychological evaluation for a wide range of psychological disorders (e.g. mood disorders, anxiety, psychotic disorders), somatoform problems (e.g. stress-related abdominal pain, chronic fatigue), personality disorders (e.g. Borderline Personality Disorder, Schizoid Personality Disorder), and cognitive impairments (e.g. organic brain dysfunction, neurological damage, difficulties with memory and concentration). Of the 66 cases available for analysis, 44 (66.7%) were male and 22 (33.3%) were female. Age ranged from 19 to 66 with a mean age of 44.8 years ( $SD = 11.0$ ). All evaluations were conducted in the metropolitan New York City area between 1997 and 2004 by or under the supervision of licensed clinical psychologists. Roughly half of the evaluations ( $n = 35$ ; 53.0%) were performed at the request of the client's insurance company after he or she filed a claim for mental disability. The remaining evaluations (31 of 66; 47.0%) were performed in the course of civil litigation (e.g. personal injury), and were requested by either the plaintiff or respondent.

## Procedures

A two-page survey instrument was developed to collect information about each case available for analysis. Information requested included demographic data about the participant, pre-referral diagnosis and final (post-evaluation) diagnosis, and raw scores for selected indices from several tests used to assess malingering. The tests analyzed included the MMPI-2 (Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989), the MCMI-III (Millon, Davis, & Millon, 1997), the Fifteen Item Test (FIT; Rey, 1964), the TOMM (Tombaugh, 1996), and the VIP (Frederick, 1997). Although some of these tests have multiple malingering indices, we focused on a subset of the most widely used and well validated indices, such as the MMPI-2's  $F$  (Butcher *et al.*, 1989) and  $F(p)$  (Arbisi & Ben-Porath, 1995, 1998) scales and  $F - K$  index (Gough, 1950) and the  $X$  and  $Z$  scales of the MCMI-III (Millon *et al.*, 1997). Evaluators occasionally reported the use of other malingering tests, such as the Structured Interview of Reported Symptoms (SIRS; Rogers, Bagby, & Dickens, 1992) and Personality Assessment Inventory (Morey, 1991), but were not used frequently enough to permit meaningful analysis.

### *Minnesota Multiphasic Personality Inventory—Second Edition (MMPI-2)*

The MMPI-2 (Butcher *et al.*, 1989) consists of 567 true/false questions contributing to several validity and clinical scales. High scores on the Infrequency ( $F$ ) scale, the most widely used MMPI-2 validity indicator, may indicate symptom exaggeration (Butcher *et al.*, 1989) or severe psychological disturbance (Arbisi & Ben-Porath, 1995, 1998). The Infrequency—Psychopathology ( $F(p)$ ) was developed to discriminate feigned from genuine psychopathology in settings in which high reported levels of psychiatric symptoms may be expected (Arbisi & Ben-Porath, 1995). The  $F - K$  index identifies a tendency to exaggerate symptoms relative to a tendency to deny them by subtracting the raw score on the  $K$  scale (which assesses defensiveness) from the raw score on the  $F$  scale (Gough, 1950). Berry *et al.* (2002)

reported a wide variance in cutoff scores employed across studies when these indices were used to identify malingering, ranging from 62 to 120 for the F scale, from 70 to 120 for the F(p) scale, and from -4 to 25 for the  $F - K$  index. Mean reported cutoff scores across studies were T scores of 106 or greater for the F scale, T scores of 96 or greater for the F(p) scale and a difference score of 15.6 or greater for the  $F - K$  index (Berry et al., 2002).

### *Millon Clinical Multiaxial Inventory—Third Edition (MCMI-III)*

The MCMI-III (Millon et al., 1997) includes 175 true/false items scored on 24 content scales intended to correspond to major DSM-IV Axis I and II disorders. The two scales most often used to detect feigning are the Disclosure Index (X), which assesses willingness to admit to difficulties, and the Debasement Index (Z), which identifies a tendency to overstate emotional and personal problems. Although the test's manual proposes that raw scores above 178 on Scale X denote excessive symptom exaggeration, no cutoff scores are recommended for use with Scale Z beyond the suggestion that base rate scores above 85 tend to be associated with malingering (Millon, 1994; Millon et al., 1997). Few studies have explored the ability of the MCMI-III to detect malingering, and those that exist have typically suggested poor classification accuracy for this measure (Daubert & Metzler, 2000; Schoenberg, Dorr, & Morgan, 2003).

### *Fifteen Item Test (FIT)*

The FIT (Rey, 1964) is a simple test of feigned visual memory impairment. Respondents are asked to memorize 15 familiar items that appear in a logical sequence (e.g. A, B, C; 1, 2, 3). Scores typically reflect the number of items correctly recalled, irrespective of the order or spatial location in which they are recalled. According to Lezak, Howieson, and Loring (2004), all but the most severely impaired respondents should be able to recall at least three lines or nine items. However, Morgan (1991) found that 20% of patients with mild to severe memory impairment failed the FIT at a cutoff score of nine, leading some researchers to advocate for a cutoff of seven items or below to identify malingering (Meyer & Deitsch, 1995). Even at this lower cutoff, the FIT has been criticized as being relatively insensitive and nonspecific to malingering (Guilmette, Hart, Giuliano, & Leininger, 1994; Schretlen, Brandt, Krafft, & van Gorp, 1991; Vallabhajosula & van Gorp, 2001). Yet the FIT remains frequently used in evaluations of malingered neurocognitive deficits (Slick, Tan, Strauss, & Hultsch, 2004).

### *Test of Memory Malingering (TOMM)*

The TOMM (Tombaugh, 1996) assesses exaggerated memory complaints using a more sophisticated approach than that of the FIT. Each of two learning trials presents 50 line drawings of common objects, followed by a 50-item test series requiring respondents to select the previously seen stimulus from a set of two-choice recognition panels. An optional retention trial, given 15 minutes after the two

learning trials, presents a test series without administration of the target drawings. Scores below 45 (90% correct) on the second or retention trial suggest that respondents have not performed to the best of their abilities (Tombaugh, 1996). Normative data indicate that, unlike the FIT, the TOMM is highly accurate in differentiating malingering individuals from normal controls (Rees *et al.*, 1998; Weinborn, Orr, Woods, Conover, & Feix, 2003). Although persons with dementia appear apt to fail the TOMM, most genuinely impaired patients perform well above cutoffs on this measure (Teichner & Wagner, 2004; Tombaugh, 1997).

### *Validity Indicator Profile (VIP)*

The VIP (Frederick, 1997) consists of verbal and nonverbal subtests that present items in randomized order of difficulty. The 78 items of the verbal subtest require respondents to select one of two words most similar in meaning to the stimulus word, while the nonverbal subtest is composed of 100 matrix reasoning items derived from a previously existing test of nonverbal intelligence. Each individual's performance is assessed using several validity measures; cutoff scores for each measure were derived such that 90% of honest responders would be accurately classified (Frederick, 1997). However, in practice, the total subtest scores may be the most salient indices attended to by clinicians. For the VIP's verbal subtest, dishonest responding is suspected when the total subtest score is 59 or below, while for the nonverbal subtest, suspicions of malingering are raised when the total subtest score is 75 or below.

### **Statistical Analysis**

Test data were classified as "probable malingering" or "not malingering" based on cutoff scores derived from the relevant literature as described above. For the MMPI-2, cases were classified as probable malingerers if they had *T* scores of 106 or greater on the *F* scale, *T* scores of 96 or greater on the *F*(p) scale, or difference scores of 16 or greater on the *F* – *K* index. For the MCMI-III, cutoff scores for determination of malingering were set at raw scores of 178 or greater for Scale *X* or base rate scores of 85 or greater on Scale *Z*. Probable malingerers on the FIT were those who obtained scores of 7 or fewer items recalled immediately after administration or following a delay. For the TOMM, probable malingerers were those who obtained scores of 44 or fewer correct responses during the second learning trial or the optional retention trial. Probable malingerers on the VIP were those who earned raw verbal subtest scores of 59 or below or raw nonverbal subtest scores of 75 or below.

Because of the wide variation in reported cutoff scores for many of the indices reviewed, and because real-world clinical decision making often cannot be reduced to a yes/no determination, cases with scores that approached but did not exceed cutoffs were classified as "indeterminate" for the purposes of description. For the MMPI-2, indeterminate cases were those with *T* scores between 90 and 105 on the *F* scale or between 70 and 95 on the *F*(p) scale, as well as those with *F* – *K* indices between 9 and 15. Cases were also classified as indeterminate, if they obtained raw scores between 149 and 177 on Scale *X* or base rate scores between 75 and 84 on Scale *Z* of the MCMI-III. For the FIT, indeterminate cases were those with scores

between 8 and 9 on either trial, whereas for the TOMM, indeterminate cases were those with scores between 45 and 46 on either trial. For the VIP, indeterminate cases were those who were classified as “careless” or “irrelevant” responders on the VIP verbal or nonverbal subscales.

Classification accuracy for each measure was determined by comparing “predicted” classifications (i.e. those based on cutoff scores) with “actual” classifications based on results from the MMPI-2 *F* scale and the TOMM. The *F* scale has repeatedly been cited as one of the best available discriminators of honest versus malingered symptom profiles (see, e.g., Bagby, Rogers, & Buis, 1994; Berry, Baer, & Harris, 1991; Rogers, Sewell, & Ustad, 1995), and published data regarding the TOMM suggest a high rate of classification accuracy (Rees et al., 1998). Because many clinicians who administer multiple tests of malingering may encounter discrepant findings, analyses comparing two of the most well validated malingering classifications against other, less widely studied measures provides important data regarding test overlap.

The data were described in terms of relative percentages of cases assigned to each classification (“honest,” “indeterminate,” and “malingering”) by each measure. To assess the degree to which measures provided overlapping information, Pearson correlations were calculated. As a means of assessing classification accuracy, sensitivity and specificity were calculated for each test using *F*-scale scores and TOMM scores as “index” variables. PPA and NPA were not calculated, as the base rate of malingering in the current sample was unknown. Cases categorized as “indeterminate” on the basis of test scores were grouped with those categorized as “malingering” based on test performance for the purposes of sensitivity and specificity calculations. Of note, correlational and classification analyses that contained fewer than ten cases were omitted, as the small sample size precluded meaningful interpretation.

## RESULTS

### Determinations Based on Indices

Table 1 reports classification data for each index based on the cutoff scores described above. Data on the MMPI-2 *F* scale were available for all 66 respondents and data on

Table 1. Percentages of response classes for malingering indices

Index	Cases with data	Cases classified honest (%)	Cases classified indeterminate (%)	Cases classified malingering (%)
MMPI-2 <i>F</i>	66	54 (81.8%)	4 (6.1%)	8 (12.1%)
MMPI-2 <i>F-K</i>	61	48 (78.7%)	4 (6.6%)	9 (14.8%)
MMPI-2 <i>F(p)</i>	52	39 (75.0%)	10 (19.2%)	3 (5.8%)
MCMI-III <i>X</i>	36	32 (88.9%)	3 (8.3%)	1 (2.8%)
MCMI-III <i>Z</i>	36	23 (63.9%)	5 (13.9%)	8 (22.2%)
FIT	26	20 (76.9%)	2 (7.7%)	4 (15.4%)
TOMM	52	28 (53.8%)	5 (9.6%)	19 (36.5%)
VIP verbal	31	22 (71.0%)	6 (19.4%)	3 (9.7%)
VIP nonverbal	29	13 (44.8%)	11 (37.9%)	5 (17.2%)

*N* = 66.

the *F* – *K* scale were available for 61 respondents, but data for the *F*(*p*) scale were available for only 52 cases (i.e., because the raw data were often unavailable, only those cases in which the scale had been calculated were available for this analysis). Of the 66 evaluatees, 52 completed at least two trials of the TOMM (i.e. not including the “retention” trial), 36 completed the MCMI-III, 26 completed the FIT, 31 completed the VIP verbal scale and 29 completed the VIP nonverbal scale.

Scale X of the MCMI-III classified the greatest proportion of respondents as honest (32 of 36; 88.9%) and classified the smallest proportion of respondents as probable malingerers (1 of 36; 2.78%); three of 36 respondents (8.33%) were classified as indeterminate. In contrast, the Nonverbal scale of the VIP classified the smallest proportion of respondents as honest (13 of 29; 44.8%). The TOMM classified the greatest proportion of respondents as probable malingerers (19 of 52; 36.5%). The *F*(*p*) scale of the MMPI-2 classified the greatest proportion of respondents as indeterminate (10 of 52; 19.2%), while the *F* scale resulted in the lowest proportion of indeterminate classifications (4 of 66; 6.1%).

### Correlations between Indices

Table 2 reports Pearson correlations among the nine indices reviewed. High correlations were obtained among the three MMPI-2 scales, as may be expected when comparing indices that are drawn from the same measure and that share several items (Greene, 2000). Similarly, a high correlation was obtained between the *X* and *Z* scales of the MCMI-III. The majority of the highest *inter*-test correlations, ranging from 0.46 to 0.78, were obtained between indices of the MMPI-2 and the MCMI-III, suggesting that these tests provide somewhat overlapping information. Indeed, almost every test analyzed was significantly correlated with the three MMPI-2 indices, although the correlations tended to be positive for tests of psychopathology and negative for tests of cognitive effort (see Table 2). Correlations between many cognitive measures were also high. For example, the correlations between the FIT and TOMM ( $r = 0.74$ ,  $p < 0.01$ ), and between the verbal and nonverbal subscales of the VIP ( $r = 0.56$ ,  $p < 0.01$ ), were large and statistically significant. Unfortunately,

Table 2. Intercorrelations among indices of malingering

Index	1	2	3	4	5	6	7	8	9
1. <i>F</i>	—	0.78**	0.66**	0.68**	0.64**	-0.41*	-0.45**	-0.25	0.11
2. <i>F</i> – <i>K</i>	(61)	—	—	0.75**	0.78**	0.71**	-0.34	-0.40**	-0.19
3. <i>F</i> ( <i>p</i> )	(52)	(48)	—	0.54**	0.46*	-0.41*	-0.53**	-0.13	-0.37
4. <i>X</i>	(36)	(33)	(33)	—	0.73**	-0.04	-0.37*	-0.170	0.16
5. <i>Z</i>	(36)	(33)	(33)	(36)	—	-0.16	-0.36	-0.13	-0.05
6. FIT	(26)	(25)	(25)	(17)	(17)	—	0.74**	X	X
7. TOMM	(52)	(48)	(44)	(29)	(29)	(23)	—	0.35	0.18
8. VIP-V	(31)	(29)	(29)	(13)	(13)	X	(25)	—	0.56**
9. VIP-N	(29)	(27)	(21)	(12)	(12)	X	(24)	(27)	—

Values enclosed in parentheses represent cell sizes.

\* $p < 0.05$ .

\*\* $p < 0.01$ .

X: statistics not reported because fewer than ten cases were available for analysis.

Table 3. Sensitivity and specificity of indices using *F*-scale scores as the index variable

Index	Sensitivity	Specificity	No. errors (% classified as indeterminate)
MMPI-2			
<i>F</i> – <i>K</i>	0.83	0.94	5 (60.0%)
<i>F</i> ( <i>p</i> )	0.69	0.92	7 (57.1%)
MCCI-III			
<i>X</i>	0.50	0.93	7 (85.7%)
<i>Z</i>	0.86	0.96	8 (75.0%)
FIT	0.17	0.74	11 (45.5%)
TOMM	0.82	0.63	17 (35.3%)
VIP			
Verbal	0.75	0.74	7 (28.6%)
Nonverbal	0.75	0.92	14 (42.9%)

several of these associations were based on small numbers of cases and were therefore not reported (i.e., statistics based on fewer than ten cases were omitted).

### Sensitivity and Specificity

Table 3 reports sensitivity and specificity for each measure using *F*-scale scores as the index variable. Overall, with the exception of the FIT, tests achieved moderate to high sensitivity in conjunction with moderate to high specificity. Both sensitivity and specificity tended to be higher for measures targeted towards feigning psychopathology, as is the *F* scale. However, cognitive malingering tests, particularly the TOMM and VIP subscales, displayed at least moderate sensitivity when compared to the *F*-scale results, indicating that these measures capture many of the same malingerers as does the MMPI-2. Yet except in the case of the VIP nonverbal scale, the moderate sensitivity estimates for cognitive malingering were accompanied by lower specificity, suggesting that the cognitive measures also identified some malingerers that were not detected by the *F* scale.

Table 4 reports the sensitivity and specificity for each measure using TOMM scores as the index variable. The pattern of findings is quite different from those

Table 4. Sensitivity and specificity of indices using TOMM scores as the index variable

Index	Sensitivity	Specificity	No. errors (% classified as indeterminate)
MMPI-2			
<i>F</i>	0.38	0.93	17 (29.4%)
<i>F</i> – <i>K</i>	0.39	0.88	17 (47.0%)
<i>F</i> ( <i>p</i> )	0.53	0.79	11 (45.5%)
MCCI-III			
<i>X</i>	0.30	0.95	8 (37.5%)
<i>Z</i>	0.58	0.72	8 (50.0%)
FIT	0.36	1.00	7 (28.6%)
VIP			
Verbal	0.75	0.93	4 (25.0%)
Nonverbal	0.75	0.50	9(44.4 %)

based on the *F* scale. Notably, many of the tests reviewed displayed much higher specificity than sensitivity. The moderate sensitivity estimates of the VIP verbal (0.75) and nonverbal (0.75) subscales may be expected given that all three tests focus on cognitive functioning. It is also not surprising that measures of exaggerated psychiatric symptomatology largely failed to detect feigning among individuals identified as malingerers by the TOMM. However, even tests of psychopathology ruled out malingerers among many of the individuals identified as honest by the TOMM.

## DISCUSSION

The practice of combining results from separate measures to detect malingerers is thought to reduce classification errors by providing convergent evidence of symptom exaggeration. However, this study represents one of the first systematic efforts to investigate this assumption, by assessing the degree of overlap between different malingerers instruments. The present study utilized data from respondents believed to be honest as well as those suspected to be malingerers, and incorporated measures of exaggerated cognitive impairment and psychopathology, modeling the full range of behavior likely to be encountered in forensic contexts.

These analyses demonstrated a high degree of concordance (i.e. high correlations) among the raw scores of these multiple measures of malingerers. Large and statistically significant correlations were not limited to different scales drawn from the same measure, although correlations were strongest within a single test (e.g. the *F* scale and *F*–*K* index of the MMPI-2). High inter-test correlations were also obtained between tests assessing similar types of malingerers (i.e. between MMPI-2 scales and MCMI-III scales and between the TOMM and the FIT). More importantly, tests intended to detect malingerers general psychopathology (e.g. the MMPI-2) also displayed significant correlations of moderate to high magnitude with those designed to assess cognitive effort (e.g. the FIT, and particularly the TOMM). This suggests that, in the current sample, malingerers tended to be a more global construct, with those who showed indications of exaggerated psychopathology also displaying evidence of exaggerated cognitive deficit.

Although the correlations between different measures of malingerers might lead one to conclude that these measures appear highly redundant, examination of classification decisions revealed important discrepancies. When the *F* scale indicated that the individual was likely to be malingerers, the majority of other measures were in agreement with this conclusion (i.e. moderate to high sensitivity). However, the reverse was not always the case—when the *F* scale suggested honest responding, other measures, particularly those targeted to cognitive effort, may well have identified probable symptom exaggeration (i.e. only moderate specificity). Of course, the finding that cognitive malingerers measures such as the TOMM, FIT, and VIP verbal scale identified probable malingerers that were not captured by the MMPI-2 may simply reflect a more sophisticated form of malingerers. Nevertheless, these data have important implications for clinical practice, particularly in civil forensic settings, where individuals may selectively exaggerate cognitive symptoms. Although neuropsychological deficits can impact MMPI-2 performance, the *F* scale is not intended to reflect cognitive effort (Greene, 2000) and is generally thought to

be insensitive to feigning of neurocognitive dysfunction (Larrabee, 2003). These data support that presumption.

In contrast to the MMPI, the TOMM showed the reverse pattern of results. When the TOMM was used as the index variable for classification accuracy, most of the instruments examined had moderate to high levels of specificity but much poorer sensitivity. Thus, individuals classified as probable malingerers by TOMM were not necessarily identified as malingerers by other measures, particularly (but not limited to) measures of exaggerated psychiatric symptoms. However, those classified as honest responders by the TOMM were rarely identified as malingering by other measures. For example, both a measure of psychopathology (i.e. scale *X* of the MCMI-III) and a test of neurocognitive functioning (i.e. the FIT) showed high levels of agreement with the TOMM in identifying dishonest responding. These findings might reflect either the greater sensitivity of the TOMM to detecting subtle forms of malingering, or might indicate a tendency to over-classify malingering. Given that other research has not supported the latter conclusion (Rees et al., 1998; Weinborn et al., 2003), the former appears more likely, although further research using analog methods is necessary to fully understand these associations.

Despite these novel results, these findings must be qualified by several methodological limitations. For example, data were drawn from a convenience sample of individuals referred for civil forensic evaluation in the New York City area. As such, it is not possible to determine the extent to which these cases are representative of malingering (or honest) evaluatees in the New York City area, let alone the U.S. more generally. Moreover, because these data were drawn from clinical practice rather than controlled research, some tests that might be desirable to include in a study of malingering (e.g. the SIRS) were rarely available, and therefore were not included in analyses. Of course, because these data were solicited from civil litigation and disability cases, tests of feigned psychosis such as the SIRS may have been much less informative in relation to the referral questions likely to be posed in such settings. Nevertheless, further research using other samples and other tests is clearly necessary.

A comparison of results across tests was also complicated by the fact that not all cases included data for all tests reviewed; the VIP subscales were particularly underrepresented in this sample. As such, some large correlation coefficients failed to reach statistical significance due to the small sample size for such comparisons (e.g., a correlation of 0.70 between the VIP verbal subscale and the FIT included only four cases in which both tests were administered concurrently). Clearly any conclusions based on such small samples must be considered tentative; however, the consistency between the pattern of these associations and those that were estimated with larger samples increases confidence that these data may present a valid picture of the inter-relationships among tests. Although such small cells might have been eliminated from analyses altogether, inclusion of these data allows clinicians and future researchers to compare or combine their data with those presented here.

Perhaps most importantly, our retrospective methodology did not permit an independent assessment of the likelihood of malingering (i.e. without reliance on these same test data). This method, coupled with our naturalistic assessment, precludes accurate categorization of these subjects as genuinely “malingering” or “honest” (and hence our analyses focused on comparing measures to one another, rather than to a “gold standard”). Of course, the use of actual forensic evaluations

(rather than an analog study where college students are asked to present exaggerated symptomatology) provides precisely the ecological validity that is lacking from many studies of malingering. Nevertheless, continued research into the overlapping natures of certain malingering measures is needed before any firm conclusions can be drawn.

Along with attempts to replicate the present study with a larger and more representative sample, future efforts should include a wider array of indices, particularly those with promise in assessing both exaggeration of psychopathology and subnormal cognitive effort (e.g. the MMPI-2 Fake-Bad Scale; Lees-Haley, English, & Glenn, 1991). Investigation into how tests are used in practical settings (i.e. the decision rules used by forensic clinicians in making determinations of malingering given inconsistent results from two or more tests) will also help clarify the real-world impact of the overlap identified in this study. Only with continued refinement can clinicians provide accurate information regarding the combinations of tests that are used in forensic evaluation.

## 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.
- Arbisi, P. A., & Ben-Porath, Y. S. (1998). The ability of Minnesota Multiphasic Personality Inventory-2 validity scales to detect fake-bad responses in psychiatric inpatients. *Psychological Assessment*, 10, 221–228.
- 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.
- Bagby, R. M., Rogers, R., & Buis, T. (1994). Malingered and defensive response styles on the MMPI-2: An examination of validity scales in a forensic population. *Journal of Personality Assessment*, 62, 191–203.
- Berry, D. T. R., Baer, R. A., & Harris, M. J. (1991). Detection of malingering on the MMPI: A meta-analytic review. *Clinical Psychology Review*, 11, 585–598.
- Berry, D. T. R., Baer, R. A., Rinaldo, J. C., & Wetter, M. W. (2002). Assessment of malingering. In J. N. Butcher (Ed.), *Clinical personality assessment: Practical approaches* (2nd ed., pp. 269–302). New York: Oxford University Press.
- Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *MMPI-2: Manual for administration and scoring*. Minneapolis, MN: University of Minnesota Press.
- Daubert v. Merrell Dow Pharmaceuticals, Inc. 113 S.Ct. 2786 (1993).
- Daubert, S. D., & Metzler, A. E. (2000). The detection of fake-bad and fake-good responding on the Millon Clinical Multiaxial Inventory III. *Psychological Assessment*, 12, 418–424.
- Frederick, R. I. (1997). *Validity Indicator Profile: Manual*. Minneapolis, MN: National Computer Systems.
- Gough, H. G. (1950). The *F* minus *K* dissimulation index for the MMPI. *Journal of Consulting Psychology*, 14, 408–413.
- Greene, R. L. (2000). *The MMPI-2: An interpretive manual* (2nd ed.). Boston: Allyn and Bacon.
- 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 Hiscock forced-choice procedure. *The Clinical Neuropsychologist*, 8, 283–294.
- Kamphuis, J. H., & Finn, S. E. (2002). Incorporating base rate information in daily clinical decision making. In J. N. Butcher (Ed.), *Clinical personality assessment: Practical approaches* (2nd ed., pp. 257–268). New York: Oxford University Press.
- Larrabee, G. J. (2003). Detection of symptom exaggeration with the MMPI-2 in litigants with malingered neurocognitive dysfunction. *The Clinical Neuropsychologist*, 17, 54–68.
- Lees-Haley, P., English, L. T., & Glenn, W. J. (1991). A fake bad scale on the MMPI-2 for personal injury claimants. *Psychological Reports*, 68, 203–210.
- Lezak, M. D., Howieson, D. B., & Loring, D. W. (1995). *Neuropsychological assessment* (4th ed.). New York: Oxford.

- Meyer, R. G., & Deitsch, S. M. (1995). The assessment of malingering in psychodiagnostic evaluations: Research-based concepts and methods for consultants. *Consulting Psychology: Practice and Research*, 47, 234–245.
- Millon, T. (1994). *The Millon Clinical Multiaxial Inventory—III manual*. Minneapolis, MN: National Computer Systems.
- Millon, T., Davis, R., & Millon, C. (1997). *The Millon Clinical Multiaxial Inventory—III manual* (2nd ed.). Minneapolis, MN: National Computer Systems.
- Morey, L. C. (1991). *Personality Assessment Inventory professional manual*. Odessa, FL: Psychological Assessment Resources.
- Morgan, S. F. (1991). Effect of true memory impairment on a test of memory complaint validity. *Archives of Clinical Neuropsychology*, 6, 327–334.
- Nelson, N. W., Boone, K., Dueck, A., Wagener, L., Lu, P., & Grills, C. (2003). Relationships between eight measures of suspect effort. *The Clinical Neuropsychologist*, 17, 263–272.
- Norris, M. P., & May, M. C. (1998). Screening for malingering in a correctional setting. *Law and Human Behavior*, 22, 315–323.
- Rees, L. M., Tombaugh, T. N., Gansler, D. A., & Moczynski, N. P. (1998). Five validation experiments of the Test of Memory Malingering (TOMM). *Psychological Assessment*, 10, 10–20.
- Rey, A. (1964). *L'examen clinique en psychologie [The clinical examination in psychology]*. 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., Salekin, R. T., & Goldstein, A. (1994). Explanatory models of malingering: A prototypical analysis. *Law and Human Behavior*, 18, 543–552.
- Rogers, R., Salekin, R. T., Sewell, K. W., Goldstein, A., & Leonard, K. (1998). A comparison of forensic and nonforensic malingerers: A prototypical analysis of explanatory models. *Law and Human Behavior*, 22, 353–367.
- Rogers, R., Sewell, K. W., & Ustad, K. L. (1995). Feigning among chronic outpatients on the MMPI-2: A systematic examination of fake-bad indicators. *Assessment*, 2, 81–89.
- Rosenfeld, B., Sands, S. A., & van Gorp, W. G. (2000). Have we forgotten the base rate problem? Methodological issues in the detection of distortion. *Archives of Clinical Neuropsychology*, 15, 349–359.
- Schoenberg, M. R., Dorr, D., & Morgan, C. D. (2003). The ability of the Millon Clinical Multiaxial Inventory—Third Edition to detect malingering. *Psychological Assessment*, 15, 198–204.
- Schretlen, D., Brandt, J., Krafft, L., & van Gorp, W. G. (1991). Some caveats in using the Rey 15-Item Memory Test to detect malingered amnesia. *Psychological Assessment*, 3, 667–672.
- Slick, D. J., Tan, J. E., Strauss, E. H., & Hultsch, D. F. (2004). Detecting malingering: A survey of expert's practices. *Journal of Clinical Neuropsychology*, 19, 465–473.
- Teichner, G., & Wagner, M. T. (2004). The test of Memory Malingering (TOMM): Normative data from cognitively intact, cognitively impaired, and elderly patients with dementia. *Archives of Clinical Neuropsychology*, 19, 455–464.
- Tombaugh, T. N. (1996). *Test of Memory Malingering*. Toronto: Multi-Health Systems.
- Tombaugh, T. N. (1997). The test of Memory Malingering (TOMM): Normative data from cognitively intact and cognitively impaired individuals. *Psychological Assessment*, 9, 260–268.
- Vallabhajosula, B., & van Gorp, W. G. (2001). Post-*Daubert* admissibility of scientific evidence on malingering of cognitive deficits. *Journal of the American Academy of Psychiatry and the Law*, 29, 207–215.
- Weinborn, M., Orr, T., Woods, S. P., Conover, E., & Feix, J. (2003). A validation of the Test of Memory Malingering in a forensic psychiatric setting. *Journal of Clinical and Experimental Neuropsychology*, 25, 979–990.
- Ziskin, J. (1984). Malingering of psychological disorders. *Behavioral Sciences and the Law*, 2, 39–49.