Psychological Assessment

Estimating the Accuracy of Neurocognitive Effort Measures in the Absence of a “Gold Standard”
Douglas Mossman, Dustin B. Wygant, and Roger O. Gervais
Online First Publication, April 30, 2012. doi: 10.1037/a0028195

CITATION
Estimating the Accuracy of Neurocognitive Effort Measures in the Absence of a “Gold Standard”

Douglas Mossman
University of Cincinnati College of Medicine and University of
Cincinnati College of Law

Dustin B. Wygant
Eastern Kentucky University

Roger O. Gervais
Neurobehavioural Associates, Edmonton, Alberta, Canada, and University of Alberta

Psychologists frequently use symptom validity tests (SVTs) to help determine whether evaluees’ test performance or reported symptoms accurately represent their true functioning and capability. Most studies evaluating the accuracy of SVTs have used either known-group comparisons or simulation designs, but these approaches have well-known limitations (potential misclassifications or lack of ecological validity). This study uses latent class modeling (LCM) implemented in a Bayesian framework to estimate SVT classification accuracy based on data obtained from real-life forensic evaluations. We obtained archival data from 1,301 outpatient evaluees who underwent testing with the Computerized Assessment of Response Bias (CARB), the Test of Memory Malingering (TOMM), and the Word Memory Test (WMT) in a forensic evaluation context. Under various data models, Markov chain Monte Carlo methods implemented via WinBUGS converged to target distributions that permitted Bayesian estimates of SVT accuracy. Under the most plausible model (conditional dependence in test results), classification accuracies (expressed as area under the “trapezoidal” receiver operating characteristic curve) were as follows: CARB = 0.765 ± 0.030, WMT = 0.929 ± 0.020, and TOMM = 0.771 ± 0.034. At decision thresholds that hold false positive rates at 0.02, the SVTs would detect invalid responses (true positives) at rates of approximately 35%, 65%, and 49%, respectively, for the 3 tests. Though LCM methods have limitations, this study suggests that they offer an approach to SVT evaluation that avoids methodological pitfalls of known-group research designs while retaining ecological validity that is absent in simulation studies.

Keywords: malingering, Bayesian inference, gold standard, neuropsychological testing, ROC analysis

Supplemental materials: http://dx.doi.org/10.1037/a0028195.supp

Most psycholegal assessments require evaluators to consider whether evaluees’ performances on tests and descriptions of their symptoms are valid representations of their functioning and capability. Failing to detect feigned problems (“false negative” errors) can lead to incorrect legal determinations (e.g., a finding of incompetence to stand trial) or the institution of treatment that is inappropriate and (if side effects occur) potentially injurious. Improperly designating persons as malingeringers (“false positive” errors) or erroneously concluding that evaluees have performed at less than their true ability can result in wrongful deprivation of social entitlements, incorrect legal verdicts, or failure to administer needed treatment.

Broadly, mental health evaluators use three approaches to detect inaccurate self-portrayals. They can compare an evaluee’s reported or claimed severity of symptoms to known patterns of symptom presentation displayed by patients who have no motivation to appear ill (see, e.g., Resnick & Knoll, 2008); large disparities between reported symptoms and typical symptom patterns suggest feigning or exaggeration. Evaluators can also use “collateral” data (information from records or the observations of other individuals) to make inferences about truthfulness by comparing the evaluee’s claims with others’ reports of or records describing the evaluee’s behavior and functioning (Resnick, West, & Payne, 2008). Finally, examiners can use one of several assessment instruments that are either specifically designed to detect test-taking behavior associated with malingering (e.g., the Structured Interview of Reported Symptoms-2; Rogers, 2008) or that use internal “validity scales” to characterize how the evaluee approached the assessment (e.g., the Minnesota Multiphasic Personality Inventory-2; Butcher et al., 2001).

No method for detecting invalid responding is perfect, but assessment instruments have the advantage of being the most easily testable and potentially amenable to quantitative evaluation. Investigators have used two main approaches (sometimes in combination) to estimate the accuracy of such instruments. In what are termed criterion-group or known-group comparisons, investiga-
tors make inferences about the instrument’s accuracy by comparing how two clinical populations—a group that the investigators believe are reporting symptoms honestly and a group that the investigators believe are malingering—perform. In simulation studies, investigators ask “normal” (non-ill) subjects to respond to the instrument’s items as they would if they really had mental disorders and compare these simulated responses to the item responses of persons who (the investigators believe) are honestly reporting symptoms of their mental illnesses.

Known-group and simulation designs have drawbacks. The chief problem with the known-group approach is that what is “known” is not known for certain. That is, a subject’s assignment to the “honest” or “malingering” subgroup depends on the investigators’ application of an independent but imperfect criterion for the truth of the subject, which means that misclassifications and misestimates of accuracy are always possible. To avoid classification problems, investigators may exclude equivocal, hard-to-classify subjects from known-group studies. Of course, using only easier cases and excluding equivocal, “tougher” cases from studies will likely lead to overestimates of accuracy in populations that contain both easy and difficult cases. Moreover, it is the hard-to-classify cases—that is, individuals whom clinicians cannot confidently categorize as valid or invalid responders based on other data sources—that often constitute the group of evaluees concerning whom clinicians most want information obtained from assessment instruments.

Simulation methods always raise the question of whether individuals who are instructed to pretend they are impaired or ill actually produce the types of responses that real-world malingerers produce. In real-world contexts, individuals who falsely portray problems or symptoms can sustain significant negative consequences (e.g., imprisonment or loss of benefits) if they are unsuccessful, whereas persons recruited for simulation studies face no penalty if they fail to convince. Investigators have tried to overcome the limitations of known-group and simulation designs through combining groups and using cross-validation methods, but these only raise the question of which method gives the “right” answer.

If it were possible to use data from actual forensic evaluations to make accuracy judgments about symptom validity tests (SVTs), investigators could examine and base inferences on the behavior of evaluees acting under real-world testing environments and real-world incentives. As we have noted earlier, however, the problem for an accuracy study is that one cannot be sure about the true condition of individuals undergoing real-life evaluations. That is, investigators rarely can know for certain (or even with high confidence) whether a particular evaluee is responding honestly. In medicine, long-term outcomes or biopsy data often provide “gold standard” diagnoses that let investigators be virtually sure whether a subject had a disease at the time of testing. But mental health diagnoses lack gold standards, and in forensic psychiatry and psychology, investigators can rarely do better than make their own best guesses about evaluees’ true status. Often, these imperfect determinations depend in part on data from the very instruments for which the investigators are trying to evaluate accuracy.

Over the last quarter century, latent class modeling (LCM; Uebersax & Grove, 1990) has allowed investigators in subject areas as diverse as imaging liver metastases (Henkelman, Kay, & Bronskill, 1990) and detecting infections in dairy cattle (Choi, Johnson, Collins, & Gardner, 2006) to evaluate diagnostic accuracy using receiver operating characteristic (ROC) analyses without gold standards. Implementing the LCM approach involves evaluating the same cases with multiple diagnostic modalities, which often permits statistical identification of models that include accuracy parameters for those modalities. Mossman et al. (2010) described the potential usefulness of this approach in examining assessment of adjudicative competence and showed that evaluators who based competence judgments on written reports appeared to be highly accurate. In this study, we investigated whether LCM methods could use data from real-life subjects evaluated in a forensic context to generate inferences about the accuracy of cognitive effort tests used to evaluate potential neuropsychological malingering.

Method

This study received approval from the institutional review boards of the University of Cincinnati and Eastern Kentucky University. The archival data used in the study came from approximately 2,700 consecutive evaluees referred to the third author’s private psychology practice over the past two decades. The subjects had no known vulnerabilities (e.g., dementia or severe cognitive disability) that would require special cautions beyond those normally used during a psychological evaluation. The focus of the present study was those evaluees who underwent testing with three commonly used and well-validated cognitive SVTs—the Computerized Assessment of Response Bias (CARB; Allen, Conder, Green, & Cox, 1997), the Test of Memory Malingering (TOMM; Tombaugh, 1996), and the Word Memory Test (WMT; Green, 2003; Green, Allen, & Astner, 1996). We focused on the trio of CARB, WMT, and TOMM in the present study because this yielded the largest number of evaluees to work with. The CARB was the first SVT introduced into the practice’s evaluation battery (beginning in the early 1990s), followed by the WMT and TOMM. Other SVTs were also introduced as they became available. Toward the end of the data accumulation period, use of the CARB and TOMM was discontinued as more effective SVTs were developed and incorporated into the test battery. This meant that some evaluees did not take all three SVTs because they were not yet available at the time of evaluation or because the practice had stopped administering the CARB or TOMM when their evaluation occurred. Also, a few evaluees may have completed only one or two of the SVTs of interest due to compliance issues, slow work speed, or limited literacy that precluded administration of the WMT.

We excluded from analysis results from evaluees who did not take all three SVTs. This left us with data from 1,301 predominately non-head injury disability claimants and counseling clients. We could identify no demographic or other systematic differences between evaluees included in and excluded from the study.

Major reasons for undergoing evaluation were assessments for workers’ compensation (54%), litigation issues (e.g., personal injury; 22%), or long-term disability. The sample was 59% men, with age $M \pm SD = 40.4 \pm 11.0$ years and education $M \pm SD = 12 \pm 2.3$ years. English was the first language of 88% of the sample. The remaining 12% spoke a variety of languages—including Punjabi, Mandarin, Arabic, Spanish, Polish, and Ukrainian—as their first language. Of these individuals, 98% were sufficiently fluent in English to undertake clinical interviews and assessments without interpreters. Native English speakers obtained a mean Wide Range Achievement Test (Wilkinson, & Robertson,
Reading standard score of 93.9 ± 11.3, compared with 85.1 ± 14.9 in individuals for whom English was a second language.

The primary psychiatric diagnoses in the sample were chronic pain (38%), anxiety or posttraumatic stress disorder (34%), and depression (19%). Diagnoses applied criteria according to the Diagnostic and Statistical Manual of Mental Disorders (4th ed.; DSM–IV; American Psychiatric Association, 1994) or the Diagnostic and Statistical Manual of Mental Disorders (4th ed.; text rev.; American Psychiatric Association, 2000) and are those rendered by the third author at the time of assessment based on all available data (including a detailed clinical interview of the evaluatee, psychological test results, and documentation that accompanied the referral). Musculoskeletal and orthopedic injuries were the most common physical conditions leading to or related to the referrals. The primary sites of pain at time of assessment were the head, face, or mouth (21%); lower back (20%); neck (16%); shoulders/upper extremities (15%); and lower extremities (8%).

Assessments with the CARB, WMT, and TOMM all produce multiple results from which one might make inferences about an evaluatee’s attempt to feign cognitive impairment. For this study, we evaluated the accuracy of three measures: a total CARB score equaling the total percentage correct across all three blocks of the instrument; a total WMT score equaling the average of the immediate recall (IR), delayed recall (DR), and consistency scores; and Trial 2 of the TOMM. We used the results of Trial 2 of the TOMM because Tombaugh (who developed the TOMM) and most published studies have used this score as the primary indicator of feigned memory impairment (see, e.g., Frederick & Bowden, 2009; Tombaugh, 2003).

We took a Bayesian approach to estimating ROC accuracy parameters from our study data. ROC graphs describe the accuracy of detection methods by plotting a method’s true positive rate ($tpr$, equal to test sensitivity) as a function of the method’s false positive rate ($fpr$, equal to $1 - test$ specificity) as the diagnostic threshold or cutoff is moved along the full range of possible test outcomes. “Trapezoidal” ROC graphs result from plotting ($fpr$, $tpr$) pairs associated with the cutoffs between data categories. In such cases, one can estimate a summary index of accuracy, the area under the ROC curve (AUC), by summing the trapezoid-shaped areas under each segment of the ROC graph. One can also construct a ROC graph with a smoothed line if one makes certain assumptions about the distributions underlying the empirical test results. Under the commonly used “binormal” assumption of ROC curve fitting, one assumes that the data arise from two Gaussian (bell-shaped) distributions with different means and variances. The binormal assumption allows one to summarize the entire ROC curve through the linear relationship $z(tpr) = A + Bz(fpr)$, where $z(*)$ is the inverse of the cumulative normal distribution function (a $z$ score), $A$ is the distance between means of the two data distributions, and $B$ is the ratio of the distributions’ standard deviations. The area under a binormal ROC curve is

$$AUC = \Phi \left( \frac{A}{\sqrt{1 + B^2}} \right),$$

where $\Phi(*)$ is the cumulative normal distribution function.

We obtained estimates of ($fpr$, $tpr$) pairs for trapezoidal ROC graphs using methods described by Mossman et al. (2010), and we estimated parameters of binormal ROC curves using methods described by Müller et al. (2009). In contrast to maximum-likelihood estimation (MLE), which provides point estimates of the parameter values most likely to have generated the observed data, Bayesian estimation summarizes knowledge about unknown parameters using “posterior” distributions representing the probability that a parameter has a particular value, given the observed data. Bayes’ Rule states that the posterior probability of a parameter’s value is proportional to the likelihood of observing the data given that parameter value, multiplied by a “prior” probability of the parameter’s value. When a prior is “non-informative,” Bayesian and MLE methods often lead to similar numerical results (Carlin & Louis, 2000), but they have important theoretical differences. Like other conventional (“frequentist”) statistical methods, MLE assumes that the parameter of interest (say, $\theta$) is a fixed but unknown constant; thus, the proper interpretation of a conventional 95% confidence interval is that, because 95% of random intervals calculated in a particular way will contain $\theta$, one has 95% confidence that the calculated interval does. By contrast, Bayesian estimation—which articulates inferences about $\theta$ based on posterior distributions that are developed from prior assumptions and observed data—summarizes beliefs about the values for $\theta$ that one would credibly hold, given the data. Credible intervals are often more meaningful than properly interpreted confidence intervals because they come from Bayesian inferences and allow direct probability statements about accuracy parameters (e.g., “the probability is 95% that parameter $\theta$ for instrument $j$ is between $a$ and $b$” or “the probability that this subject feigned cognitive impairment is more than 90%”).

In our study, both parametric and nonparametric accuracy analyses involved preparing code for the free software program WinBUGS (Lunn, Spiegelhalter, Thomas, & Best, 2009) to find posterior distributions from which we could make Bayesian inferences about the values of accuracy parameters. Markov chain Monte Carlo (MCMC) methods (Gelfand & Smith, 1990; Geman & Geman, 1984; Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953) permit inferences on posterior distributions for which (as was true in our work) lack of analytic methods would make Bayes’ Rule intractable, because under mild regularity conditions, a Markov chain converges to a unique invariant or target distribution. One can therefore use MCMC methods for Bayesian analysis by constructing the transition kernel so that the target distribution of the resulting Markov chain will be the joint posterior distribution of interest. After discarding input from initial “burn-in” iterations, one can use the remaining draws to make inferences about model parameters. A full description of models and procedures appears in the online Supplemental Materials, accompanied by examples of the WinBUGS code written for our analyses.

Given the illustrative aims of this study, we implemented two approaches toward the true status of subjects, which we term agnostic and partial truth. Under the agnostic approach, one assumes that the only information available to an evaluator is the subjects’ SVT scores and asks WinBUGS to seek the best values for model parameters based on those SVT values alone. The partial truth approach recognizes that for a subset of cases, features of the referral context combined with SVT results yield incontrovertible evidence of malingering or honest responding, and this information can be included when WinBUGS estimates a model’s parameters. For example, an evauluee who travels to an evaluation on his
own (which requires relatively intact cognitive functioning) but who scores below chance on the TOMM must be responding invalidly during testing. Similarly, one can be confident that an evaluatee who is not involved in litigation, has been referred by his physician, and who scores in the total passing range on all SVTs is responding honestly. The agnostic approach has the advantage of avoiding any potential argument about whether a particular subject is "really, positively" responding genuinely or not—the data simply speak for themselves. But if a subset of subjects can be classified into one group or the other with virtual certainty, the partial truth approach has the advantage of incorporating this knowledge into one's accuracy determinations.

For our partial truth analyses, we classified as "definitely responding invalidly" 67 subjects who scored below chance on one or more SVTs (i.e., less than 37.5% on the DR or IR sections of the WMT, less than 45% on the total CARB score, or less than 36% correct on Trial 2 of the TOMM) or who scored in the random responding range on two SVTs (37.5%–62.5% on the DR or IR sections of the WMT, 45%–55% on the CARB, or 36%–64% on the TOMM). We reasoned that such scores could not reflect valid responding, given that our data were accumulated in an outpatient setting to which evaluatees traveled independently, and the evaluatees did not have the kinds of impairments (e.g., dementia or severe cognitive disability) that could lead to genuine no-better-than-chance responding. We classified as "definitely not malingering" 79 subjects who passed all SVTs (i.e., by scoring at least 82.5% correct on all three of the WMT primary effort measures, at least 90% correct on the total CARB, and at least 90% correct on Trial 2 of the TOMM) and who were neither referred by the workers' compensation board nor involved in litigation. Our reasoning was that these subjects had no external motive to appear impaired and the evaluatees maintained by other considerations (e.g., factitious disorders). This "definitely not malingering" group consisted mainly of privately referred subjects who were paying their own way for evaluation.

Results

Our 1,301 subjects achieved the following results (calculated as error percentages): CARB $M \pm SD = 5.28 \pm 11.53$ (range = 0–84.7); WMT $M \pm SD = 11.70 \pm 13.52$ (range = 0–65); and TOMM $M \pm SD = 3.70 \pm 10.44$ (range = 0–80). Our Bayesian analyses produced results summarized in Tables 1 and 2 and Figures 1 and 2.

Table 1 shows the Bayesian estimates of AUC (ROC areas) for all three SVTs under all six sets of model and data assumptions. Here, AUC equals the probability that the score (calculated as a monotonic function of error rate) obtained by a randomly chosen invalidly responding subject would be higher than the score obtained by a randomly chosen honestly responding subject. An AUC of 1.0 would imply perfect sorting, and an AUC of 0.5 would imply no-better-than-chance discrimination between invalidly responding and validly responding subjects. The results imply that all the SVTs do far better than chance at feigned cognitive impairment, but the WMT appears superior to the CARB and TOMM. Our WinBUGS code included a feature that allowed direct monitoring of the differences among AUCs for the three SVTs. In all six analyses, the 95% credible intervals for the AUC differences between WMT and the other SVTs did not include zero, implying a high probability that the accuracy of WMT exceeds the accuracies of other SVTs.

Looking at Table 1, one might wonder whether one of the models is “better,” given the data. One way to compare results involves the Akaike (1974) information criterion, calculated as AIC = $-2\ln L + 2p$, where $\ln L$ is the log-likelihood function (two times, which is the WinBUGS mean deviance shown in Table 1; see Spiegelhalter et al., 2002) evaluated at each MCMC iteration given the current sampled values of the parameters and $p$ is the number of model parameters. In calculating the AIC, the superior fit (measured by the $-2\ln L$ term) that one expects from including additional parameters is offset by a penalty (the 2p term). Minimum AIC thus can serve as a basis for choosing, from among several models with different numbers of parameters, a model that the data best support. Albert (2007) noted that one can estimate cutoff points for trapezoidal ROC graphs under assumptions of conditional independence (CI) and conditional dependence (CD). The CI model assumes that measures (in this case SVTs) are independent in the scores they produce for validly and invalidly responding subjects—which is implausible, because all three SVTs utilize the same basic premises for detecting response bias. The CD model lets one factor into accuracy measures any peculiarities of the malingering and nonmalingering subject pools, such as whether either pool contained cases that were especially easy or hard. In Table 1, AIC values are substantially smaller for CD.
models than for CI models. This is consistent with expectations: Subjects who wish to feign cognitive impairment might well do so on multiple measures and to similar degrees on those measures.

The negative deviance values for the binormal models make direct comparison with the trapezoidal models difficult. Negative deviances can arise when the posterior distribution for a parameter is very nonnormal, so that the posterior mean is a poor summary statistic for the parameter (Spiegelhalter, Best, Carlin, & Van der Linde, 2002), a situation that arises here for the means of the nonmalingering populations (due to a ceiling effect for all three SVTs). Another basis for evaluating the binormal models emerges from Figure 1, which contains plots of estimated binormal ROC curves for all three SVTs under the agnostic assumption. The ROC curves for all three SVTs have visible “hooks” in the upper right corner of the ROC square and (especially for the TOMM) have portions in which the curve falls well below the “no information” diagonal. The curves’ shapes imply that some malingering individuals make fewer errors than nonmalingeringers do, which is not true: Many nonmalingeringers make no errors (especially on the TOMM). When we attempted to rectify this by transforming the distributions of SVT scores to normalize their distributions (via Box-Cox and other techniques), WinBUGS either failed to identify a single model or repeatedly identified models in which the TOMM had perfect accuracy and any wrong answer implied malingering—a result we rejected as incorrect. In sum, the “binormal” ROC curve-fitting assumptions (at least as these are realized by Choi et al., 2006, and Müller et al., 2009) might prove valuable for evaluating many data sets, but they were ill suited to our data.

Figure 2 and Table 2 provide detailed results from the agnostic approach under the nonparametric, conditional dependence assumption. The figure and the table show that all three SVTs have operating points that allow for very low false positive rates (false identifications of invalid responding). Often, low fprs are desirable features of tools for assessing impression management, but low fprs also limit the fraction of invalid responders who are identified correctly. For example, an evaluator who believed that the false positive rate should be held at 0.02 and who therefore used decision thresholds of 89%, 82%, and 96% for the CARB, WMT, and TOMM, respectively, would produce invalid response detection (true positive) rates of approximately 35%, 65%, and 49%, respectively.

We note two other interesting features of these results:

First, the WMT appears superior to the other SVTs. We confirmed this by comparing differences between AUCs of the WMT’s and the other two SVTs’ statistical testing and found that the 95% credible intervals for these differences did not include zero. AUCs of the CARB and TOMM did not differ significantly from each other, however.

Second, when evaluating persons similar to members of our subject population (i.e., ambulatory adult evaluees), one might use a stringent cutoff for the TOMM, which would identify a larger fraction of malingeringers without misidentifying many nonmalingering evaluatives. For example, using a cutoff of ≤96% would detect nearly half of the malingeringers, whereas using ≤90% would detect only two out of seven malingeringers.

Another way to interpret these results involves recognizing that the outcomes of SVTs actually provide graded evidence for or against invalid responding—that is, the more incorrect answers, the stronger the evidence for feigning of cognitive deficits. With this in mind, we created Table 3 using the accuracy characteristics listed in Table 2 and the malingering base rate (0.325) obtained with the agnostic, nonparametric CD model (see Table 1) to calculate probabilities of invalid responding associated with scores in specific ranges. One way to think about these results is to ask what fraction of SVT results would let an evaluator be at least 90% sure whether the subject was responding honestly or not. For the CARB, this would occur in fewer than one out of 10 cases (for scores below 85%), and for the TOMM, in just one out of eight cases (for scores less than 96% on Trial 2); neither of these SVTs produces results that support 90% confidence that a subject is...
responding validly. For the WMT, about two out of nine subjects would have scores low enough (81% or less) to indicate a probability of invalid responding above 90%, and more than half the subjects would have scores high enough to let one be confident that they were responding honestly.

### Discussion

Our results support Mossman and colleagues’ (2010) contention that latent class modeling (i.e., *ROC analysis without truth*; Henkelman et al., 1990) might help investigators gauge the accuracy of assessment techniques used in forensic psychology and psychiatry. The present study has shown that results from three symptom validity tests completed by a large number of actual neuropsychological evaluees permitted estimation of those tests’ accuracy parameters despite the absence of an infallible criterion for ascertaining whether the subjects were performing at less than their true ability. Our findings suggest that LCM methods offer a solution to methodological limitations of known-group research designs (imperfect truth criteria and/or exclusion of ambiguous cases) while retaining “external” or “ecological” validity (i.e., the use of results from “real” evaluees who experience all the motivations, contingencies, and stresses of “real” evaluations) that is absent from simulation studies. Yet we ask readers to view our results with skepticism and cautiousness, for several reasons.

First, we have reported findings based on a single data set and a single evaluation context. Though our results appear consistent with those reported in other evaluations of the SVTs we examined (Fox, 2011; Gervais, Rohling, Green, & Ford, 2004; Green, Flaro, Courtney, 2009), our findings are not definitive judgments about the performance of these SVTs. In addition to expectable variation caused by sampling error, results would likely look different if developed using subjects from different treatment or evaluation settings (e.g., psychiatric inpatients or individuals with developmental disabilities).

Second, our findings show that reasonable statistical models for our data can generate somewhat different results. As we noted earlier, the results from the binormal assumption yield problematic and perhaps misleading conclusions about the data distributions.

### Table 3

**SVT Range (Percentage of Correct Responses), Fraction of Subjects With Scores in Those Ranges, and Probability of Invalid Responding P(IR+)**

<table>
<thead>
<tr>
<th>CARB</th>
<th>Range</th>
<th>Fraction</th>
<th>P(IR+)</th>
<th>WMT</th>
<th>Range</th>
<th>Fraction</th>
<th>P(IR+)</th>
<th>TOMM</th>
<th>Range</th>
<th>Fraction</th>
<th>P(IR+)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–59.9</td>
<td>0.010</td>
<td>0.996</td>
<td></td>
<td>0–50</td>
<td>0.007</td>
<td>0.987</td>
<td></td>
<td>0–50</td>
<td>0.004</td>
<td>0.99996</td>
<td></td>
</tr>
<tr>
<td>60–74.9</td>
<td>0.037</td>
<td>0.993</td>
<td></td>
<td>50.1–63</td>
<td>0.035</td>
<td>0.997</td>
<td></td>
<td>52–60</td>
<td>0.006</td>
<td>0.99988</td>
<td></td>
</tr>
<tr>
<td>75–84.9</td>
<td>0.050</td>
<td>0.955</td>
<td></td>
<td>63.1–73</td>
<td>0.081</td>
<td>0.974</td>
<td></td>
<td>62–70</td>
<td>0.009</td>
<td>0.9997</td>
<td></td>
</tr>
<tr>
<td>85–88.9</td>
<td>0.040</td>
<td>0.685</td>
<td></td>
<td>73.1–81</td>
<td>0.095</td>
<td>0.925</td>
<td></td>
<td>72–78</td>
<td>0.021</td>
<td>0.991</td>
<td></td>
</tr>
<tr>
<td>90–93.9</td>
<td>0.038</td>
<td>0.573</td>
<td></td>
<td>81.1–88</td>
<td>0.104</td>
<td>0.504</td>
<td></td>
<td>80–84</td>
<td>0.024</td>
<td>0.995</td>
<td></td>
</tr>
<tr>
<td>94–95.9</td>
<td>0.098</td>
<td>0.455</td>
<td></td>
<td>88.1–91</td>
<td>0.095</td>
<td>0.375</td>
<td></td>
<td>86–90</td>
<td>0.030</td>
<td>0.985</td>
<td></td>
</tr>
<tr>
<td>98–97.9</td>
<td>0.068</td>
<td>0.340</td>
<td></td>
<td>91.1–93</td>
<td>0.049</td>
<td>0.198</td>
<td></td>
<td>92–94</td>
<td>0.034</td>
<td>0.941</td>
<td></td>
</tr>
<tr>
<td>98–99.9</td>
<td>0.088</td>
<td>0.153</td>
<td></td>
<td>93.1–95</td>
<td>0.144</td>
<td>0.0226</td>
<td></td>
<td>96</td>
<td>0.037</td>
<td>0.894</td>
<td></td>
</tr>
<tr>
<td>99–99.9</td>
<td>0.002</td>
<td>0.518</td>
<td></td>
<td>95.1–98</td>
<td>0.127</td>
<td>0.0249</td>
<td></td>
<td>98</td>
<td>0.086</td>
<td>0.388</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>0.568</td>
<td>0.173</td>
<td></td>
<td>98.1–100</td>
<td>0.263</td>
<td>0.0466</td>
<td></td>
<td>100</td>
<td>0.749</td>
<td>0.178</td>
<td></td>
</tr>
</tbody>
</table>

*Note.* Data are based on the findings reported in Table 2 and a base rate of 0.325. SVT = symptom validity test; CARB = Computerized Assessment of Response Bias; WMT = Word Memory Test; TOMM = Test of Memory Malingering.
The nonparametric conditional dependence results from agnostic and partial truth assumptions show similar patterns and tend to confirm impressions from other publications about the comparative accuracy of these three SVTs. Yet when the partial truth assumption is applied, all three SVTs appear more accurate and the rate of invalid responding declines noticeably. Also, as Uebesax (1988) noted, LCM provides upper bounds for accuracy under certain conditions. LCM chooses underlying classes that minimize error rates defined within the model, but these error-minimizing, empirically generated latent classes can differ from the true classes when probabilities of the empirical classes depend on covariates. Knowing whether this actually has occurred is difficult to ascertain (Spencer, 2011), but we must acknowledge its potential presence.

A third point relates to potential limitations caused by unintentionally mistaking reliability of test results for validity of interpretations based on test results. We assumed that the results from SVTs reflect valid (if implicit) conceptions about responses to testing, including the notion that individuals inclined to attempt impression management would likely do so on more than one SVT and on other cognitive evaluation measures besides SVTs. The SVTs we examined employ similar detection strategies (“performance worse than that given by persons with genuine impairment implies feigning”), and they can be scored reliably. A problem with our statistical methods is that irrelevant assessment methods that yield reliable test scores can appear accurate, though they really have no-better-than-chance accuracy. Thus, if SVTs elicited reliable patterns of responses that were irrelevant to invalid responding, they would look more accurate than they really were. This could happen if, for example, most malingering evaluators knew about these tests and “played it straight” on them but intentionally gave inaccurate responses on other measures in the test battery. Our response to this criticism is that we based our findings on SVTs commonly used by neuropsychologists, but we recognize this limitation in our approach.

A fourth (and related) point involves the absence of any information external to the SVT data. That is, our statistical approach involved making judgments about invalid responding from SVT data alone, something mental health professionals should not and ordinarily would not need to do. In most evaluation settings, psychologists and psychiatrists obtain ample (if unsystematic) data from their clinical encounters with evaluees or from collateral sources (e.g., past treatment records or family members) that inform their final judgments about how honestly evaluees are portraying their problems and capacities. We are not suggesting that mental health professionals ignore such data. We do think, however, that mental health professionals should scrutinize such data with the same techniques that we used in our study.

For example, in addition to obtaining information from multiple SVTs, future investigators could ask evaluators to provide graded judgments (on a Likert scale) about the likelihood of invalid responding based only on clinical and collateral data. The accuracy of the evaluators’ judgments could then be treated as an additional “effort measure” and be evaluated simultaneously with SVTs using the same statistical approaches we used here. In other words, rather than treat a clinician’s opinion or any other criterion as infallible, future studies of symptom validity assessment should recognize that clinical opinion is just one of many sources of information about evaluees’ true status, the accuracy of which deserves scrutiny. Also, future studies might incorporate more complex models of codependence within SVT results and between SVT results and clinical data (see Menten, Boelaert, & Lesaffre, 2008, for a discussion with examples). Moreover, recent advances in clinical neuropsychology have offered structured approaches to incorporate various sources of clinical data (including SVTs) in the systematic detection of neurocognitive and pain malingering (Bianchini, Greve, & Glynn, 2005; Slick, Sherman, & Iverson, 1999).

Our approach to our data presupposed, as many studies of impression management do, that it was valid to assume our subjects either gave valid responses or did not and to assess the accuracy of SVTs based on this dichotomy. Of course, real-world evaluees include individuals who try to exaggerate or feign cognitive impairment to varying degrees, in various styles, and with various conceptions about real disability. Also, as Frederick and Bowden (2009) pointed out, some evaluees may produce invalid responses because of disengagement from the testing process (low effort) rather than a (conscious) intention to respond incorrectly. This implies that malingering is a dimensional construct (one that admits of degrees) and that invalid responding has more potential causes than simple intention to mislead evaluators (see Rogers, 2008, for a discussion of various motivations for malingering). Our model does not capture these features, nor does it provide any insight about what invalidly responding evaluees are thinking.

We assumed that invalid versus genuine responding was a both a logical and empirically legitimate dichotomy, but evaluating such assumptions is a typical application of LCM. The boundary between being honest and other forms of responding is fuzzy. Moreover, accumulating research utilizing taxometric analyses (e.g., Walters, Berry, Lanyon, & Murphy, 2009; Walters, Berry, Rogers, Payne, & Granacher, 2009) has suggested that malingering does indeed exist as a dimensional construct. Forensic psychiatry and psychology might benefit from empirically based explorations of response “taxa” and from comparing those taxa to theoretical taxonomies that have been proposed for malingering (e.g., Rogers, 2008) or that are assumed to exist for purposes of test construction (e.g., Frederick & Crosby, 2000). Yet it is not known whether an empirical taxonomy would track what is thought to be an intrinsically meaningful, logical boundary between compliant/honest responding and other response styles. Whatever an empirically based malingering taxonomy might reveal, it still seems reasonable to evaluate the accuracy of measures by assuming a distinction between invalid and honest responding, a distinction that uses the commonsense but reasonable assumption that ultimately, evaluees respond either in ways that reflect their true ability or otherwise.

Under the presumption that these two response styles are mutually exclusive, we aimed our study at quantifying and characterizing how well three often-used SVTs could distinguish them. The valid-responding-or-not dichotomy can accommodate the notion that degrees of neurocognitive impression management exist and manifest themselves in greater and lesser error rates on SVTs. Because all three SVTs had low sensitivity in the score ranges usually used to identify malingers, our study suggests that evaluees who exaggerate impairment only slightly stand a good chance of escaping detection with these instruments.

**Conclusion**

The statistical approach we describe in the article is applicable to evaluating many types of measures used in psycholegal determina-
tions. We suspect that among the readers of this article are several colleagues who have created or could create data relevant to other forensic problems for which no diagnostic gold standards exist. We encourage other investigators to use and improve upon our work to address a broad variety of questions concerning the accuracy of assessment methods in forensic psychology and psychiatry.

References


