Setting empirical cut-offs on psychometric indicators of negative response bias: a methodological commentary with recommendations

Kevin W. Greve a, b, *, Kevin J. Bianchini a, b

a Department of Psychology, University of New Orleans, New Orleans, LA 70147, USA
b Jefferson Neurobehavioral Group, Metairie, LA, USA

Accepted 5 August 2003

Abstract

Malingering in neuropsychological assessment has been the subject of intense research for more than a decade and the detection methods arising from this work are diverse and sophisticated. However, the empirical findings are often presented in ways that limit the clinical utility of these techniques and may threaten their admissibility into legal proceedings. The purpose of this paper is to outline an approach for setting cut-offs on techniques designed to identify the presence of negative response bias. The use of this approach will result in the explicit specification of the error rate(s) of a given technique which can easily be applied by clinicians in the course of their practice and be admissible in court.

Keywords: Malingering; Assessment; Brain injury; Response bias

Neuropsychologists are increasingly called upon to evaluate persons for whom there is substantial incentive to appear impaired. Thus, the detection of negative response bias and the diagnosis of malingering are important questions that must be addressed (Binder & Rohling, 1993). Moreover, when the answers to these questions take the form of expert testimony they must meet certain standards of admissibility. In Federal Court admissibility of expert testimony is governed by the Federal Rules of Evidence (FRE 702) as articulated in the case of Daubert v. Merrell Dow Pharmaceuticals (1993) and other cases. Most state jurisdictions have adopted rules like FRE 702. In the remaining jurisdictions the “Frye” standard generally applies.

*Corresponding author. Tel.: +1-504-280-6185; fax: +1-504-280-6049.
E-mail address: kgreve@uno.edu (K.W. Greve).

0887-6177/– see front matter © 2003 National Academy of Neuropsychology.
The “Daubert” standard requires that expert evidence be both relevant and reliable. “Reliability” is related to the science upon which the evidence or opinion is based (Berger, 2000). In determining whether a technique or theory is “scientific” the judge, as gatekeeper, must consider: (1) whether the theory or technique can or has been tested; (2) evidence of peer review or publication; (3) known or potential error rate and standards; and, (4) general acceptance of the methodology (the old “Frye” standard). The scientific literature on the detection of response bias and malingering with neuropsychological techniques is extensive. However, it is often difficult to derive “error rates” from individual studies and methodological problems sometimes raise questions about the relevance of research results (see Bianchini, Mathias, & Greve, 2001a, for a discussion of these issues related to the use of Symptom Validity Tests). Moreover, even in studies where the accuracy of a technique is clearly described, the data are often not presented in clinically useful manner.

The purpose of this paper is to outline an approach for setting cut-offs on techniques designed to identify the presence of negative response bias. The use of this approach will result in explicit specification of the error rate(s) of a given technique which can easily be applied by clinicians in the course of their practice and be admissible in court.

1. Defining “error rates”

The relevant indices of classification accuracy (and therefore error rate, the reciprocal of accuracy) are Sensitivity, Specificity, and Predictive Power (Hennekens & Buring, 1987; see also Gouvier, Hayes, and Smiroldo, 1998, for a discussion of these factors). Sensitivity is the true positive (Hit) rate for a test (number of persons with a condition who had a positive test result divided by all persons with the condition). Specificity is the true negative rate (number of persons without a condition who had a negative test result divided by all persons without the condition). Poor Sensitivity means a given cut-off produces a large number of False negative errors, meaning that some true malingerers go undetected. In contrast, poor Specificity results in higher numbers of False positive errors which means that a legitimate case may be classified as invalid. Sensitivity and Specificity are directly dependent on the classification scheme (cut-off) employed with a given diagnostic technique and are independent of base-rates. Predictive Power is an index of the confidence one can have that an individual test result is accurate (Hennekens & Buring, 1987); it is dependent on both accuracy of the test itself and the base rate of the target condition in the population of interest.

Positive Predictive Power (+PP, measured: True positives divided by the sum of True positives and False positives; see Millis & Volinsky, 2001, for an alternative formula for the calculation of +PP or “posttest odds”) indicates the probability that someone has the condition given a positive test result. +PP is related to Specificity in that, regardless of Sensitivity, if persons without the condition never have a positive finding, one can be very confident that a positive finding reflects the presence of that condition. However, calculation of +PP does require knowing Sensitivity for a given score and everything else being equal, higher Sensitivity will result in higher +PP. Negative Predictive Power (−PP, measured: True negatives divided by the sum of True negatives and False positives) indicates the probability that a patient does not have the condition given a negative test result. −PP is related to Sensitivity in that, regardless
of Specificity, if a test never misses an individual with the condition, then one can be very confident that a negative finding indicates the absence of the condition. Predictive Power in general is related to base-rates in that the rarer the condition in the population of interest, the more likely a positive test result is to be a false positive.

In short, Sensitivity and Specificity are indices of overall classification success or accuracy while Predictive Value is an index of confidence one can have that the classification of an individual is correct.

2. The importance of Specificity versus Sensitivity

The Sensitivity of all individual indicators of negative response bias will always be less than perfect (false negative errors will occur) if one wishes to guard against excessive false positive errors. False negative errors using individual indicators for the detection of response bias cannot be eliminated for three reasons.

1. The degree of transparency and apparent difficulty of the test can significantly affect Sensitivity. Transparency refers to how obvious the means for detecting malingering is. For example, value and effectiveness of forced-choice Symptom Validity Tests (SVTs) rests on the fact that the probability of a correct answer is known (usually .50); the more obvious this is, the less likely a person is to be fooled by the task. SVTs are also designed to look much harder than is actually the case (apparent difficulty). Almost all variations of SVTs are designed to enhance Sensitivity by reducing transparency and increasing apparent difficulty (Bianchini et al., 2001a).

2. Poor Sensitivity may also be related to an individual’s strategic approach to malingering. For example, Greve, Bianchini, Mathias, Houston, and Crouch (2002) demonstrated three different approaches used by malingerers in taking the Wisconsin Card Sorting Test (WCST; Heaton, Chelune, Talley, Kay, & Curtiss, 1993), one of which included performing normally. Similarly, Greve, Bianchini, Mathias, Houston, and Crouch (2003) argued on theoretical and empirical grounds that the Mittenberg formula (Mittenberg, Theroux-Fichera, Zielinski, & Heilbronner, 1995; Mittenberg et al., 2001) is sensitive to only a subset of malingered performances. Sensitivity can be enhanced by the use of multiple detection techniques that are sensitive to different approaches to malingering (Greve et al., 2002). This may be the most effective means of improving Sensitivity in the long run.

3. Sensitivity can be reduced by attorney coaching. Youngjohn (1995) published a case of documented coaching. Allen and Green (2001) documented the declining Sensitivity of one commonly used SVT over a period of six years and speculated that it may have been due to its increased familiarity to plaintiff’s attorneys.

Thus, because Sensitivity of individual indicators will likely never be perfect, attempts to set cut-offs that detect all true malingerers will inevitably result in an unacceptably high number of false positive errors. Sensitivity might be improved through the development of multiple indicator sets sensitive to different malingering strategies. However, attempts to establish cut-offs that maximize either Sensitivity or overall classification accuracy will inevitably compromise Specificity and lessen the value of the indicator.
In contrast, it is possible to set cut-offs on a test or indicator beyond which no patients with legitimate performance fall. That is, cut-offs with perfect Specificity can be determined though, of course, cut-offs set to enhance Specificity will inevitably reduce Sensitivity. However, many consider false negative errors more acceptable than false positives. For example, Wasyliw and Golden (1985) consider “malingering” as one of the most pejorative terms that can be applied to an individual. Practically, the primary distinction between false positive and false negative errors is who bears the burden of the error. In the case of false positives, it is the individual patient and their dependents who are affected, while the price of false negatives is distributed across a society as a whole. Specifically, the false diagnosis of malingering can have important financial, occupational, and personal (emotional, interpersonal) consequences for the patient. On the other hand, the failure to recognize malingering may reduce the availability of services for legitimate claims (Franzen, Iverson, & McCracken, 1990), may lead to avoidance of prosecution for criminal action (Franzen et al., 1990), and may damage the credibility of the individual clinician and the profession (Haines & Norris, 1995). Of course, a tendency toward false positive errors, like false negative errors, is also harmful to the clinician and profession. When an indicator has perfect Specificity, whatever its Sensitivity and the base rate, one can be extremely confident (+PP = 1.00) that a positive result reflects response bias and may indicate the presence of malingering.¹

3. Setting cut-offs

The main conceptual point of this paper is that given the importance of Specificity and the limitations of Sensitivity, empirical cut-offs for indicators of response bias should be based on Specificity while letting the Sensitivity chips fall where they may. The primary practical point is that any paper or manual describing a technique designed to detect malingering must explicitly report the Sensitivity and Specificity (error rate) of the technique (regardless of the method by which the cut-off was established) or the technique will be less clinically useful and more vulnerable to Daubert challenge.

If only a single cut-off is established it should be the one associated with perfect Specificity (a false positive rate of .00). Ideally, a range of cut-offs associated with empirically defined Specificity levels (e.g., .80, .90, .95, .99, and 1.00) should be developed and the associated Sensitivity levels subsequently reported. This will allow the development of a continuum of Predictive Power below 100 percent confidence. Predictive Power (especially +PP) associated with these cut-offs for a range of likely base-rates of malingering (e.g., .10-.50 at .10 increments) should then follow. If only one level of PP is presented, we recommend that it be calculated using a base-rate of .30. This value is consistent with the findings of Mittenberg, Patton, Canyock, and Condit (2002). Other estimates have ranged as high as 40 percent (see Mittenberg et al., 2002, for discussion). However, clinicians can easily calculate PP based on their local base-rates if Sensitivity and Specificity data are provided.

¹Note that the diagnosis of malingering is a clinical decision that takes into account a range of clinical, psychometric, and other information and should follow guidelines like those provided by Slick, Sherman, and Iverson (1999) in which indicators of response bias are an important, but not sole, criterion for the diagnosis.
3.1. Cut-off precision and validity

It is very easy to mechanistically set cut-offs following the above guidelines. However, the ultimate value of the decision rule is dependent on a number of factors. These factors must be addressed by both the researchers developing the cut-offs and evaluated by the clinician who intends to use a given indicator. Specifically, the precision and validity of empirically established cut-points are dependent on at least three factors:

3.2. Score distribution

It will not always be possible to establish cut-offs that result in the exact Specificity/False positive rates recommended above. Scores that are more continuous and less heavily skewed will be more precise. Discriminant function scores, such as those developed by Millis (Millis, Putnam, Adams, & Ricker, 1995) and Mittenberg (Mittenberg et al., 1995, 2001) are often nearly continuous with a broad distribution, so precise cut-offs should be relatively easy to develop. However, many indicators do not have a broad range, are heavily skewed, and/or are very rare. For example, the overall range of the Reliable Digit Span test (RDS; Greiffenstein, Baker, & Gola, 1994) is 0 to 17 with the relevant portion ranging from 8 down to 0 (Mathias, Greve, Bianchini, Houston, & Crouch, 2002). On many SVTs even patients with documented brain damage and cognitive deficits make few errors so the distribution is skewed. For example, on the Test of Memory Malingering (Tombaugh, 1996) even dementia patients were over 90 percent accurate on average. Finally, while the number of Unique Responses on the WCST can theoretically range from 0 to 128, in reality most patients have none (Greve et al., 2002).

3.3. Sample size

The ability to precisely set a range of cut-offs using the method described above is also dependent on the size of the nonmalingering control group. For example, a control group with only 20 persons would not allow cut-offs at the .99 level because each individual represents five percent of the sample. In contrast, in a control group in which \( N = 500 \), each observation represents .2 percent of the sample. Moreover, the larger and more diverse the nonmalingering control group, the more likely the distribution will reflect the full range of possible scores. Cut-offs based on small sample sizes may also be less stable, so confidence in the classification result would necessarily be lower. Finally, a large sample will allow some patients to be held back for cross-validation of the new cut-offs. Inclusion of a cross validation data in the same paper reporting the cut-offs would allow the cut-offs to be used clinically without waiting for an independent cross-validation to be published. Of course large samples are always desirable but often hard to find. The use of a formal multi-site format for the study of malingering detection techniques is one way to overcome this difficulty.

3.4. Sample composition

At minimum there must be two criterion samples in any complete examination of a malingering test: a suspected malingering sample (index) and a nonmalingering (control) group. Sen-
Sensitivity and Specificity are a function of the performance of the individuals in these two groups. In order to accurately determine Sensitivity one must be very confident that only persons who are malingering have been included in the sample (thus, the weakness of the “differential prevalence” approach). There is no perfect test of malingering and the vast majority of malingerers will not admit to such or be caught by surveillance. Therefore, the use of multiple external criteria (e.g., those of Slick et al., 1999) for assignment to the malingering group is essential for accurately determining Sensitivity (Bianchini et al., 2001a) as has been clearly illustrated in Bianchini, Mathias, Greve, Houston, and Crouch (2001b). Moreover, when studying the accuracy of a psychometric indicator of response bias, one must be cautious when using other psychometric indicators of response bias as a criterion for group assignment. Finally, while simulators (e.g., college students asked to fake cognitive deficits) are often used in the preliminary development of malingering detection techniques, “sensitivity” data derived from simulators cannot be reliably applied in clinical settings and will likely be vulnerable to Daubert challenge.

In evaluating Specificity, the nonmalingering control group must, at least, be comprised of persons who are similar to the index sample in terms of etiology as well as the usual demographic characteristics. If possible, groups of diverse clinical patients and healthy control subjects would be included to help establish the range of possible nonmalingering scores. Under ideal circumstances the control subjects will be persons without identifiable external incentive to perform poorly. These people are, by definition, not malingering. This criterion does not rule out Factitious Disorder, however, so control subjects with extreme scores should be closely examined to insure that their performance does not reflect either a factitious symptom presentation or the presence of a previously unidentified external incentive before they are included in the final data set. When studying malingering in some populations (e.g., neurotoxic exposure, pain-related disability) etiologically similar patients without external incentive may not be available. In those cases, malingering must be ruled out clinically. The Slick criteria can serve as an effective starting point but are not sufficient. In such cases, anyone who is suspicious should be excluded from the control group. The major point here is that in developing empirical cut-offs, the purity of the control and malingering groups is critical to the validity of the cut-offs.

Table 1 presents a concrete example of the approach recommended in this paper. This table was adapted from Larrabee (2003) for the MMPI-2 Fake Bad Scale (FBS; Lees-Haley, English, & Glenn, 1991). The criterion group was composed of 26 litigants claiming brain injury of varying etiologies who had performed significantly below chance on at least on part of the Portland Digit Recognition Test (PDRT; Binder, 1993). This finding contributed to a diagnosis of Definite Malingered Neurocognitive Deficit (MND; Slick et al., 1999). The contrast group was composed of 29 persons who had sustained well-documented moderate to severe brain trauma who were not malingering. Table 1 shows Sensitivity, Specificity, and +PP (using a malingering base-rate of .30). As can be seen, FBS scores of 30 or higher were associated with a Specificity of 1.00 and thus a +PP of 1.00. Thus, scores of 31 or higher are exclusively associated with malingering. Even scores as low as 26 were associated a False positive error rate of less than five percent. This table indicates that the probability that a mild TBI patient with a score of 26 is malingering is about 97 percent. As noted previously, this single finding is not sufficient for a diagnosis of malingering. However, by using this table and tables like it, the clinician can easily determine how much confidence he/she can place in a given finding.
Table 1
Sensitivity and Specificity for a range of scores on the Fake Bad Scale

<table>
<thead>
<tr>
<th>Cut-off score</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>+PP</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>.808</td>
<td>.862</td>
<td>.72</td>
</tr>
<tr>
<td>22</td>
<td>.808</td>
<td>.862</td>
<td>.72</td>
</tr>
<tr>
<td>23</td>
<td>.692</td>
<td>.862</td>
<td>.68</td>
</tr>
<tr>
<td>24</td>
<td>.615</td>
<td>.862</td>
<td>.66</td>
</tr>
<tr>
<td>25</td>
<td>.615</td>
<td>.931</td>
<td>.79</td>
</tr>
<tr>
<td>26</td>
<td>.538</td>
<td>.966</td>
<td>.87</td>
</tr>
<tr>
<td>27</td>
<td>.500</td>
<td>.966</td>
<td>.86</td>
</tr>
<tr>
<td>28</td>
<td>.500</td>
<td>.966</td>
<td>.86</td>
</tr>
<tr>
<td>29</td>
<td>.423</td>
<td>.966</td>
<td>.84</td>
</tr>
<tr>
<td>30</td>
<td>.385</td>
<td>.966</td>
<td>.83</td>
</tr>
<tr>
<td>31</td>
<td>.308</td>
<td>1.000</td>
<td>1.00</td>
</tr>
<tr>
<td>32</td>
<td>.192</td>
<td>1.000</td>
<td>1.00</td>
</tr>
<tr>
<td>33</td>
<td>.115</td>
<td>1.000</td>
<td>1.00</td>
</tr>
<tr>
<td>34</td>
<td>.115</td>
<td>1.000</td>
<td>1.00</td>
</tr>
<tr>
<td>35</td>
<td>.038</td>
<td>1.000</td>
<td>1.00</td>
</tr>
<tr>
<td>36</td>
<td>.000</td>
<td>1.000</td>
<td>1.00</td>
</tr>
</tbody>
</table>

This table is adapted from Larrabee (2003). +PP: positive predictive value using a base-rate of .30.

In short, a wealth of clinically relevant information regarding the meaning of a positive or negative result on a measure of response bias and the necessary data to address questions related to admissibility can be provided in one relatively simple table.

4. Conclusions

This paper has attempted to do three things. The first was to emphasize the practical point made previously by Bianchini et al. (2001a) that the classification accuracy of all tests, scores, or other indicators designed to detect malingering should be presented explicitly in the form of Sensitivity, Specificity, and Predictive Power. Not only has this often not been done in the past, but many published articles and test manuals continue to be weak in this respect. Second, this paper attempted to make the conceptual point that in the detection of malingering in clinical settings, emphasis should be placed on maximizing Specificity rather than Sensitivity. If Specificity is known, then the clinician can conclude, with known confidence, whether negative response bias is present. Finally, we have attempted to offer practical, scientifically rigorous guidelines for developing valid empirical decision rules. Ultimately, the test results and the expert opinions derived from them are only as good as the method used to establish and test the cut-offs. The detection of malingered neurocognitive deficits using neuropsychological techniques is sophisticated and advanced relative to other areas in which malingering is a serious problem (e.g., pain-related disability). We hope the ideas and recommendations offered here are of value to other researchers in the on-going development of malingering detection techniques.
References


