Determination of the smoking gun of intent: significance testing of forced choice results in social security claimants

Laurence M. Binder & Michael D. Chafetz

To cite this article: Laurence M. Binder & Michael D. Chafetz (2017): Determination of the smoking gun of intent: significance testing of forced choice results in social security claimants, The Clinical Neuropsychologist, DOI: 10.1080/13854046.2017.1337931

To link to this article: http://dx.doi.org/10.1080/13854046.2017.1337931

Published online: 15 Jun 2017.
Determination of the smoking gun of intent: significance testing of forced choice results in social security claimants

Laurence M. Binder\textsuperscript{a} and Michael D. Chafetz\textsuperscript{b}

\textsuperscript{a}Independent Practice, Beaverton, OR, USA; \textsuperscript{b}Independent Practice, New Orleans, LA, USA

\textbf{ABSTRACT}

\textbf{Objective}: Significantly below-chance findings on forced choice tests have been described as revealing “the smoking gun of intent” that proved malingering. The issues of probability levels, one-tailed vs. two-tailed tests, and the combining of PVT scores on significantly below-chance findings were addressed in a previous study, with a recommendation of a probability level of .20 to test the significance of below-chance results. The purpose of the present study was to determine the rate of below-chance findings in a Social Security Disability claimant sample using the previous recommendations. 

\textbf{Method}: We compared the frequency of below-chance results on forced choice performance validity tests (PVTs) at two levels of significance, .05 and .20, and when using significance testing on individual subtests of the PVTs compared with total scores in claimants for Social Security Disability in order to determine the rate of the expected increase. 

\textbf{Results}: The frequency of significant results increased with the higher level of significance for each subtest of the PVT and when combining individual test sections to increase the number of test items, with up to 20% of claimants showing significantly below-chance results at the higher p-value. 

\textbf{Conclusions}: These findings are discussed in light of Social Security Administration policy, showing an impact on policy issues concerning child abuse and neglect, and the importance of using these techniques in evaluations for Social Security Disability.

\section*{Introduction}

In neuropsychological and medical evaluations, examinees may malinger by intentionally underperforming on measures of ability or by exaggerating symptoms (Heilbronner et al., 2009). Malingering is defined as the intentional production or exaggeration of symptoms or abnormal behavior, or the manipulation of test performance, in order to deceive the examiner into incorrectly concluding that the claimed dysfunction is greater than what actually exists, for the purpose of obtaining an external incentive (Slick, Sherman, & Iverson, 1999; Bianchini, Greve, & Glynn, 2005; Heilbrunner et al., 2009; Chafetz et al., 2015). For the incentive of Social Security Disability (SSD) benefits, claimants who are malingering may
exaggerate their functional limitations in order to meet SSD disability requirements (Chafetz et al., 2015). Performance validity tests (PVTs) are useful for assessment of the validity of cognitive testing (Armistead-Jehle & Denney, 2015; Bianchini et al., 2005; Larrabee, 2012). However, for several years, the Social Security Administration (SSA) has actively discouraged the use of validity testing in its consultative examinations (Chafetz et al., 2015) for reasons that are outside the scope of this paper (see Chafetz, 2010).

PVTs that employ established cut-offs are designed to assess the validity of performance in neuropsychological evaluations (Larrabee, 2012). Many PVTs utilize a forced choice format in which the examinee must choose between targets and foils. Apart from the use of established cut-offs, any forced choice test score falling below chance can also be tested for statistical significance. Testing determines if the score is sufficiently low that it is unlikely it could have been generated by unlucky guessing of the incorrect answers by truly impaired individuals. If the result is statistically significant, then it is highly probable that the examinee intentionally generated incorrect answers (Binder, 1990; Pankratz & Erickson, 1990; Slick et al., 1999).

A PVT score that is significantly below chance provides the strongest psychometric evidence of deliberate underperformance that is consistent with a diagnosis of malingering because it provides evidence that the examinee intentionally gave wrong answers. The alternative explanation that this score was obtained by consistently unlucky guessing of the answers by impaired individuals is less likely because the cutting scores obtained on samples composed of impaired individuals with well-documented brain injuries or disease are well above chance. For example, the below-chance score might be compared to the much higher scores obtained in a sample with well-documented moderate to severe traumatic brain injuries (Tombaugh, 1996).

The original Slick et al. (1999) criteria for the determination of malingering, as well as the revised criteria (Slick & Sherman, 2013), indicated that a PVT score significantly below chance at the .05 level provides ‘definite’ evidence of malingering in the context of external incentives and symptoms and behaviors that cannot be fully explained by neurological, psychiatric, or developmental factors (D Criteria). Pankratz and Erickson (1990) described a significantly below-chance result on a forced choice PVT as ‘the smoking gun of intent’ because a higher (chance-level) score would be obtained if the claimant was blind to the choices and guessed the answer for each item. Guessing is likely to generate a chance-level performance. Intentionally providing a sufficient number of wrong answers leads to a score that is significantly below chance.

Chafetz (2008) provided base-rate evidence of the frequency of significantly below-chance PVT failure in SSD claimants for Supplemental Security Income (SSI; children and adults) and Social Security Disability Insurance (SSDI; adults) income. This study showed that 12.3–13.6% of adult claimants and 10% of child claimants had PVT scores that were significantly below chance, defined with a binomial probability level of \( p < .05 \). As these figures provide unequivocal evidence of a large number of adult and child claimants who are deliberately misrepresenting their claimed abilities in a public disability program, they potentially have policy implications for the SSA system.

Given the medicolegal importance of statistically significant below-chance results on PVTs, it is important to optimize the method of testing of significance. Binder, Larrabee, and Millis (2014) addressed the issues of probability levels, one-tailed vs. two-tailed tests, and combination of PVT scores. The probability level of .05 used for testing hypotheses about
issues such as the effects of pathogens and treatments or the associations between historical factors and dependent variables, using two or more independent samples from different populations, had been arbitrarily applied to the completely different statistical issue of determining if a forced choice test result was significantly worse than chance. As Binder et al. (2014) pointed out, the significance level for forced choice testing was appropriated from research on groups without receiving a logical or statistical test of its accuracy or appropriateness. The widespread acceptance of a probability level of .05 for comparison of groups was not a sufficient reason to use it for analysis of the results of a forced choice test in a single case. To determine an empirically based probability level for testing significance, they examined data from the Warrington Recognition Memory Test-Words (WRMT-W) from a sample of 127 adult participants with traumatic brain injuries (TBIs) ranging from mild to severe, none of whom had financial incentives for their test performance. None scored worse than 23/50 correct (46%). The cut-off score from this sample of 22/50 yielded a one-tailed p level of .24 when the binomial was calculated, rather than estimated from the z approximation. This same cut-off score was also supported on the WRMT-W in another sample of participants with mixed etiologies cited by Binder et al. (Kim et al., 2010). Based on these data, Binder et al. recommended a probability level of .20 for testing the hypothesis that a below-chance score was significant.

Binder et al. recommended one-tailed significance testing for below-chance scores to test the hypothesis that the results were significantly below chance. Their review of the literature showed no consensus about the issue of one- or two-tailed testing. They reasoned that the hypothesis of below-chance performance was uni-directional and one-tailed; there was no need for significance testing when scores were above chance, as low scores that were 50% correct or better were compared with empirically based cut-off scores.

Combining PVT scores also was recommended in order to increase the number of items and thus the statistical power of significance testing with the binomial theorem (Binder et al., 2014). Scores from tests divided into more than one subtest could be added together. Consistent with the recommendation of the authors of the Victoria Symptom Validity Test (VSVT; Slick, Hopp, Strauss, & Thompson, 2005) and of the Portland Digit Recognition Test (PDRT; Binder, 2007), significance could be tested on either section of the tests (easy or hard items) or on the total score. As long as they are independent, subtest scores could be combined (Binder et al., 2014) to increase statistical power. For example, on the Test of Memory Malingering (TOMM; Tombaugh, 1996), with three sections, each containing 50 items, for 21 correct of 50 items (42%), \( p = .161 \). If the same percentage correct (42%) was obtained on the total score for all three sections of the test (63 correct of 150 items), \( p = .03 \).

Subtests that are not independent cannot be combined. For example, the Word Memory Test (WMT; Green & Astner, 1995) and the Medical Symptom Validity Test (MSVT; Green, 2004) have two forced choice subtests: Immediate Recognition (IR) and Delayed Recognition (DR). Scores on IR and DR are independent and can be combined. However, the Consistency score for both tests is dependent on IR and DR subtest behavior, and thus cannot be combined with these other scores to test the significance of below-chance performance.

As reviewed in Binder et al. (2014), most reports of the frequency of significantly below-chance findings in the literature were based on \( p \) levels of .05. Both one-tailed and the more conservative two-tailed tests of significance were reported. While most of these reports are about single subtests, some investigators (Binder, 2007; Green, 2008; Slick et al., 2005) have
taken the opportunity to combine test scores when reporting the frequency of significantly below-chance scores.

While Chafetz (2008) already reported on the frequency of significantly below-chance (p < .05) findings using the TOMM and MSVT in SSD samples of adults and children, this prior study utilized only this conservative probability level, analyzing only single (and not combined) subtests. The present study was designed to determine the frequency of significantly below-chance findings using more recent, empirically derived probability levels while also examining combined subtest scores. Obviously, a higher p level of .20 as recommended by Binder et al. (2014) will produce a higher frequency of significantly below-chance findings, but it was of interest to determine just how much higher. As these findings are potentially of interest for SSA policy concerning misrepresentation of a claimant’s disabling problems, they are discussed in relation to SSA rules. Policy concerning child disability was also of concern, as Chafetz and Dufrene (2014) have implicated below-chance PVT findings in child abuse and neglect.

**Methods**

**Participants**

As previously described in Chafetz, Abrahams, and Kohlmaier (2007) and Chafetz (2008), archived records from consecutive Disability Determinations Services (DDS) referrals to the practice of Michael Chafetz for the psychological consultative examination (PCE), most alleging low cognitive functioning, were used. When the claimants signed the HIPAA notification, they were asked for permission to use the scores from their examination for research and assured that their identity would be protected. Permission was granted if the claimant initialed next to the research notification. Only one claimant declined, and her scores were not used. The first author presented preliminary findings of this research at a statewide Louisiana DDS meeting (April 28, 2003), and the local DDS Medical Liaison Officer provided written acknowledgment about this presentation and the research involved.

Table 1 shows the demographic tabulations of the records. As indicated (Chafetz, 2008), participants in the TOMM study were 232 consecutive DDS referrals for the PCE. Of the 221 referrals who were administered the TOMM, 129 were WAIS-III-age, and 92 were WISC-III-age, and the two data-sets were separated into samples of adults and children. When DDS shifted to the use of the WISC-IV, the MSVT was then used for validity testing. In the MSVT study, there were 55 WAIS-III-age participants and 27 WISC-IV-age participants, which were again separated into samples of adults and children. As the present study had different objectives.

<table>
<thead>
<tr>
<th>Table 1. Sample characteristics for adults and children on the Test of Memory Malingering and Medical Symptom Validity Test (adults only) studies.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Test of Memory Malingering</strong></td>
</tr>
<tr>
<td><strong>Adult</strong></td>
</tr>
<tr>
<td>N</td>
</tr>
<tr>
<td>Age</td>
</tr>
<tr>
<td>Education</td>
</tr>
<tr>
<td>Gender</td>
</tr>
<tr>
<td>Ethnicity</td>
</tr>
</tbody>
</table>

Notes: B = Black/African-American; W = White; O = Other; M = Male; F = Female.
1M (SD).
than in Chafetz (2008), there were differences in the number of records involved. For example, the Chafetz (2008) study had more restrictive inclusion rules for the frequencies in each condition, requiring each case to have all PVT and IQ data. Except for the Adult MSVT table, the tables that show the rate of significantly below-chance findings in the base-rate study (Chafetz, 2008) have fewer cases than in the present study. The Child MSVT data were not included in the current study due to low sample size.

**Performance validity tests**

TOMM (Tombaugh, 1996): In the original SSD study (Chafetz et al., 2007), two trials of the TOMM were used in the formal testing of performance validity. The TOMM is not described here because it is well known to the readers. The Retention trial was not used. MSVT: The MSVT is a forced choice test that contains a list of 10 pairs of words presented twice to each claimant on a computer screen. Claimants are then asked to choose the correct target words from pairs consisting of a target and a foil. Because most claimants – adults and children – complained of reading difficulties, the examiner read all the words and choices on the screen, thereby effecting a combined computer and oral administration (Chafetz, 2008). This accommodation of impaired claimants, removing a possible source of error from reading or executive dysfunction and leaving only the claimant’s (forced) choices, was further discussed in Chafetz (2015). We know of no research examining the effect of this modification of the standard administration, though we note that the origin of the WMT changed from oral to computer administration without change of cut-offs, and the equivalence of the oral and computer administrations has been demonstrated (Hoskins, Binder, Chaytor, Williamson, & Drane, 2010). However, any hypothetical change of scores from this hybrid method would likely be in the positive direction (e.g. eliminating impulsive responding errors), which would render our results in this study less likely to be due to inadvertent responses in truly impaired individuals. Moreover, Dr. Paul Green, the author of the MSVT, suggested in a personal communication (May 11, 2017) that this hybrid method would assist the very few with severe reading difficulties, thereby minimizing failures.

**Binomial probability significance levels**

Statistical significance was tested with the binomial theorem, as calculated online at the Vassarstats website [http://vassarstats.net/textbook/ch5apx.html](http://vassarstats.net/textbook/ch5apx.html), which provides an exact binomial probability calculator. The exact binomial probability ($P_b$) is obtained using various applications of the formula:

$$P_b(k \text{ out of } N) = \frac{N!}{k!(N-k)!} \cdot p^k \cdot q^{N-k}$$

In this formula: $N =$ number of opportunities for an event to occur (e.g. choosing the target word), $k =$ number of times that the event occurs or is stipulated to occur; in this case, the number of correct target choices, $p =$ probability the event will occur on any particular occasion (e.g. .5 for 2 forced choice), $q =$ probability the event will not occur on any particular occasion (e.g. .5 for 2 forced choice).

We used the $P_b$ value for a one-tailed test, applying the formula to all values of $k$ equal to or smaller than the stipulated $k$ amount, which is the probability of getting $k$ or fewer
events. For the MSVT and the TOMM (or any other PVT with two forced choice alternatives), Pb refers to the probability of obtaining such a result by chance. For example, for 20 (or fewer) correct of 50 items on the TOMM, \( k = 20, N = 50 \), the one-tailed exact binomial probability \( (P_b) \) is .1013.

**Results**

**Adult sample**

In the adult MSVT sample, 7 (12.7%) individuals obtained scores on Immediate Recognition (IR) or Delayed Recognition (DR) that were worse than chance at the .05 level, compared with 9 (16.4%) individuals at the .20 level (see Table 2). When MSVT IR and DR scores were summed, 6 (10.9%) individuals performed significantly less than chance at .05, compared with 8 (14.5%) at .20. Considering the typical cut-off on the MSVT (<90%), 33 individuals (60%) performed worse than the cut-off score on either IR or DR.

For either trial on the TOMM, 18 (14%) individuals scored worse than chance at the .05 level, compared with 26 (20.2%) at the .20 level. When scores on both trials were summed, the number of participants with significantly below-chance scores increased from 19 (14.7%) at the .05 level to 22 (17.1%) at the .20 level. For the conventional level of failure on the TOMM (Trial 2 < 90%), 72 individuals (55.8%) scored below cut-off (see Table 3). The reader will notice from the tables that the number of individuals in the summed trials with significantly below-chance results was the same as the highest MSVT trial alone and between the two TOMM trials. While the statistical power increases with summed trials, this finding is an expected result when one of the trials has several fewer claimants obtaining below-chance results.

**Child sample**

For either of the TOMM trials, 11 (12%) children scored worse than chance at the .05 level, compared with 12 (13%) children at the .20 level. Summing the scores on Trials 1 and 2 yielded 8 (8.7%) children at the .05 level and 11 (12%) children at the .20 level. For the traditional failure cut-off (<90%) on Trial 2 of the TOMM (used in Chafetz et al., 2007), 26 (28.3%) children scored below cut-off (see Table 4).

**Table 2.** Frequencies of significantly below-chance scores at different binomial probability levels: adults MSVT.

<table>
<thead>
<tr>
<th>Cut score (≤X)</th>
<th>Percent correct (20 trials)</th>
<th>Binomial Probability</th>
<th>IR% (#)</th>
<th>DR% (#)</th>
<th>IR or DR% (#)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSVT IR or DR (20 trials)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>25</td>
<td>.0207</td>
<td>9.1 (5)</td>
<td>7.3 (4)</td>
<td>12.7 (7)</td>
</tr>
<tr>
<td>7</td>
<td>35</td>
<td>.1316</td>
<td>14.5 (8)</td>
<td>10.9 (6)</td>
<td>16.4 (9)</td>
</tr>
<tr>
<td>17</td>
<td>85</td>
<td>.9998</td>
<td>52.7 (29)</td>
<td>58.2 (32)</td>
<td>60 (33)</td>
</tr>
<tr>
<td>Cut score (≤X)</td>
<td>Percent (40 trials)</td>
<td>Binomial Probability</td>
<td>IR + DR% (#)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSVT IR plus DR (40 trials)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>35</td>
<td>.0403</td>
<td>10.9 (6)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>40</td>
<td>.1341</td>
<td>14.5 (8)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: MSVT = Medical Symptom Validity Test; TOMM = Test of Memory Malingering; IR = Immediate Recognition; DR = Delayed Recognition; T1 = Trial 1; T2 = Trial 2. In these tables, the one-tailed binomial probabilities are given for all successes (# items correct) up through the cut score. For example, if the cut score is ≤7, the binomial probability for 7 or fewer successes is given. Traditional failure cut score is less than 90% (for TOMM just T2). Adults (MSVT: \( n = 55 \); TOMM: \( n = 129 \)).
Discussion

The use of the less conservative probability level for significance testing of .20 led to a small increase in the frequency of significant results, compared with a probability of .05. In the adult MSVT sample, for either recognition trial, the increase was from 12.7 to 16.4%. Using the combined MSVT recognition trials, the increase was from 10.9 to 14.5%. In the TOMM adult study, using either trial, the increase was from 14 to 20.2%. With summed trials, the increase was from 8.7 to 12%. In comparison, on the VSVT, in a sample not known to have external incentives for poor performance, 0.3% were below chance at the .05 level and 2.3% at the .20 level (Loring, Larrabee, Lee, & Meador, 2007). The current results also are consistent with a prior estimate (Binder et al., 2014) that was based on data from the PDRT (Greve & Bianchini, 2006).

When testing the level of significance, we recommend the exact calculation of the binomial, using an online calculator such as VassarStats, rather than the z approximation that was used in the prior paper (Binder et al., 2014). In some instances, the approximation gives a non-precise estimate.

Both the original Slick criteria (Slick et al., 1999) and the revised criteria (Slick & Sherman, 2013) specified a .05 level of significance for definite malingering. Binder et al. (2014)

Table 3. Frequencies of significantly below-chance scores at different binomial probability levels: adults TOMM.

<table>
<thead>
<tr>
<th>Cut Score (≤X)</th>
<th>Percent (50 trials)</th>
<th>Binomial Probability</th>
<th>T1% (#)</th>
<th>T2% (#)</th>
<th>T1 or T2% (#)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOMM T1 or T2 (50 Trials)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>36</td>
<td>.0325</td>
<td>6.2 (8)</td>
<td>13.2 (17)</td>
<td>14.0 (18)</td>
</tr>
<tr>
<td>21</td>
<td>42</td>
<td>.1611</td>
<td>11.6 (15)</td>
<td>19.4 (25)</td>
<td>20.2 (26)</td>
</tr>
<tr>
<td>44</td>
<td>88</td>
<td>.9999</td>
<td>–</td>
<td>55.8 (72)</td>
<td>–</td>
</tr>
<tr>
<td>Cut Score (≤X)</td>
<td>Percent (100 trials)</td>
<td>Binomial Probability</td>
<td>T1 or T2% (#)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOMM T1 plus T2 (100 trials)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>41</td>
<td>.0443</td>
<td>14.7 (19)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>45</td>
<td>.1841</td>
<td>17.1 (22)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: MSVT = Medical Symptom Validity Test; TOMM = Test of Memory Malingering; IR = Immediate Recognition; DR = Delayed Recognition; T1 = Trial 1; T2 = Trial 2. In these tables, the one-tailed binomial probabilities are given for all successes (# items correct) up through the cut score. For example, if the cut score is ≤ 7, the binomial probability for 7 or fewer successes is given. Traditional failure cut score is less than 90% (for TOMM just T2). Adults (MSVT: n = 55; TOMM: n = 129).

Table 4. Frequencies of significantly below-chance scores at different binomial probability levels: children TOMM.

<table>
<thead>
<tr>
<th>Cut Score (≤X)</th>
<th>Percent (50 trials)</th>
<th>Binomial Probability</th>
<th>T1% (#)</th>
<th>T2% (#)</th>
<th>T1 or T2% (#)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOMM T1 or T2 (50 Trials)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>36</td>
<td>.0325</td>
<td>6.5 (6)</td>
<td>12 (11)</td>
<td>12 (11)</td>
</tr>
<tr>
<td>21</td>
<td>42</td>
<td>.1611</td>
<td>8.7 (8)</td>
<td>12 (11)</td>
<td>13 (12)</td>
</tr>
<tr>
<td>44</td>
<td>88</td>
<td>.9999</td>
<td>–</td>
<td>28.3 (26)</td>
<td>–</td>
</tr>
<tr>
<td>Cut score (≤X)</td>
<td>Percent (100 trials)</td>
<td>Binomial Probability</td>
<td>T1 + T2% (#)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOMM T1 plus T2 (100 trials)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>41</td>
<td>.0443</td>
<td>8.7 (8)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>45</td>
<td>.1841</td>
<td>12 (11)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: TOMM = Test of Memory Malingering; IR = Immediate Recognition; DR = Delayed Recognition; T1 = Trial 1; T2 = Trial 2. In these tables, the one-tailed binomial probabilities are given for all successes (# items correct) up through the cut score. For example, if the cut score is ≤ 7, the binomial probability for 7 or fewer successes is given. Traditional failure cut score is less than 90% (for TOMM just T2). Children (TOMM: n = 92).
L. M. Binder and M. D. Chafetz proposed broadening this criterion to .20, based on statistical principles, cutting scores, and data from two samples of brain injured persons, including one sample found in the literature (Kim et al., 2010). Of potential concern is the possibility that impaired individuals who guess on all items are more likely to obtain results significantly worse than chance with the higher p level of .20. Such performances, if they were to occur, should not be classified as deliberate provision of wrong answers. Our current data on the frequency of below-chance results in our sample do not directly address the validity of the higher .20 probability level.

However, the empirical literature does not support the concern that misclassification is likely with the higher probability level in clinical samples with no known external incentives for poor performance. We reviewed forced choice test data that could be extracted for individual participants from several papers and two test manuals (Binder, 2007; Binder & Kelly, 1996; Carone, 2008; Chafetz & Biondolillo, 2012; Greve & Bianchini, 2006; Greve, Bianchini, & Doane, 2006; Haber & Fichtenberg, 2006; Howe, Anderson, Kaufman, Sachs, & Loring, 2007; Kim et al., 2010; Loring et al., 2007; MacAllister, Nakhutina, Bender, Karantzoulis, & Carlson, 2009; Macciocchi, Seel, Alderson, & Godsall, 2006; Schroeder et al., 2012; Singhal, Green, Ashaye, Shankar, & Gill, 2009; Slick et al., 2003). We also contacted one of these authors of a pediatric study and obtained previously unpublished details about the results of individual cases in one published study cited here on the TOMM, additional unpublished data on 93 pediatric cases that followed the original study of the TOMM, and 25 unpublished pediatric cases with the MSVT (William MacAllister, personal communications, February 23 and March 1, 2017). In a published study of children with fetal alcohol syndrome (Larson et al., 2015), 6 of 120 cases performed below the usual cut-offs on the MSVT; we were unable to learn via personal communication how many cases performed below chance at the .20 level, and we did not include this study in our aggregate summary. Some of the samples in the studies reviewed above include children with low intelligence. In aggregate, including the unpublished cases, 9 of 1630 cases (0.6%) with no known external incentives for poor performance scored worse than chance (p < .20) on a forced choice test.

The proportion of individuals obtaining significantly worse than chance results at the .20 level in these published and unpublished studies is much less than the proportion of .20 that would be expected if everyone was guessing the answers to the PVTs. We analyzed data from the Loring et al. (2007) study because this was the study with the highest frequency of below-chance responding; 8 of the 346 people (2.3%) in the clinical sample performed worse than chance. The obtained percentage in the Loring et al. (2007) sample is significantly less than the 20% expected if every participant guessed on all items, z = −8.24, p < .00001. For all 1630 cases, z = −19.62. If below-chance results in these published and unpublished samples result from random guessing, then only a small proportion of the participants are guessing. It also is possible that some of the participants in these studies who were classified as lacking external incentives actually were misclassified and had external incentives for their scores.

We also reviewed some of the experimental studies that asked participants to simulate cognitive impairment (Armistead-Jehle & Denney, 2015; Binder & Willis, 1991; Frazier, Frazier, Busch, Kerwood, & Demaree, 2008; Green, Flaro, & Courtney, 2009; Powell, Gfeller, Hendricks, & Sharland, 2004; Singhal et al., 2009; Vickery et al., 2004). The mean scores on forced choice PVTs for groups simulating cognitive impairment in these seven studies including nine groups were above 50% correct, except for a group simulating reading disorder on the VSVT
hard items (Frazier et al., 2008) and a group simulating TBI on the PDRT hard items that averaged just under 50% correct (Binder & Willis, 1991).

In summary, experimental studies have demonstrated that most simulators instructed to feign brain dysfunction perform better than chance. The clinical studies have shown that only 0.6% of various samples with no known external incentives, many of them with obvious cognitive impairment, performed significantly worse than chance \( (p < .20) \). Therefore, significantly below-chance findings \( (p < .20) \) have great clinical significance.

We found that a high rate of SSD claimants performed worse than chance at the .20 level, with up to 20% of adult claimants and up to 13% of child claimants performing significantly worse than chance on a performance validity test. The performance validity tests, in isolation, do not necessarily indicate that the examinees were malingering. The diagnosis of malingering includes the criterion of an external incentive for poor performance, which also is met in all SSD claimants.

Our findings potentially have an impact on policy and practice within the SSA. Administrative Law Judge Michael Gilbert (Personal Communication, 2016) highlighted that SSA and the Office of the Inspector General (OIG) have developed the Cooperative Disability Investigations (CDI) program (https://oig.ssa.gov/cooperative-disability-investigations-cdi), which works with the state DDSs and other departments of SSA to prevent and investigate fraud in various programs, including disability. Since the CDI program was established in 1998, investigative efforts have resulted in a reported $3.2 billion in projected savings to SSA’s disability programs. Judge Gilbert indicated that it is unlikely that many of these investigations have been about malingering per se. Most investigations concerned issues such as inconsistencies in the record, concealing work after getting benefits, college activities, other activities of daily living, or lying about other facts of the claim. The use of this neuropsychological performance validity technology could benefit the CDI investigations if the DDSs were to refer based upon consultative examination findings.

We emphasize that psychologists do not have the legal authority to equate any of their findings with a determination of fraud. The authors are merely providing education to the fact-finders (e.g. administrative law judges) within SSA for their own determinations or referrals to OIG. The mere evidence of significantly below-chance results, along with other information in the record, may demonstrate a pattern that is factually meaningful to the fact-finder charged with assessing myriad data points throughout the medical record (Michael Gilbert, Personal Communication, 2016).

It is important to determine the practical significance of the current findings, as potentially the relaxation of the below-chance criterion may have an impact on SSA’s assessment of its evidentiary requirements. This is a difficult proposition. The publicly reported actuarial tables of adult disabled beneficiaries by diagnostic group (only mental disorders) show that in SSDI there were 2,727,043 Disabled Worker Beneficiaries and in SSI 2,819,246 Adult Recipients, ages 18–64 for 2015 (SSA, 2016a, 2016b). Compared to the 2011 numbers reported in Chafetz and Underhill (2013), the 2015 SSI report did not break out the numbers by diagnostic category. Therefore, the previously reported percentage of mental disorders (58.6% of adult SSI cases) was used. Under the current rules, these figures represent beneficiaries for whom no validity evidence was systematically obtained (Chafetz, 2010; Chafetz et al., 2015). However, not all beneficiaries were sent for PCEs from which the current below-chance data were obtained, and it is known within the system that not all consultants make their recommendations fully from the results of the PCEs.
Another way of examining this problem is to study the disposition of applications for SSI disability benefits in 2015 (Table V.C1; SSA, 2016b). Of the 1,426,419 SSI cases filed in 2015, with 1,167,472 total decisions in that year, 29% were provided allowances (with some occurring after appeals and reconsiderations), while 71% were denied for various reasons (including: the impairment was not expected or did not last 12 months; the impairment was not severe or did not cause severe functional limitations; the claimants were able to do their past work or another type of work; or there was insufficient evidence, failure to cooperate, did not want to continue the claim, or returned to work). Of these denials, 17.7% (206,463) were still on appeal for the next step in the appeals process. The knowledge that up to 20% of claimants sent for PCEs to garner evidence for the fact-finding in these cases were showing ‘the smoking gun of intent’ to misrepresent impairment would have been informative in any stage of the process, including the further appeals at the Administrative Law Judge level.

The high rate of significantly below-chance findings in the child sample is of special concern. Feigning at the direction or pressure by others in the context of a financial incentive has been labeled malingering-by-proxy (MBP) (Slick et al., 1999; Slick & Sherman, 2013). As Chafetz and Dufrene (2014) asserted from their reviewed cases and from prior examples (see Cassar, Hales, Longhurst, & Weiss, 1996), MBP can be a form of abuse and may be treated as such by protective agencies. Chafetz and Dufrene (2014) proposed that in cases of MBP, which are especially signaled by below-chance findings, that the cases be referred to the local Child Protection Agency for investigation as a matter of policy.

One of the limitations of the current study comes from the definition of impairment in the study of Binder et al. (2014) as ranging from mild to severe TBI, without knowing the percentages or impact of the TBI levels. While none of the TBI subjects, including severe TBI, obtained a below-chance finding beyond the $p < .20$ level, these TBI subjects are not likely comparable to the claimants with low intellectual functioning in the current SSD sample. However, this is not likely to have created any interpretive problem in the current study, as it is known that well-motivated claimants with low intellectual functioning (IQ: 60–75) do not typically fail the MSVT or other PVTs designed for low-IQ claimants, much less obtain below-chance scores (Chafetz & Biondolillo, 2012). Furthermore, in this study, we aggregated literature showing that below-chance PVT findings at a probability level of .20 had a frequency of only 0.6% in clinical samples.

As with any study, it is possible that the frequency of below-chance findings is sample dependent. The frequency of below-chance findings should be reported in future studies of different populations. Another limitation is the non-random placement of targets and foils in the MSVT. A more conservative approach to the testing of significance may be indicated for tests with topographic locations of targets and foils that are non-random.

Neuropsychologists are aware that malingering at any level cannot be defined by a test result alone, and that it takes adherence to established guidelines (e.g. Slick & Sherman, 2013) for appreciation of the context of the findings. Moreover, we caution that neuropsychologists should not liberalize the established cut-offs without carefully validated empirical justification, such as Greve and Bianchini’s work on the TOMM (2006).

In summary, the recommended use of a probability level of .20 for determining if a forced choice test score is significantly below chance (Binder et al., 2014) led to a small but meaningful increase in the frequency of significant results in disability claimants. Significant results in the context of an external incentive likely signify deliberate efforts to mislead the examiner,
i.e. malingering. The frequency of significant results in this sample of applicants for Social Security benefits has implications for the administration of these benefits.

Acknowledgments

Authorship is alphabetical. We thank Jonathan Ord for his description of SSA administrative law on fraud, and Glenn Larrabee and Scott Millis for their advice on statistical matters. We thank three anonymous DDS consultants for their discussion about DDS process. We would also like to thank the TCN reviewers, whose comments helped improve the manuscript. Thanks to Melissa Greenfeld for her technical help. Note: We thank Judge Michael Gilbert for his analysis concerning the reporting of fraud within the SSA-OIG system. The views he expressed in this paper are his, made in his personal capacity as a private citizen. He does not in this paper represent the views of the Social Security Administration or the United States Government. The information he has provided in this paper is based on publicly available sources unrelated to his service with the Social Security Administration. Judge Gilbert is not acting as an agent or representative of the Social Security Administration or the United States Government in this activity. There is no expressed or implied endorsement of his views or activities by either the Social Security Administration or the United States Government.

Disclosure statement

Laurence Binder is the author and publisher of the Portland Digit Recognition Test. Michael Chafetz authored Intellectual Disability: Criminal and Civil Forensic Issues, published by Oxford University Press, from which he receives royalties. The authors perform medicolegal evaluations.

References


