Malingering on the Social Security Disability Consultative Exam: A New Rating Scale

Michael D. Chafetz a,*, Joel P. Abrahams b, Joy Kohlmaier a

a 3520 General DeGaulle Dr. #4098, New Orleans, LA 70114, United States
b River Oaks Hospital, 1525 River Oaks Road W., Harahan, LA 70123, United States

Accepted 4 October 2006

Abstract

In disability examinations, benefits may depend on the findings of a psychological consultative examination (PCE), which in Louisiana usually involves a mental status examination and a Wechsler Scale. The disability determinations service (DDS) requires a warning that failure to do one’s best may result in an unfavorable decision on the claim, but psychologists are officially discouraged from determining effort by the use of formal effort tests. Consequently, there is a need for internal indicators of effort. Formal testing of effort was undertaken in order to identify indicators of effort within the PCE in WAIS-age and WISC-age claimants. Our findings indicated that the total score of indicators was more predictive of effort than any single indicator. Regression equations yielded information on how much effort contributes to IQ. Classification accuracy for the new rating scale was described for a “dose-response” of effort. Disincentives for malingering detection in the PCE were identified.

© 2006 National Academy of Neuropsychology. Published by Elsevier Ltd. All rights reserved.

Keywords: Malingering; Disability; Social security; Mental retardation

1. Introduction

Social security disability compensation is dependent in part on the findings of the psychological consultative examination (PCE). The disability determinations service (DDS) office requires that a written warning be given to the claimant. This warning states, “Failure to do your best on these tests may result in an unfavorable decision on the claim.” Moreover, psychologists must also include a statement of the validity of their test findings, and the Social Security examiners often ask psychologists to comment on malingering. However, official guidance from Disability states: “SSA’s position is that results derived from tests of malingering are not programmatically useful in determining if an individual meets the SSA definition of disability” (Medical Liaison Officer, personal communication). Indeed, the term “malingering” is not found in the entire Blue Book publication (US Dept. of Health and Human Services, 1994, revised 2006).

The finding of malingering on the PCE is a serious event, as it constitutes an allegation of fraud by a claimant who is seeking compensation from the government for problems that are allegedly disabling. According to the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV, American Psychiatric Association, 1994), “malingering is the intentional production of false or grossly exaggerated physical and psychological symptoms for...
external incentives such as obtaining monetary compensation...”. Although one may conceive of the needy claimant within an adaptational model (Rogers, 1990) in which the claimant is struggling to adapt to indifference or opposition to his or her needs, it is not the stated purpose of the social security disability system to provide compensation when a compensable disability does not exist.

The seriousness of the problem can be brought into perspective when one considers the total estimated outlay of Social Security Administration Disability Insurance in 2004 was $80.33 billion. In addition, the projected outlay for Disability Insurance benefits for the old age survivors and disability insurance (OASDI) program in 2004 was $77.95 billion, and the 2004 Supplemental Security Income outlay was $37.36 billion (Social Security Administration, 2003). The base rate range of malingering in private disability cases is estimated to be 30% (Mittenberg, Patton, Canyock, & Condit, 2002). If the base rate of malingering in the public-sector disability determinations is as high as in the private sector, then a large amount of money likely is being spent on fraudulent claims.

While there exists a need to separate fraudulent from valid claims, there is no category of payment to psychologists for effort testing. The decisions about disabling problems are made from behavioral observations, mental status results, Wechsler scales, and other evidence gathered by examiners in the DDS office. Given that Social Security discourages formal effort testing, it is not surprising to find that a PsycInfo search through 2005 revealed no studies of effort on the PCE. This deficit is starting to change as investigators have recognized the importance of studying effort in the PCE (Chafetz & Lambert, 2005; Chafetz & Abrahams, 2005; Miller, Boyd, Cohn, Wilson, & McFarland, 2006; Yanez, Fremouw, Tennant, Strunk, & Coker, 2006). In the period from 2002 (when the first author introduced formal effort testing into his PCE) until this writing, evidence from PCE reports (and personal communication from PCE colleagues) have indicated that blatant floor violations and inconsistencies are the primary tools used to examine effort. However, in many, if not most, cases, there are no comments regarding effort. Because effort can account for up to 50% of the variability in neurocognitive findings (Green, Rohling, Lees-Haley, & Allen, 2001), it is critical to have a systematic method to examine effort in the PCE for Social Security, where claimants are all seeking compensation.

The psychological exam itself provides numerous opportunities to study effort. As Slick, Sherman, and Iverson (1999) have indicated, studying the pattern of performance is an effective way to detect malingering within evaluations. Methods used to study pattern of performance include inspecting “floor” items for rare mistakes, highlighting unusual patterns or responses, and looking at the magnitude of errors (i.e., errors more than expected). Within a Wechsler scale (WAIS-III or WISC-IV), one can derive the reliable digit span (RDS), use indices from Mittenberg, Theroux, Zielinski, and Heilbronner (1995) (e.g., vocabulary > digit span by three scaled score points), examine inconsistencies within the test (e.g., missing coding items by both horizontal and vertical inversions), and examine Ganser-like incorrect answers on Arithmetic or in the mental status examination. Ganser-like incorrect answers (Drob & Meehan, 2000) belie the truth by being consistently close to the correct answers (e.g., $2 + 3 = 6$; $3 + 4 = 8$; $10 - 2 = 7$), thus giving information that the claimant actually knows the correct answers. During the mental status examination, one can note whether the claimant incorrectly reports his age or birthday, or whether he picks the wrong name of the current president out of a short list. A systematic way to examine these issues in the psychological consultative examination would be helpful to the psychologist who is discouraged from using a formal test of effort.

Moreover, individuals who malinger do not necessarily do so in a consistent manner within or between examinations (Greiffenstein, Gola, & Baker, 1995; Meyers & Volbrecht, 2003). Such examinees may be attempting to malinger different types of impairment (e.g., “slow” versus “poor memory”) or may simply have a different style (e.g., missing easy items versus attempting odd errors on harder stimuli).

Due to the lack of consistency from DDS workers in referral questions (e.g., “learning and mental”; “mental and back problems”), it is difficult, if not impossible, to categorize the referral problems. However, it is clear that low intellectual functioning is the central focus of most PCE referrals even when there may be other psychological problems. In the adult sample, fully 76.3% of the consults had received Special Education services, and 85.8% had less than a 12th grade education. Also, when prior PCE evaluations are available, they usually indicate low IQ levels. As the Wechsler tests usually provide the only source of objective data in the PCE, the current inferences of functional limitations are largely made from low scaled scores and low IQ scores. Thus, the examination of effort in the social security disability PCE provides an opportunity for understanding effort in mostly low functioning adults and children.

The purpose of this study is to validate the use of internal indicators within the psychological consultative examination so that psychologists can provide reliable and valid statements to DDS about malinger. A rating scale for the PCE is developed and validated against the TOMM and the medical symptom validity test (MSVT—formerly the MACT: memory and concentration test).
2. Methods

2.1. Subjects

Subjects were claimants from DDS referrals for the PCE, most alleging low cognitive functioning. At the time 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 initialized next to the research notification. Guardians or parents provided permission if the claimant was unable or under age of majority, and children were asked for assent. Only one claimant declined, and her scores were not used. The preliminary findings of this research were presented by the first author at a statewide Louisiana DDS meeting (April 28, 2003), and written acknowledgement of this research was provided by the Medical Liaison Officer (personal communication).

The subjects in the TOMM study were 232 consecutive DDS referrals for the PCE. Of these referrals, 136 were WAIS-III-age, and 96 were WISC-III-age, and the two data sets were separated into “Adults” and “Children.” When DDS shifted to the use of the WISC-IV, the MSVT was then used for effort testing. In the MSVT samples, there were 58 WAIS-III-age subjects, and 27 WISC-IV-age subjects, which were again separated into “Adults” and “Children”. Sample characteristics for both the TOMM and MSVT study are shown in Table 1. Demographics were consistent across samples, and no significant differences were found for age or education among adults and children.

2.1.1. Classification of subjects

The four Slick et al. (1999) criteria include: (A) the existence of a substantial external incentive; (B) evidence from psychometric testing; (C) evidence from self-report; (D) the evidence from B and C is not fully accounted for by psychiatric, neurological, or developmental factors. The classifications arising from these criteria (assuming external incentive) include: definite malingered neurocognitive dysfunction (MND) in which an established symptom validity test (SVT) was failed at a significantly ($p < .05$) below-chance level; probable MND in which an established SVT is simply failed (B) or there are other discrepancies between test data and known patterns of brain functioning, behavioral observations, information from collaterals, or documented history (B), or a discrepancy in B and C (a discrepancy in C alone [self-report] is not sufficient for a diagnosis of malingering in the absence of a B criteria); possible MND in which the criteria for probable MND have been met but there are D factors present.

Because of the nature of the DDS samples in which the claimants were low-functioning and documentation and collateral reports were unreliable, we felt we could not utilize C criteria. Because the FSIQ scores were largely in the mental retardation range, the question naturally arose as to whether these “developmental disabilities” mitigated most of the probable group into a possible group, according to Slick D criteria. On the other hand, if these claimants were indeed malingering, then their FSIQ scores were not valid representations of their true abilities, and the Slick criteria for possible malingering would also not work. Therefore, a “dose-response” set of classification levels were achieved by defining a definite group (TOMM2 < 18; MSVT < 30), and then splitting the probable MND claimants into chance-level failure on an SVT (e.g., TOMM2: 18–32; MSVT: 30–70) and simple failure on an SVT. This method was initially based on a personal communication from Dr. Kevin Greve (2005), who suggested using a subsample of definite malingerers. We then developed the rest of the graded level groupings.

Table 1
Sample characteristics for adults and children on the TOMM and MSVT

<table>
<thead>
<tr>
<th></th>
<th>TOMM Adult</th>
<th>TOMM Child</th>
<th>MSVT Adult</th>
<th>MSVT Child</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>136</td>
<td>96</td>
<td>58</td>
<td>27</td>
</tr>
<tr>
<td>Age a</td>
<td>29.0 (10.9)</td>
<td>10.6 (2.7)</td>
<td>29.5 (12.1)</td>
<td>11.5 (2.6)</td>
</tr>
<tr>
<td>Education a</td>
<td>9.5 (2.0)</td>
<td>3.5 (2.5)</td>
<td>9.1 (2.2)</td>
<td>4.6 (2.6)</td>
</tr>
<tr>
<td>Gender</td>
<td>76 M, 60 F</td>
<td>66 M, 30 F</td>
<td>36 M, 22 F</td>
<td>21 M, 6 F</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>107 B, 27 W, 2O</td>
<td>85 B, 11 W</td>
<td>50 B, 8 W</td>
<td>25 B, 2 W</td>
</tr>
</tbody>
</table>

B = black; W = white; O = other; M = male; F = female.

a $M$ (S.D.).
Table 2
Disability determinations malingering rating scale

<table>
<thead>
<tr>
<th>Item</th>
<th>Description</th>
<th>Score 0</th>
<th>Score 1</th>
<th>Score 2</th>
<th>Score 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Three simple arithmetic (one each)</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>2.</td>
<td>Three simple sequences (one each)</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>3.</td>
<td>DK Pres/pick wrong from list</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td>Miss personal information (age, Bday)</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>5.</td>
<td>Ganser-like answers (0, 2, 3, &gt;3)</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>6.</td>
<td>WAIS – miss items before start – subtests</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>7.</td>
<td>WAIS—low average scaled scores (≥2, [1], 0)</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>8.</td>
<td>WAIS—RDS (&lt;7, &lt;6)</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>9.</td>
<td>WAIS—Voc or PiArrang &gt; DigSp by 3</td>
<td>0</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10.</td>
<td>WAIS—coding (0, 2, &gt;2 errors)</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>11.</td>
<td>Any highly improbable response (ball = triang)</td>
<td>0</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12.</td>
<td>Claims improbable pathology (seeing ghosts)</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Item 9 was subsequently removed from the rating scale due to its unpredicted positive correlation with reference variables. Note on scoring: (1) Score a point for each arithmetic item missed; (2) one point for each sequence missed; (3) one point for not knowing current US president; two points for missing from list; (4) one point for misstating a personal item; two for a second misstatement; (5) one point for at least 2 Ganser-like answers; two points for three answers; three points for more; (6) a point for each subtest in which an item (or more) are missed before the start point (up to three subtests); (7) if two at least low average scaled scores on WAIS-III (ss = 6+), score is 0; 1 low average score gives a one; none gives a 2; (8) RDS score <7 gives a 1; <6 gives a 2; (10) two coding errors gives a 1; >2 errors gives a 2; (11) and (12) can occur anywhere in the exam. There may also be some overlap—e.g., stating there are “11” months in a year gives a point for #6 and two points on #11; stating that 3 + 4 = “53” gives a point on #1 and #2 points on #11.

2.2. Procedures

All participants underwent standard interview, mental status examination, and testing procedures for the PCE. Floor probes and effort testing were administered during the mental status examination. Ganser-like answers and other rating scale material were archived after the exam.

2.2.1. Scale development

Items from the scale were chosen primarily for their utility and ease of use within the PCE for the identification of poor effort or symptom exaggeration. The selection of items initially arose from the first author’s observations of behavior on the mental status examination and Wechsler scales in cases with egregiously low scores on the TOMM.

Table 2 shows the DDS rating scale. Floor items consisted of simplistic tasks based on overlearned information that should be retained even in a population with limited functioning (Slick et al., 1999). The floor task items included: three simple (orally presented) arithmetic calculations (2 + 3; 3 + 4; and 10 − 2), and three simple sequences (13, 14, 15, ___; 21, 22, 23, ___; 2, 4, 6, ___). The floor information items were: Who is the president of the United States? (If the respondent did not know, a list was given to choose from: Albert Brooks, George Bush, or Lafcadio Hearn.) During the orientation questions, another floor item included a notation of whether the respondent missed the question about age or birthday.

Ganser-like incorrect answers were also noted (Drob & Meehan, 2000). For consistency, at least two Ganser-like answers had to be observed before they were rated. They were observed throughout the exam—e.g., arithmetic and information (e.g., 11 months in a year) subtests, or mental status calculation and sequence items. In some cases, one square in block design would be consistently off, but these were not counted due to lack of research information.

Any highly improbable response was noted. For example: What is the shape of a ball? “A triangle”. Also, an improbable drawing of a simple alternating figure was counted. These drawings strained credibility when they included the basic forms but were drawn in grotesque ways, or when the basic forms were grossly distorted. A case example is shown below:
Claiming improbable pathology was another category. These were usually psychotic-like claims that did not fit with the presentation or were not similar to known pathology (e.g., seeing ghosts or perceiving the office plant weeping).

Target items from the Wechsler scales included: the number of subtests in which items were missed before the start point; having less than two scaled scores that were greater than 5; RDS less than 7 or less than 6 (Greiffenstein, Baker, & Gola, 1994; Meyers & Volbrecht, 2003); the vocabulary or picture arrangement (PA) scaled score greater than the digit span scaled score by three or more scaled score points (modification of Mittenberg et al., 1995, 2001, 2002), and two or more coding errors. Coding errors were considered as strong indicators if both horizontal and vertical reversals were made (as perceptual deficits do not usually work both ways), but only the errors were counted.

The DDS Rating Scale originally contained two forms: a dichotomous method to facilitate ease of scoring, and a scaled method to take into account the seriousness of the offense. Point biserial correlation methods were used to account for the associations between dichotomous variables and continuous variables (Cohen & Cohen, 1983)—primarily, the effort variables. The dichotomous variables were eventually dropped due to comparatively lower correlations with the effort variables, as was item 9, the Mittenberg variable (Voc or PA > DigSp by 3), due to its unpredicted positive correlation with effort reference variables.

Two trials of the test of memory malingering (TOMM) (Tombaugh, 1996) were initially used as the formal testing of effort. This visual-based symptom validity test was initially preferred, as reading difficulties were frequently claimed by the DDS claimants. Questions about the relative sensitivity of the TOMM versus the word memory test (WMT) and the medical symptom validity test (Green, 2004), and an interest in using a second validating tool led to the use of the MSVT. The changeover occurred when DDS required the change from the WISC-III to the WISC-IV (October 2004). To obviate potential claims of illiteracy, the examiner modified the normal procedure by reading the directions and the stimuli to the claimants as they were presented on the computer screen. Thus, the procedure was tantamount to a combined oral and computer administration.

Data collection was stopped in August 2005 for two reasons. First, it had become clear to the author that DDS examiners were potentially coaching claimants. One examiner admitted to the first author that the new policy is to call the parent when there is a finding of malingering in a child. The examiner then told the parent, “We know the child is not working up to capacity. If you go to the exam and act like you are 6, I know you are not and you will be taken off.” The DDS examiner then scheduled another exam for the claimant the author had called about. Before the author could investigate this problem more systematically, Hurricane Katrina changed the entire living climate and set of conditions in New Orleans. It was thought that the magnitude of these changes was so extreme that the sample would not be the same.

2.2.2. Item validation

The first step was to identify how well the individual items shared variance with formal tests of effort. The effort test variables included the second trial of the TOMM (T2), the sum of the two trials (T1 + T2), and the main effort variables of the MSVT: immediate recognition (IR), delayed recognition (DR), and consistency (CON). Table 3 shows these results.

As can be seen in Table 3, when T2 is used as the reference variable, all zero order correlations for adults were significant (p < .01). When T1 + T2 was used as the reference variable in adults, again all the item zero order correlations were significant (p < .01). From the table, one can see which items share at least 25% of their variance with T2 and with T1 + T2.

In children, a few variables were not significantly correlated with T2. These included personal information (r = -.19), vocabulary (or PA) > DigSpan (r = .04), and improbable pathology (r = -.05). The variables in children not correlated with T1 + T2 included vocabulary (or PA) > DigSpan (r = .04), and improbable pathology (r = -.05). The table shows items sharing at least 25% of their variance with T2 and with T1 + T2.

When MSVT IR was used as the reference variable in adults, only the variables vocabulary (or PA) > DigSpan (r = -.08), and improbable pathology (r = -.19) were not significant. All other items on the rating scale were correlated significantly with IR (p < .01).

With the MSVT DR as the reference variable in adults, only the variables vocabulary (or PA) > DigSpan (r = -.05), and improbable pathology (r = -.19) were not significant (as with IR). All other items on the rating scale were correlated significantly with DR (p < .01).

When consistency was entered as the reference variable for adults, the items DK President (r = -.21), missed personal information (r = -.20), vocabulary (or PA) > DigSpan (r = -.10), and improbable pathology (r = -.23) were
Table 3
Correlations among rating scale items and effort variables for adults (A) and children (C)

<table>
<thead>
<tr>
<th></th>
<th>T2</th>
<th>T1 + T2</th>
<th>IR</th>
<th>DR</th>
<th>CONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Simple arithmetic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>−.48**</td>
<td>−.47**</td>
<td>−.61**</td>
<td>−.70**</td>
<td>−.48**</td>
</tr>
<tr>
<td>C</td>
<td>−.22</td>
<td>−.24</td>
<td>−.53**</td>
<td>−.44</td>
<td>−.58**</td>
</tr>
<tr>
<td>2. Simple sequences</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>−.49**</td>
<td>−.48**</td>
<td>−.49**</td>
<td>−.55**</td>
<td>−.33</td>
</tr>
<tr>
<td>C</td>
<td>−.46**</td>
<td>−.49**</td>
<td>−.85**</td>
<td>−.78</td>
<td>−.59**</td>
</tr>
<tr>
<td>3. Do not know president</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>−.41**</td>
<td>−.44**</td>
<td>−.40**</td>
<td>−.49**</td>
<td>−.21</td>
</tr>
<tr>
<td>C</td>
<td>−.29**</td>
<td>−.29**</td>
<td>−.39</td>
<td>−.30</td>
<td>−.20</td>
</tr>
<tr>
<td>4. Miss personal information</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>−.33**</td>
<td>−.33**</td>
<td>−.37**</td>
<td>−.36</td>
<td>−.20</td>
</tr>
<tr>
<td>C</td>
<td>−.19</td>
<td>−.23</td>
<td>.18</td>
<td>.18</td>
<td>.11</td>
</tr>
<tr>
<td>5. Ganser answers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>−.60**</td>
<td>−.60**</td>
<td>−.49**</td>
<td>−.58**</td>
<td>−.35**</td>
</tr>
<tr>
<td>C</td>
<td>−.49**</td>
<td>−.52**</td>
<td>−.84**</td>
<td>−.79**</td>
<td>−.65**</td>
</tr>
<tr>
<td>6. Miss items before start</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>−.68**</td>
<td>−.68**</td>
<td>−.50**</td>
<td>−.62**</td>
<td>−.39**</td>
</tr>
<tr>
<td>C</td>
<td>−.37**</td>
<td>−.40**</td>
<td>−.87</td>
<td>−.90</td>
<td>−.50**</td>
</tr>
<tr>
<td>7. Low average scaled scores</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>−.60**</td>
<td>−.59**</td>
<td>−.67**</td>
<td>−.66**</td>
<td>−.41**</td>
</tr>
<tr>
<td>C</td>
<td>−.55**</td>
<td>−.55**</td>
<td>−.68**</td>
<td>−.70**</td>
<td>−.40**</td>
</tr>
<tr>
<td>8. RDS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>−.49**</td>
<td>−.50**</td>
<td>−.51**</td>
<td>−.54**</td>
<td>−.34**</td>
</tr>
<tr>
<td>C</td>
<td>−.35**</td>
<td>−.34**</td>
<td>−.51**</td>
<td>−.52**</td>
<td>−.27**</td>
</tr>
<tr>
<td>9. Vocab &gt; Digit Span by 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>.24**</td>
<td>.24**</td>
<td>−.08</td>
<td>−.05</td>
<td>−.10</td>
</tr>
<tr>
<td>C</td>
<td>.04</td>
<td>.06</td>
<td>.17</td>
<td>.19</td>
<td>.22</td>
</tr>
<tr>
<td>10. Coding errors</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>−.57**</td>
<td>−.56**</td>
<td>−.62**</td>
<td>−.65**</td>
<td>−.32**</td>
</tr>
<tr>
<td>C</td>
<td>−.78**</td>
<td>−.76**</td>
<td>−.70**</td>
<td>−.72**</td>
<td>−.40**</td>
</tr>
<tr>
<td>11. Highly improbable responses</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>−.58**</td>
<td>−.59**</td>
<td>−.49**</td>
<td>−.46**</td>
<td>−.30**</td>
</tr>
<tr>
<td>C</td>
<td>−.54**</td>
<td>−.55**</td>
<td>−.75**</td>
<td>−.76**</td>
<td>−.51**</td>
</tr>
<tr>
<td>12. Improbable pathology</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>−.29**</td>
<td>−.29**</td>
<td>−.19</td>
<td>−.19</td>
<td>−.23</td>
</tr>
<tr>
<td>C</td>
<td>−.05</td>
<td>−.07</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13. Rating scale total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>−.77**</td>
<td>−.77**</td>
<td>−.73**</td>
<td>−.79**</td>
<td>−.50**</td>
</tr>
<tr>
<td>C</td>
<td>−.61**</td>
<td>−.63**</td>
<td>−.89**</td>
<td>−.85**</td>
<td>−.60**</td>
</tr>
</tbody>
</table>

* Correlation is significant at the 0.05 level.
** Correlation is significant at the 0.01 level.

not significant, and none of the items shared at least 25% of their variance with consistency. Therefore, consistency was not used further as a validating variable.

In children, with IR as the reference variable, only the items missed personal info \((r = .18)\), vocabulary > DigSpan \((r = .17)\), and improbable pathology (constant) were not significant. With DR as the reference variable, the nonsignificant items included DK president \((r = −.30)\), missed personal info \((r = .18)\), vocabulary > DigSpan \((r = .19)\), and improbable pathology (constant). When consistency was used as the reference variable in children, the items DK President
Table 4
Standardized regression coefficients for rating scale items and effort variables among adults (A) and children (C)

<table>
<thead>
<tr>
<th>Item</th>
<th>T2</th>
<th>T1 + T2</th>
<th>IR</th>
<th>DR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Simple arithmetic</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>.05</td>
<td>.07</td>
<td>-.12</td>
<td>-.16</td>
</tr>
<tr>
<td>C</td>
<td>.31**</td>
<td>.35**</td>
<td>-.08</td>
<td>.06</td>
</tr>
<tr>
<td>2. Simple sequences</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>-.12</td>
<td>-.09</td>
<td>.05</td>
<td>.01</td>
</tr>
<tr>
<td>C</td>
<td>-.26*</td>
<td>-.29**</td>
<td>-.84**</td>
<td>-.48</td>
</tr>
<tr>
<td>3. Do not know president</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>.02</td>
<td>-.01</td>
<td>.05</td>
<td>-.03</td>
</tr>
<tr>
<td>C</td>
<td>-.11</td>
<td>-.08</td>
<td>-.13</td>
<td>-.03</td>
</tr>
<tr>
<td>4. Miss personal information</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>.03</td>
<td>.02</td>
<td>-.01</td>
<td>.08</td>
</tr>
<tr>
<td>C</td>
<td>.10</td>
<td>.04</td>
<td>.08</td>
<td>.01</td>
</tr>
<tr>
<td>5. Ganser answers</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>-.30**</td>
<td>-.30**</td>
<td>-.02</td>
<td>-.03</td>
</tr>
<tr>
<td>C</td>
<td>-.30**</td>
<td>-.33**</td>
<td>.75</td>
<td>.46</td>
</tr>
<tr>
<td>6. Miss items before start</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>-.32**</td>
<td>-.31**</td>
<td>.03</td>
<td>-.17</td>
</tr>
<tr>
<td>C</td>
<td>.06</td>
<td>.02</td>
<td>-.39*</td>
<td>-.64**</td>
</tr>
<tr>
<td>7. Low average scaled scores</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>-.05</td>
<td>-.03</td>
<td>-.39**</td>
<td>-.20</td>
</tr>
<tr>
<td>C</td>
<td>-.17</td>
<td>-.15</td>
<td>.29</td>
<td>.21</td>
</tr>
<tr>
<td>8. RDS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>.01</td>
<td>-.02</td>
<td>-.10</td>
<td>-.13</td>
</tr>
<tr>
<td>C</td>
<td>.06</td>
<td>.09</td>
<td>.10</td>
<td>.10</td>
</tr>
<tr>
<td>10. Coding errors</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>-.29**</td>
<td>-.29**</td>
<td>-.34**</td>
<td>-.44**</td>
</tr>
<tr>
<td>C</td>
<td>-.57**</td>
<td>-.54**</td>
<td>-.64**</td>
<td>-.57**</td>
</tr>
<tr>
<td>11. Highly improbable responses</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>-.08</td>
<td>-.09</td>
<td>-.25*</td>
<td>-.09</td>
</tr>
<tr>
<td>C</td>
<td>-.04</td>
<td>-.06</td>
<td>-.29</td>
<td>-.22</td>
</tr>
<tr>
<td>12. Improbable pathology</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>-.15*</td>
<td>-.14*</td>
<td>.00</td>
<td>.06</td>
</tr>
<tr>
<td>C</td>
<td>-.03</td>
<td>-.04</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*$t$-Test is significant at the 0.05 level.

**$t$-Test is significant at the 0.01 level.

(r = -.20), personal information (r = .11), RDS (r = -.26), vocabulary > DigSpan (r = .22), and improbable pathology (constant) were not significant.

2.2.3. Best predictors, accounting for multicollinearity

Linear regression was used to identify the best predictors of effort, while accounting for multicollinearity. The remaining 11 rating scale items (after removal of the Mittenberg variable) were regressed on T2, T1 + T2, IR, and DR. Standardized regression coefficients for rating scale items and effort tests are shown in Table 4.

The regression of the rating scale items on T2 in adults produced a significant result ($R = .83$, $R^2_{adj} = .66$, $F_{(11,105)} = 21.7$, $p < .001$), with 66% of the (adjusted) variance in T2 accounted for by the 11 rating scale items. The best predictors may be seen from the table. The regression of the rating scale items on T1 + T2 in Adults produced a significant result ($R = .83$, $R^2_{adj} = .66$, $F_{(11,105)} = 21.2$, $p < .001$), with 66% of the (adjusted) variance in T1 + T2 accounted for by the 11 rating scale items. The regression of the rating scale items on T2 in children produced a
significant result ($R = .86$, $R^2_{adj} = .70$, $F_{(11,64)} = 17.0$, $p < .001$), with 70% of the (adjusted) variance in T2 accounted for by the 11 rating scale items. The regression of the rating scale items on T1 + T2 in children produced a significant result ($R = .86$, $R^2_{adj} = .70$, $F_{(11,64)} = 17.2$, $p < .001$), with 70% of the (adjusted) variance in T1 + T2 accounted for by the 11 rating scale items.

The regression of the rating scale items on IR in adults produced a significant result ($R = .82$, $R^2_{adj} = .59$, $F_{(11,43)} = 7.9$, $p < .001$), with 59% of the (adjusted) variance in IR accounted for by the 11 rating scale items. The regression of the rating scale items on DR in adults produced a significant result ($R = .86$, $R^2_{adj} = .66$, $F_{(11,43)} = 10.6$, $p < .001$), with 66% of the (adjusted) variance in DR accounted for by the 11 rating scale items. As the table depicts, the same items account for slightly more variance in DR than in IR, which now is chosen as the best variable to represent effort on the MSVT. As can be seen from the standardized coefficients (Beta), the only significant predictor while controlling for covariation was coding errors.

The regression of the rating scale items on IR in children produced a significant result ($R = .97$, $R^2_{adj} = .90$, $F_{(10,15)} = 23.0$, $p < .001$), with 90% of the (adjusted) variance in IR accounted for by the 10 rating scale items (Mittenberg variable removed; improbable pathology removed as a constant). The regression of the rating scale items on DR in children produced a significant result ($R = .97$, $R^2_{adj} = .91$, $F_{(10,15)} = 25.0$, $p < .001$), with 91% of the (adjusted) variance in IR accounted for by the 10 rating scale items (Mittenberg variable removed; improbable pathology removed as a constant).

### 2.2.4. Exploratory factor analysis

At this stage of the research, data reduction methods were exploratory, and principal components factor analyses with varimax rotations were utilized, as there was no need to assume the existence of hypothetical factors (Kim & Mueller, 1978). The Mittenberg variable was removed from each of these analyses due to its positive or noncorrelation with effort variables in all samples. Only factors with eigenvalues greater than one were extracted. Rating scale data from the two data sets, MSVT and TOMM, were combined to provide a larger $N$ for analysis (adult: $N = 194$; child: $N = 123$).

The initial analysis for adults yielded a two-factor solution. The first principal component accounted for about 45% of the variance, while the second component accounted for about 10% of the remaining unexplained variance. In the rotated solution, the two principal components account for about 32% and 22% of the variance, respectively. The component matrix shows that items sharing at least 50% their variance with the first component include simple arithmetic, simple sequences, miss items before start, low average scaled scores, and RDS. Highly improbable responses shares about 49% of its variance with the first component. Missed personal information (36%) and coding errors (25%) share a good proportion of variance with the second component. In the rotated solution, simple arithmetic and RDS shared at least half their variance with the first component, while missed personal information and coding errors shared at least half their variance with the second component. The communality estimates show that all the rating scale variables have a good proportion of variance that is accounted for by these two components, although improbable pathology has the lowest.

The initial solution for children yielded a three-factor solution, but the third component was largely founded upon one low-frequency variable—improbable pathology. The first principal component accounted for about 45% of the variance, while the second component accounted for about 12% of the remaining unexplained variance. The third component accounted for 9% of the remaining variance. In the rotated solution, the three principal components account for about 39%, 17%, and 11% of the variance, respectively. The component matrix shows that items sharing at least 50% their variance with the first component include simple arithmetic, simple sequences, Ganser-like answers, missed items before start, low average scaled scores, and highly improbable responses. RDS and coding errors share about 49% of their respective variance with the first component. Do not know president (50%), missed personal information (37%) and highly improbable pathology (29%) share a good proportion of variance with the second component. Only highly improbable pathology (46%) shares a good proportion of variance with the third component. In the rotated solution, missed items before start, low average scaled scores, coding errors, and improbable responses shared at least half their variance with the first component, while do not know president and missed personal information shared at least half their variance with the second component. Again, only highly improbably pathology shared at least half its variance with the third component. The communality estimates show that all the rating scale variables have a good proportion of variance accounted for by these three components, although RDS has the lowest.
Table 5
Correlations among the corrected rating scale total score and effort indicators on the TOMM and MSVT in adults (A) and children (C)

<table>
<thead>
<tr>
<th></th>
<th>TOMM</th>
<th></th>
<th>MSVT</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>T2</td>
<td>T1 + T2</td>
<td>RDS</td>
<td>IR</td>
<td>DR</td>
</tr>
<tr>
<td>Rating scale total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>−.78**</td>
<td>−.78**</td>
<td>−.75**</td>
<td>−.73**</td>
<td>−.80**</td>
</tr>
<tr>
<td>C</td>
<td>−.60**</td>
<td>−.62**</td>
<td>−.75**</td>
<td>−.89**</td>
<td>−.85**</td>
</tr>
</tbody>
</table>

**p < .001.

2.2.5. Convergent validity

Table 5 shows the correlations between the corrected rating scale total score (with Mittenberg variable now removed) and the various indicators of effort for adults and children. All correlations are significant (p < .001) and indicate a moderately strong amount of variance accounted for (>50%).

2.2.6. Sensitivity/specificity

The classification accuracy of this rating scale (Greve & Bianchini, 2004) was evaluated with respect to sensitivity, specificity, and predictive power. Sensitivity refers to the true positive (Hit) rate for a test. In this case, sensitivity refers to the number of malingerers who had a positive rating scale result divided by all malingerers. Thus, the sensitivity of this rating scale is the proportion of malingerers identified by this rating scale. Poor sensitivity (for a given cutoff) means that there are many false negative errors, which means that the rating scale is not detecting many malingerers. Specificity is the true negative rate—the number of non-malingerers rejected by this scale (negative test result) divided by the number of non-malingerers. Thus, specificity is the proportion of non-malingerers who obtained a negative test result on this rating scale. Poor specificity for a given cutoff yields a higher number of false positive errors—a lot of non-malingerers are getting tagged as malingerers.

Positive predictive power (+PP) is derived as the number of true positives divided by all positive results (true and false positives). It is an index of confidence that an individual test result is accurate. Negative predictive power (−PP) is derived as the number of true negatives divided by all negative results (true and false). It is a measure of the confidence that a rejection of the construct (e.g., “malingering”) is correct.

Because this was not a known-groups design (all claimants were seeking compensation), sensitivity cutoffs were initially set using high specificity (>85%). Ultimately, because the appeals process within DDS would provide a mislabeled individual a second chance, a lower specificity (77%) was tolerated. This error rate (.23) taken over two evaluations provides for an overall mislabeling error rate of p = .05.

Table 6 shows the relative sensitivity and specificity for different cutoff scores of the rating scale, with calculations of predictive power. We suggest a conservative approach be used in which the higher cutoffs at each level define the label for the individual (failed: adult, 5; child, 8; chance: adult, 7; child, 9; definite: adult, 12; child, 12). The various failure rates at each cutoff level are shown in Table 7.

It is interesting to note that the sensitivity, specificity, and predictive power of individual items on the rating scale is somewhat less than that of the total score (as would be predicted from lower correlations of individual items with the gold standard tests). For example, at Below-Chance levels on the TOMM (adults), a rating of three or more Ganser answers shows: sensitivity (.88), specificity (.69), +PP (.29), and −PP (.98). Greater than two coding errors shows: sensitivity (.56), specificity (.89), +PP (.43), and −PP (.93).

2.3. Relationship to IQ

Effort accounts for a large proportion of variance in cognitive test scores (Green et al., 2001). The relationship between the rating scale and IQ was examined to determine how much variance in IQ is accounted for by effort, as measured by the new rating scale.

In the adult TOMM sample, the correlation between the rating scale total (TOT1) and full scale IQ was $r = −.83 (F_{1,119} = 269; p < .001)$. The regression equation (FSIQ = −1.14TOT1 + 70.7) indicates that for every point increase in the rating scale, IQ drops a little over one point (starting at IQ = 71).
Table 6
Sensitivity, specificity, and predictive power (±) for various cut off scores on the rating scale, as identified by score category on the TOMM and MSVT

<table>
<thead>
<tr>
<th>Cut</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>P/N PP</th>
<th>Cut</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>P/N PP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Below chance&lt;sup&gt;a&lt;/sup&gt;</td>
<td>TOMM 10</td>
<td>88</td>
<td>82</td>
<td>44/98</td>
<td>MSVT 12</td>
<td>75</td>
<td>78</td>
</tr>
<tr>
<td></td>
<td>MSVT 12</td>
<td>100</td>
<td>79</td>
<td>41/100</td>
<td></td>
<td>100</td>
<td>88</td>
</tr>
<tr>
<td>Chance or below&lt;sup&gt;b&lt;/sup&gt;</td>
<td>TOMM 7</td>
<td>81</td>
<td>80</td>
<td>69/88</td>
<td>MSVT 9</td>
<td>72</td>
<td>79</td>
</tr>
<tr>
<td></td>
<td>MSVT 6</td>
<td>84</td>
<td>77</td>
<td>75/85</td>
<td></td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>At least failing&lt;sup&gt;c&lt;/sup&gt;</td>
<td>TOMM 5</td>
<td>71</td>
<td>77</td>
<td>80/67</td>
<td>MSVT 8</td>
<td>64</td>
<td>74</td>
</tr>
</tbody>
</table>

Note: We suggest a conservative approach be used in which the higher cutoffs at each level define the label for the individual (failed—adult, 5; child, 8; chance—adult, 7; child, 9; definite—adult, 12; child, 12).

<sup>a</sup> Note: TOMM T2 < 18; MSVT IR or DR < 30%
<sup>b</sup> Note: TOMM T2 < 33; MSVT IR or DR ≤ 70%
<sup>c</sup> Note: TOMM T2 < 45; MSVT IR or DR < 90%

Table 7
Failure rates for adults and children on effort tests (TOMM, MSVT) and the rating scale

<table>
<thead>
<tr>
<th>Effort tests</th>
<th>TOMM</th>
<th>MSVT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adult (%)</td>
<td>Child (%)</td>
<td>Adult (%)</td>
</tr>
<tr>
<td>Below chance</td>
<td>12.4</td>
<td>8.7</td>
</tr>
<tr>
<td>Chance or below</td>
<td>33.3</td>
<td>22.8</td>
</tr>
<tr>
<td>Failing</td>
<td>55.8</td>
<td>28.3</td>
</tr>
</tbody>
</table>

Rating scale

<table>
<thead>
<tr>
<th>TOMM</th>
<th>MSVT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adult (%)</td>
<td>Child (%)</td>
</tr>
<tr>
<td>Below chance</td>
<td>13</td>
</tr>
<tr>
<td>12</td>
<td>21.3</td>
</tr>
<tr>
<td>11</td>
<td>32.5</td>
</tr>
<tr>
<td>Chance or below</td>
<td>9</td>
</tr>
<tr>
<td>8</td>
<td>36.9</td>
</tr>
<tr>
<td>7</td>
<td>42.6</td>
</tr>
<tr>
<td>Failing</td>
<td>8</td>
</tr>
<tr>
<td>7</td>
<td>43.8</td>
</tr>
<tr>
<td>6</td>
<td>48.4</td>
</tr>
<tr>
<td>5</td>
<td>51.6</td>
</tr>
</tbody>
</table>
In the child TOMM sample, the correlation between TOT1 and FSIQ was $r = -0.75$ ($F_{(1,78)} = 99.7, p < .001$). The regression equation ($FSIQ = -1.62TOT1 + 74.6$) indicates a steeper drop in IQ for every point increase in the rating scale (starting at IQ = 75).

In the Adult MSVT sample, the correlation (TOT1 versus FSIQ) was $r = -0.83$ ($F_{(1,54)} = 115.7, p < .001$). The regression equation ($FSIQ = -0.96TOT1 + 70.7$) indicates a similar drop in IQ for each Rating Scale point as in the other adult sample (starting at IQ = 71).

In the Child MSVT sample, the correlation (TOT1 versus FSIQ) was $r = -0.75$ ($F_{(1,24)} = 29.8, p < .001$). The regression equation ($FSIQ = -1.67TOT1 + 74.0$) indicates a similar drop in IQ for every point increase in the rating scale (starting at IQ = 74) as in the other child sample.

2.4. Failure rates

Table 7 shows the failure rates at each level of effort for adults and children. There is good consistency, with failure of formal tests of effort at 50–60% levels for Adults on the TOMM and MSVT. The two samples—TOMM and MSVT—provide an adequate replication of these results. The DDS rating scale showed similar rates of failure, which is not surprising given that the cut scores were based upon the “gold standard” SVTs. Children failed effort testing at about 27–40% rates. When chance-level effort is considered, Adults fail at 33–47% rates, and children fail at 18–35% rates. With respect to definite malingering, adults fail at about 12–26% rates, while children fail at 7–30% rates. There was some tendency for effort failure to occur less frequently in children in the MSVT sample, which occurred later in time.

3. Discussion

This study describes the development of a social security disability malingering rating scale whose items all individually measure effort. Collectively, the total scale score is highly correlated with formal tests and indicators of effort in adults and children.

Items placed on this DDS malingering rating scale were chosen based upon published and experiential utility in adducing evidence for poor quality of effort within the PCE. These scale items individually and collectively share good amounts of variance with formal tests of effort. The DDS rating scale total score shows strong convergent validity using the TOMM and MSVT for validation. Reasonable sensitivity and specificity were obtained at different “doses” of effort so that the psychologist can determine from the total scale score whether the claimant is malingering at levels consistent with: (1) failed formal tests of effort; (2) chance-level scores on formal tests of effort; or (3) below chance levels (definite malingering). Cut-off scores for adults are >5, >7, and >12, respectively. Cut-offs for children are >8, >9, and >12, respectively. Given that there is an appeals process, the overall mislabeling rate corresponds to $p = .05$ or better at each level of poor effort. With this dose-response method, a large number of claimants were observed to fail at levels far below the norms for mentally retarded individuals (Green, 2004).

An extreme DDS case is useful to illustrate the working of the scale. A 31-year-old woman was seen for “mental retardation and nervous”. Her TOMM I and II scores were 13 and 2. She did not know the name of the US president, and she picked Lafcadio Hearn from a list. She missed all three simple arithmetic problems, and two simple sequences. She could not do a three-digit span, and her RDS score was 0. Her simple alternating figure was drawn so poorly as to strain credibility (improbable response). She told the examiner that the lobby plant “is crying” (improbable pathology). On the WAIS-III, she obtained no low average scaled scores, and she missed items before the start point on several subtests. She missed several coding items. She knew her birthdate and age. Her total score on the rating scale of 23 was in the definite malingering range, and was consistent with her TOMM scores of 13 and 2. It was noted that if she had been blindfolded on the TOMM, she would likely have obtained a better score.

Observations of exam behavior among definite malingerers suggested that other indicators of malingering would include having both vertical and horizontal reversals on the coding subtest, missing sample items on the TOMM, starting to point to the correct answer and then to the incorrect answer on the TOMM, and having 4 or more Ganser-like answers.

In previous studies, tests of effort have been highly correlated with neurocognitive test findings (Constantinou, Bauer, Ashendorf, Fisher, & McCaffrey, 2005; Green et al., 2001). In this study, the DDS rating scale measure of effort (whose items are not based upon the recognition memory paradigm) accounted for 69% of the variance in IQ in adult samples, and 56% in child samples. These findings were replicated, and the correlations held up consistently in adult and child TOMM and MSVT samples. The regression equations were also similar in both adult samples and both
child samples. With a DDS rating scale score of 0 (no evidence of malingering), a WAIS-age claimant is expected to have a full scale IQ score of 71, and a WISC-age claimant a full scale IQ score of about 74–75, thus confirming this to be a low-functioning sample. With every increase of a point in the DDS rating scale, a WAIS-age claimant would be expected to drop about one IQ point. A WISC-age claimant would be expected to drop about 1.65 IQ points for every rating scale point. The psychological examiner now has a tool to predict the expected IQ score, given full effort. For example, a WAIS-age claimant who scores a Full Scale IQ of 60 with a score of 10 on the DDS Rating Scale would have been expected with full effort to have obtained an IQ of 70. A WISC-age claimant with a score of 10 on the DDS Rating Scale would be expected to obtain a full scale IQ of 76. These relationships highlight the nature of beliefs among claimants about how to ensure benefits, for even with full effort the IQ scores would be in a range for consideration along with other evidence of disability.

The running debate over mental retardation and effort testing is at least partially addressed in the development of this rating scale. The question of whether low functioning individuals actually understand the task demands of a formal test of effort is essentially obviated by constructing an effort scale with multiple items that are built into the procedures used in the psychological consultative examination. We note that each retained indicator (item) has a significant correlation with two different gold standard SVTs. Of course, this is not unique and independent variance, but the sum of indicators into a DDS malingering rating scale total has an even larger correlation with the two gold standard SVTs. A score on the rating scale that indicates merely a failing effort grade is equivalent to failing at least 2–3 indicators of poor effort, which further lends support for multiple measures of a construct such as effort. Several more indicators are failed in chance-level and definite malingering.

The use of multiple measures is traditionally a more effective measure of any construct (Evans, Gage, & Chafetz, 1984), and it is critical to confirm the diagnosis of malingering with more than one measure, as malingering is multidimensional (Greiffenstein et al., 1994; Slick et al., 1999; Larrabee, 2000, 2005). Multi-strategy assessments have been effective in the detection of neurocognitive feigning (Bender & Rogers, 2004). Thus, failing the DDS malingering rating scale may occur through Ganser-like answers, an improbable response, and missing items before the start on several subtests of the WAIS-III, or it may be obtained through choosing Albert Brooks as the current US President (from a list), missing simple arithmetic problems, and missing several coding items (particularly with horizontal and vertical reversals). A definite malingerer, whose score on the Rating Scale is consistent with below chance performance on the TOMM or MSVT, may well get poor effort points on 5–6 separate indicators (items on the scale). Low functioning individuals who do so are thus scoring on several indicators, rendering it less likely that the claimant simply misunderstood the task demands.

Low functioning examinees may lack insight or judgment that would allow appreciation of their deficits, and they may be poorly motivated to perform to the best of their ability even in the absence of secondary gain (Greve, Bianchini, Mathias, Houston, & Crouch, 2003). While it is clear that at some level of mental retardation, the ability to perform and even understand the task demands of a formal test of effort will fail, it is also clear that low functioning claimants can and do mangle. Regression equations from the combined adult sample and the combined child sample indicate that 25.8% of adults and 9.5% of children who failed the rating scale would have obtained IQ scores below 70 even if they had obtained a score of 0 on the rating scale (no evidence of malingering). Their rating scale scores, obtained through multidimensional assessment, make it more probable that the diagnosis of malingering has validity.

On the other hand, 49 adults (TOMM 24.6 valid% and MSVT 29.3 valid%) and 43 children (TOMM 38.7 valid% and MSVT 26.9 valid%) who made perfect or near perfect scores on the TOMM II (48–50) or MSVT (95–100%) had IQ scores below 70, thus illustrating the ability for good quality of effort in low functioning claimants. For example, a 9-year-old girl with an IQ of 64 made perfect scores on the TOMM, and had a rating scale total score of 4. Thus, even children with low intellectual functioning are capable of performing in the normal range on these scales.

Hayes et al. (1997, 1998) have suggested a need for new norms or possibly simplified tasks or instructions when working with a mentally retarded examinee. They have suggested that the Rey 15-item test, dot counting test, and the M-test should not be used to detect malingering in defendants with mental retardation. Hurley and Deal (2006) tested mentally retarded individuals in residential settings who had no involvement with the criminal justice system and were not attempting to obtain compensation. They found that 41% of the sample obtained trial 2 scores less than 45, and that there was no relationship between trial 2 scores and IQ \( (r = -0.09) \). They suggested that because these subjects were not attempting to claim compensation a large proportion of MR individuals might simply fail effort testing without actually attempting to feign deficits. The lack of correlation with IQ was suggested to support this claim. On the other hand, there appears to have been no determination of how many subjects had been through the DDS system and were
currently accepting payments from Social Security. It is possible that a mixed group of subjects – some on SSDI and some not – could have flattened out the correlation with IQ while providing for failed scores on TOMM trial 2.

**Rohling (2004)** has argued that motivations for poor effort on the part of children would include feeling the need for accommodations for “disabilities”, and obtaining medications that have abuse and resale potential. Nevertheless, he believes that the concept of “secondary gain” as applied to adults may be too narrow for children, who often “play ill” to avoid negative consequences. Within the context of a DDS evaluation, there may be implicit or even explicit demands from parents to satisfy the parents’ needs for financial gain or stability. This essentially raises the issue of “malingering by proxy”, as enunciated in **Slick et al. (1999)** and described in **Lu and Boone (2002)**. Children may be unwittingly becoming part of a system of disability in which their grades are suffering in service to a financial reward to the family. Further evaluation of these problems would be most important because, if validated, there are enormous policy implications not only for DDS but also for the public school system (i.e., “no child left behind”).

Disincentives for effort testing are apparently intertwined with policy. DDS does not consider formal tests of effort to be “programmatically useful.” With regard to fees for psychologists, there is no category of compensation for formal tests of effort. If malingering is found prior to administering the Wechsler scale, the psychologist is supposed to stop and report only the mental status, thereby decreasing the fee. Moreover, there is the suggestion that budget outlays are based on the number of claimants in the catchment area, which would certainly decrease if findings of malingering were acted upon in a manner consistent with the discovery of fraudulent claims. Also, a finding of malingering can potentially harm DDS examiners financially: pay raises to examiners can be withheld for not processing cases in a timely fashion or having a pending case load that is too high, the latter being affected by cases of malingering which remain in the case load awaiting a new examination. Faced with disincentives, DDS examiners may be coaching claimants not to act as if they are too young. Apparently, the system of disincentives works to keep the system in place, with large budget increases every year.

Moreover, these are not isolated findings limited only to the region. It has become clear that Dr. Miller’s group has been working on the effort problem within the PCE over the same time period (since 2002) in Atlanta, GA **(Miller et al., 2006)**. They have used two different symptom validity tests (SVTs), CARB and the WMT, and have found results that are strikingly similar to ours, with over 50% of claimants failing conservative criteria on SVTs. Thus, without knowing about each other, two independent research efforts were launched at approximately the same time, in two different cities, using different symptom validity tests, and have both found high rates of effort problems in DDS psychological evaluations. We sincerely hope that this presents a challenge to current DDS policy.

**Acknowledgements**

We wish to thank Drs. Paul Lees-Haley, David Fox, and F. William Black for their early discussions and interest in this work. Dr. Lees-Haley provided further comment and critique. Dr. Glenn Larrabee has provided considerable encouragement for these data. Dr. Kevin Greve has also been most helpful in his discussions of this topic with me over the years. He provided a critical comment that led to the development of the “dose-response” analysis. We would also like to thank Jesse Lambert for his considerable help with data collection, databasing, and early discussions. We are grateful for the comments of the three reviewers, who made this a much better paper.

**References**


US Department of Health & Human Services, Social Security Administration (1994, revised 2006). Disability evaluation under social security. SSA publication no. 64-039.