Detecting memory performance validity with DETECTS: A computerized Performance Validity Test

Rui M. Paulo
College of Liberal Arts, Bath Spa University
School of Psychology, University of Minho

Pedro B. Albuquerque
School of Psychology, University of Minho

Abstract

Evaluating performance validity is essential in neuropsychological and forensic assessments. Nonetheless, most psychological assessment tests are unable to detect performance validity and other methods must be used for this purpose. A new Performance Validity Test (DETECTS – Memory Performance Validity Test) was developed with several characteristics which enhance test utility. Moreover, precise response time measurement was added to DETECTS. Two groups of participants (normative and simulator group) completed DETECTS and three memory tests from the Wechsler Memory Scale III. Simulators achieved considerably lower scores (hits) and higher response times in DETECTS compared with the normative group. All participants in the normative group were classified correctly and no simulator was classified as having legitimate memory deficits. Thus, DETECTS seems to be a valuable computerized Performance Validity Test with reduced application time and effective cut-off scores as well as high sensitivity, specificity, and positive and negative predictive power values. Lastly, response time may be a very useful measure for detecting memory malingering.

Keywords: Memory; Malingering; Detection; Performance Validity Test; Response Time

Professionals such as forensic psychologists and neuropsychologists often use psychological assessment tests to evaluate memory, personality, intelligence or verbal fluency in patients and interviewees. However, the utility of these tests is influenced by respondents' motivation to respond sincerely and perform well, which is in turn influenced by many other variables, such as respondents' goals (Simões, 2006), task payoff (Wang, Proctor, & Pick, 2009) and external incentives (Tan, Slick, Strauss, & Hultsch, 2002). Moreover, individuals are often motivated to malinger cognitive deficits or conceal their psychological functioning to obtain secondary gains such as monetary compensation. Patients with moderate brain injuries tend to exaggerate their conditions showing more intense and frequent symptoms than patients with severe brain injuries (Green, 2011; Green, Rohling, Lees-Haley, & Allen, 2001). Thus, evaluating malingering, i.e., intentional production of false or grossly exaggerated physical or psychological symptoms, usually motivated by external incentives (American Psychiatric Association, 2013), is critical during neuropsychological and forensic assessments (Boone, 2009; Greve & Bianchini, 2004; Lange, Pancholi, Bhagwat, Anderson-Barnes, & French, 2012; O'Bryant, Engel, Kleiner, Vasterling, & Black, 2007; O'Bryant & Lucas, 2006). Consequently, measuring and controlling the extent to which psychological test results reflect an attempt to deceive an assessor is a subject that is gaining increasing attention (Bauer, O'Bryant, Lynch, McCaffrey, & Fisher, 2007; Bush et al., 2005; Green et al., 2001; Greve & Bianchini, 2004; Iverson, 2003; Whiteside, Dumbar-Mayer, & Waters, 2009).

Patients can malinger very different symptoms, such as pain, disorientation, depression, lack of concentration, personality changes, or memory loss (Iverson, 2003; Leppma, Long, Smith, & Lassiter, 2017). Memory malingering is commonly used, for instance, to obtain

monetary compensation (Oorsouw & Merckelbach, 2010; Porter & Woodworth, 2007; Simões, 2006). Today, there are several methods for evaluating memory performance validity. Performance Validity Tests (PVTs) are one of the most commonly used methods. PVTs are usually forced-choice recognition tests, in which a patient, in the presence of a foil, must identify the stimuli which have previously been presented (Blaskewitz, Merten, & Kathmann, 2008; Gervais, Rohling, Green, & Ford, 2004). Although PVTs have been widely studied (Armistead-Jehle, Lange, & Green, 2017; Bianchini, Mathias, & Greve, 2001; Erdodi et al., 2017; Gast & Hart, 2010; Reslan, & Axelrod, 2017; Simões, 2006), a few research issues have emerged over the years.

Firstly, when most of these tests were developed, they were all considered Symptom Validity Tests (SVTs). However, several authors (Larrabee, 2012) recently suggested that researchers should distinguish between Symptom Validity Tests (SVTs) and Performance Validity Tests (PVTs). SVTs are tests which assess the validity of symptomatic complaints on a self-report measure, while PVTs assess the validity of a patient's performance. Secondly, the term 'effort' may have several meanings (Bigler, 2012; Bigler, 2014; Van Dyke, Millis, Axelrod, & Hanks, 2013). Although earlier PVTs (at the time called SVTs) were considered to measure poor effort (Tombaugh, 1996), nowadays, PVTs are considered to measure performance validity because, as stated above, the term 'effort' might have several meanings. Thirdly, there are now two major research designs to study memory malingering: the simulation design, where non-injured subjects (simulators) are instructed to feign deficits, and the 'known groups' or 'criterion groups' design, where a group of litigating subjects (probable malingerers) are evaluated. Lastly, there are at least four key aspects researchers need to consider in evaluating PVT utility: specificity, sensitivity, positive predictive power, and negative predictive power (Greve &

Bianchini, 2004; Lippa, Lange, Bhagwat, & French, 2017). Specificity refers to the percentage of individuals who do not have a condition of interest and are correctly classified. Therefore, low specificity is related to a high number of false positives (e.g., people with real memory impairments, who are classified as having a non-credible performance). Specificity is related to positive predictive power (PP+), which is the probability of an evaluated subject having a condition of interest when identified as having such condition. These are both important concepts, particularly because a false positive diagnosis can have serious consequences for a person's life. Therefore, researchers sometimes place more relevance on these criteria than on sensitivity and negative predictive power (Iverson, 2007). Sensitivity refers to the percentage of individuals with a condition of interest who are properly classified. This concept is related to negative predictive power (PP-), which is the probability of an evaluated subject not having a condition of interest when identified as having such condition (e.g., a simulator diagnosed with real memory problems).

Moreover, a PVT should meet two basic requirements. First, it should evaluate performance validity instead of other variables, such as, intelligence, memory capacity, or even brain damage, which may all be related to retrieval capacity (Tombaugh, 1996). Second, unlike most psychological assessment tests, PVTs must have low face validity, in other words, they should not be easily identified as a performance validity test (e.g., a memory PVT must appear to be measuring memory capacity). Nonetheless, trained subjects may identify a PVT more easily and malinger in a more sophisticated manner. Thus, the effect of different types of training has been assessed in several studies (Powell et al., 2004; Weinborn, Woods, Nulsen, & Leighton, 2012). Moreover, developing, testing and using different PVTs is essential because the greater the number of PVTs available, the less likely an individual is to have time and ability to study

and recognize them (Bianchini et al., 2001; Chafetz, 2011; Haber & Fichtenberg, 2006; Oorsouw & Merckelbach, 2010). Lastly, some PVTs (e.g., TOMM) use drawings as stimuli. This can be an advantage as they may require fewer adjustments when applied to different populations, unlike tests that use words as stimuli, which require stimuli translation and adaptation. (MacAllister, Nakhutina, Bender, Karantzoulis, & Carlson, 2009; Powell, Gfeller, Hendricks, & Sharland, 2004; Simões, 2006; Tombaugh, 1996).

Computerized PVTs are becoming more common and can have several advantages over traditional PVTs (Vanderslice-Barr, Miele, Jardin, & McCaffrey, 2011) such as precise measuring of response time (Bianchini et al., 2001; Haines & Norris, 1995; Vagnini, Berry, Clark, & Jiang, 2008; Willison & Tombaugh, 2006). Measuring response time can be very important, since a common strategy used by malingerers and simulators is to intentionally respond more slowly. Also, malingerers/simulators may need more time to respond, as they have to decide what the correct response is and whether they will provide a correct or an incorrect response (Tan et al., 2002; Willison & Tombaugh, 2006). Thus, measuring response time may increase test discriminative power. Different computerized PVTs are available today, such as the Nonverbal Medical Symptom Validity Test (Green, 2008), the Medical Symptom Validity Test (Green, 2004), the Word Memory Test (Green, 2003), the Computerized Assessment of Response Bias (Allen, Conder, Green, & Cox, 1997), and the computerized version of the Test of Memory Malingering (Tombaugh, 1996); the latter will be addressed below.

The Test of Memory Malingering - TOMM (Tombaugh, 1996) is a widely studied and used PVT (Bauer et al., 2007; Bianchini et al., 2001; Etherton, Bianchini, Greve, & Ciota, 2005; Greiffenstein, Greve, Bianchini, & Baker, 2008; O'Bryant, Finlay, & O'Jile, 2007; Slick, Tan, Strauss, & Hultsch, 2004). This test has a high correct classification rate, high sensitivity, high

specificity, high positive predictive power, and high negative predictive power (Fazio, Denning, & Denney, 2017; Powell et al., 2004; Tombaugh, 1996). Several characteristics of this instrument may explain why it usually allows accurate detection of memory malingering. Firstly, TOMM is not sensitive to variables such as age or education and, because of the number and nature of the stimuli, it looks far more difficult than it actually is, encouraging malingerers and simulators to exhibit low performance (Batt, Shores, & Chekaluk, 2008; Blaskewitz, et al., 2008; Gast & Hart, 2010; Greve et al., 2006; Simon, 2007). Secondly, the structure of this test is very similar to a legitimate memory capacity test, and feedback is provided immediately after each response. This allows subjects who are purposely trying to exhibit low performance to have control over the number of wrong responses they provide, thereby encouraging them to exhibit low performance (Haines & Norris, 1995). Lastly, patients with brain injuries (or other disorders, such as affective and psychotic disorders, chronic pain, or epilepsy) usually have a high capacity for storing and recognizing common pictures. Consequently, this test is not sensitive to these impairments and does not classify such patients as having non-credible performance (Duncan, 2005; Etherton et al., 2005; Iverson, LePage, Koehler, Shojania, & Badii, 2007; MacAllister et al., 2009; O'Bryant et al., 2007; Tombaugh, 1996).

Current Study

In this study, a new PVT based on TOMM was developed with several new characteristics (e.g., precise response time measurement) to enhance overall utility and specificity, sensitivity, PP+, and PP-. Our aim was to develop a short PVT suitable for distinguishing simulators who have been previously warned and informed about performance validity testing. Lastly, the impact of PVT administration order on performance validity

evaluation was studied.

Thus, a new PVT (DETECTS – Memory Performance Validity Test) was developed. As addressed in the method section, DETECTS was used with three memory tests and applied to two different groups: a control group and a simulator group. Hits and response times were measured and trial and test administration order were controlled. Lastly, proper cut-off scores were defined.

Method

Participants

A total of 65 Caucasian psychology students from Portugal took part in this study. Participants were recruited through a course credit program by which students earned credits by participating in research experiments implemented in our University. Participants were randomly assigned to one of two groups: a normative group and a simulator group. The normative group had 41 participants, 37 females and four males, with an age range from 18 to 28 years (M = 19.71, SD = 1.93). The simulator group had 24 participants, 21 females and three males with an age range from 21 to 35 years (M = 23.33, SD = 3.66).

Design

A between-participants design was used with the group of participants as an independent variable with two levels: normative group and simulator group. Correct responses and reaction times on DETECTS were measured in hits and milliseconds, respectively. Wechsler Memory Scale III memory subtests were scored according to the WMS III manual (Wechsler, 1997).

Materials

To conduct this study a new PVT (DETECTS) and a computerized version of three Wechsler Memory Scale III memory subtests was developed (Faces I, Visual Reproduction and Spatial Span).

DETECTS. This test is a fully computerized PVT consisting of two trials lasting approximately 15 minutes (total). In each trial, the same 50 drawings of common objects are presented randomly and immediately after a forced-choice recognition test composed of 50 pairs of drawings. Each pair contains a previously shown drawing and a new one. Thus, participants must choose the drawing presented before (target). Moreover, the only difference between trials 1 and 2 concerns foils: each trial has a different foil assigned to each target drawing.

DETECTS was programmed with Superlab 4.5 (Cedrus Corporation, San Pedro, CA, USA). Participants responded to the test by pressing one of two computer keyboard keys assigned to each drawing (A or B), and received immediate visual feedback (Right! or Wrong!). All images presented were chosen from the Snodgrass and Vanderwart database (1980), previously used in similar tasks (Nishimoto, Miyawaki, Ueda, Une, & Takahashi, 2005). To ensure a very low level of difficulty, all drawings were classified into narrow categories (e.g., mammals, insects, birds, aromatic plants, decorative plants, fruits, etc.) and no drawing was ever in the same category as the remaining stimuli. Foils were always in a category other than the target. However, the two foils presented for the same target were always in the same category. For example, the target image ‘Turtle’ was presented during the trial 1 recognition test with the foil ‘American Football *Helmet’*, and presented during the trial 2 recognition test with the foil ‘American Football *Ball’*, which means both foils paired with ‘Turtle’ belonged to the American Football category. All images were adjusted to have the same visual characteristics (e.g., all images were presented at the centre of the screen with the same contrast and dimensions).

Wechsler Memory Scale III subtests. Three memory subtests from the Wechsler Memory Scale III (Wechsler, 1997) were used: Faces I, Visual Reproduction and Spatial Span. All subtests were computerized with Superlab 4.5 and all subtest features were maintained exactly the same (e.g., in terms of stimuli, instructions, training trials, presentation times, and scoring). All subtests were acquired previously by our department, which was granted permission for clinical and research purposes. Faces I is a non-verbal visual memory test in which several faces are presented. Immediately afterward, participants must respond to a recognition test where they are asked to state which faces have, or have not, been previously presented. Visual Reproduction is a non-verbal visual memory test in which participants must perform a visuographic reconstruction task after seeing different geometric figures. This test was only partially computerized, i.e., the visuographic reconstruction task was performed with a regular pencil and paper to avoid any influence of digital drawing software, such as task difficulty. Spatial Span evaluates components of working memory by asking participants to use their hands to repeat several sequences, in direct and reverse order, previously executed by the researcher using a set of plastic blocks. The computerized version was identical (e.g., all sequences were from the original test), although these were presented, and responses were provided, on the computer screen. These three tests were chosen because they evaluate different types of memory (e.g., non-verbal visual memory and working memory) and use different kinds of stimuli (e.g., faces, geometric figures, etc.). Moreover, these tests already have standard application and scoring norms.

Procedure

Participants gave their informed consent to participate in this study and were assessed

individually in a soundproof booth. The normative group was informed they would be given four tests to evaluate their memory capacity and were asked to give their best effort. Application order was counterbalanced using four experimental conditions; DETECTS would be immediately preceded, and followed, by each WMS-III subtest. Trial order (1 or 2) was also counterbalanced: half of the participants in each test order condition responded first to DETECTS trial 1, followed by trial 2, and another half responded to DETECTS trials in the opposite order. Therefore, phase 1 was labelled as the first trial to which a given participant responded first and phase 2 was labelled as the second trial to which a given participant responded second.

A similar procedure was implemented for the simulator group, who were given the following instructions: *Imagine you were involved in a car accident where you suffered a minor head trauma but you have not experienced any cognitive or memory problems. To obtain a large compensation from your insurance company you want to simulate memory problems. Your insurance company has scheduled you for a psychological assessment to see if your memory problems are real. Imagine this assessment is happening today and I am the expert who will evaluate you. Your goal is to make me believe you have real memory problems when I look at your test results*. Since we wanted to have an informed group of simulators, a 15-slide presentation was shown to this group of participants before responding to the tests. These slides were designed to be similar to the information a real malingerer would find online, although more accurate and informative because our participants would have less time to research the subject. These slides included concise information about memory problems (e.g., amnesia), common causes and consequences of these problems on patients' memory and daily routines, what malingering is, and how it can be evaluated during forensic or neuropsychological assessments. Also addressed in this presentation were different techniques used to detect memory malingering,

including what a PVT is and how it works.

Correct responses (hits), incorrect responses (false alarms), and reaction times to DETECTS were automatically recorded by Superlab 4.5. WMS-III subtests were scored according to the authors' norms (Wechsler, 1997). The entire procedure usually took 40 minutes.

Results

An alpha level of .05 was used for all statistical tests. Exploratory data analysis was performed. Based on the recommendations by Fife-Schaw (2006), whenever violations of the assumptions were found, both parametric and equivalent non-parametric tests were conducted. As the conclusions drawn from both tests were always identical, only parametric tests were reported.

DETECTS

Administration Order. The serial position of a PVT within a test battery may affect performance (Ryan, Glass, Hinds, & Brown, 2010). Therefore, the influence of DETECTS administration order (four administration orders) on participants' performance (hits) was first tested. Two independent one-way ANOVAs were conducted, one for each group of participants. No administration order effect on participants' performance (hits) was found for either the normative group, F (3, 37) = 1.49, p = .232, η_p^2 = .11, or the simulator group, F (3, 20) = 1.11, p = .367, η_p^2 = .14. Therefore, administration order effects were not considered in further analysis.

Accuracy. A 2 X 2 ANOVA was conducted to see if group of participants (normative or simulator) and DETECTS phase (one or two) had an effect on the number of hits in DETECTS.

A main effect for group of participants was found, F (1, 63) = 135.63, p < .001, η_p^2 = .68, with simulators presenting fewer hits (M = 30.48, SD = 10.30, 95% CI [27.92, 33.04]) than participants in the normative group (M = 49.23, SD = .90, 95% CI [47.28, 51.19]). No main effect

of DETECTS phase, F (1, 63) = .80, p = .375, η_p^2 = .01, or interaction between DETECTS phase and group of participants, F (1, 63) = 3.05, p = .086, η_p^2 = .05, was found (see Table 1).

Insert Table 1

Response times. A 2 X 2 X 2 ANOVA was used to analyze whether the group of participants (normative or simulator), DETECTS phase (1 and 2), and response type (hit or false alarm) had an effect on participants' response time in DETECTS.

A main effect of group of participants on response time (measured in milliseconds) was found, F (1, 28) = 7.59, p = .01, η_p^2 = .21 (see Table 2). Results showed participants in the normative group (M = 1385, SD = 311, 95% CI [921, 1849]) responded faster than participants in the simulator group (M = 2114, SD = 718, 95% CI [1834, 2394]). There was also a main effect of response type (hit or false alarm), F (1, 28) = 20.75, p < .001, η_p^2 = .43. Participants were faster when responding correctly (M = 1601, SD = 660, 95% CI [1348, 1854]) than when responding incorrectly (M = 1899, SD = 660, 95% CI [1595, 2202]). No main effect of DETECTS phase on response time was found, F (1, 28) = 3.62, p = .068, η_p^2 = .11.

An interaction effect between response type and group of participants was found, F (1, 28) = 7.88, p = .009, η_p^2 = .22. Simulators exhibited slow response times for hits (M = 2057, SD = 769, 95% CI [1796, 2318]) and false alarms (M = 2172, SD = 756, 95% CI [1858, 2485]). As anticipated, the normative group exhibited faster and different mean response times for hits (M = 1144, SD = 262, 95% CI [711, 1578]) and false alarms (M = 1626, SD = 939, 95% CI [1106, 2145]).

Insert Table 2

There was no interaction effect between DETECTS phase and group of participants, F (1, 28) = .20, p = .662, η_p^2 = .01, nor between DETECTS phase and response type, F (1, 28) = 4, p = .055, η_p^2 = .13. Lastly, no triple interaction between DETECTS phase, group of participants, and response type was found, F (1, 28) = .05, p = .828, η_p^2 = .002.

Cut-off Scores. Next, specificity, sensitivity, PP+, and PP- for several possible cut-off scores were calculated. Table 3 and 4 show values for each of these parameters when considering different cut-off scores for both DETECTS phases.

Insert Table 3 and Table 4

Based on the values described in Table 3 and Table 4, a cut-off score of 44 hits for Phase 1 and 47 hits for Phase 2 was established. This means that anyone with an equal or inferior value of hits in both DETECTS phases is considered to have a non-credible performance.

Diagnostic Confirmation. After establishing cut-off scores, they were applied to the entire sample and each participant was reclassified accordingly. This was a blind procedure, with participants' group information hidden from the database. This procedure allowed us to further evaluate DETECTS' ability to detect simulators within our sample.

Only one participant was incorrectly reclassified (0.015% of the sample): a simulator who was incorrectly reclassified as a normative group participant since he had a score of 50 hits (maximum score) in both DETECTS phases. After assessing this participant's responses to the WMS-III subtests, we found he obtained high scores not only in DETECTS but also in all other

WMS-III subtests (Spatial Span: 12 points, Faces: 35 points, Visual Reproduction: 83 points). Thus, he failed to successfully simulate memory problems, achieving results very similar to the normative group (high scores in all applied tests).

WMS-III Memory Subtests Performance. Lastly, three Student's t-tests were used to compare the normative and simulator group regarding their scores in all WMS-III memory subtests. Regarding the Spatial Span subtest, simulators achieved lower scores (M = 8.25, SD = 5.02, 95% CI [6.13, 10.37]) in comparison with the normative group (M = 16.39, DP = 2.33, 95% CI [15.65, 17.13]), t (63) = 8.91, $p < .001$, d = 2.08. In the Faces I subtest, simulators also obtained lower scores (M = 28.50, DP = 7.56, 95% CI [25.31, 31.69]) than the normative group (M = 38.02, DP = 4.14, 95% CI [36.72, 39.33]), t (63) = 6.57, $p < .001$, d = 1.56. Similar results were found for the Visual Reproduction subtest, i.e., simulators (M = 59.17, DP = 20.16, 95% CI [50.65, 67.68]) obtained lower scores compared with the normative group (M = 90.98, DP = 9.29, 95% CI [88.04, 93.91]), t *(63)* = 8.68, $p < .001$, d = 2.02. In sum, simulators achieved considerably lower scores in all memory tasks.

Discussion

Memory malingering is frequent during forensic and neuropsychological assessments. Therefore, developing effective Performance Validity Tests (PVTs) can be essential in this field; the more PVTs made available to professionals, the less likely it is for a malingerer to have the time and ability to study and recognize all the tests (Bianchini et al., 2001; Chafetz, 2011; Haber & Fichtenberg, 2006; Oorsouw & Merckelbach, 2010). The present study describes the development of an efficient and time-saving PVT and the impact of participant response time and test administration order on performance validity testing.

Measuring hits can be crucial for evaluating performance validity when using certain PVTs. The simulator group exhibited a considerably lower hit rate in DETECTS in comparison with the normative group. As hypothesized, possibly due to its characteristics (e.g., high expected difficulty and low actual difficulty), DETECTS was able to differentiate simulators from normative respondents. Immediate feedback might have influenced this result, allowing simulators to have control over their performance, thus exhibiting lower scores (Tombaugh, 1996). Moreover, using a two-phase test such as DETECTS may improve performance validity assessments, allowing for cross-trial analysis. Whereas simulators obtained low scores in both DETECTS phases, participants from the normative group exhibited high scores, which are very close to the maximum possible score (ceiling effect), in both DETECTS phases. Moreover, these two groups showed distinct response patterns. For instance, while 42% of simulators achieved higher scores in DETECTS phase 1 in comparison with DETECTS phase 2, 93% of participants in the normative group increased or maintained their performance in DETECTS phase 2. The lowest score a participant from the normative group exhibited in DETECTS phase 1 was 45 hits, and in DETECTS phase 2, 48 hits. Both scores were considerably higher than the scores obtained by simulators. In sum, even though all simulators were previously informed about memory problems, malingering behaviour, and about the different techniques which can be used to detect memory malingering (e.g., PVTs), DETECTS made it possible to clearly differentiate between simulators and non-simulators.

Response time analysis may also be another very important measure for performance validity testing (Willison & Tombaugh, 2006). Our study supports this premise, as participants from the normative group exhibited faster response times than participants from the simulator group. Research suggests that one strategy used by simulators is to respond more slowly than

usual to convince the researcher that their slow performance is due to their cognitive impairments (Tan et al., 2002; Willison & Tombaugh, 2006). This is consistent with our results, which showed simulators exhibited longer response times. Moreover, participants from the normative group and simulators showed different response time patterns depending on the type of response provided. Participants in the normative group were faster when providing correct responses, possibly because they were unsure of their wrong answers and this decision-making process under uncertainty may take additional time. Simulators, however, provided incorrect responses as fast as correct responses because, as stated above, their simulation strategy may be to purposely try to always respond slowly (Vagnini et al., 2008). Thus, the simulators' response time was intentionally higher than the time they would actually need to discriminate between a correct and an incorrect response regardless of accuracy. Lastly, a tendency (marginally significant effect) for participants to provide faster responses in DETECTS phase 2 in comparison with DETECTS phase 1 was found, particularly when providing incorrect responses. This may be explained by practice. However, these differences were not significant and should be further tested in upcoming studies. In sum, while our study does not show that response times can be singly used to evaluate performance validity, they may be very helpful for confirming a diagnosis, given the different mean response times and response time patterns exhibited in DETECTS by participants in the normative group and the simulators group. This may be an advantage over other PVTs (Bianchini et al., 2001; Bigler, 2014; Haines & Norris, 1995; Vagnini et al., 2008; Willison & Tombaugh, 2006).

Well-defined cut-off scores are central to any PVT, allowing technicians to objectively interpret test results and different technicians to make similar diagnostics (Greve & Bianchini, 2004; Rickards, Cranston, Touradji, & Bechtold, 2017). A cut-off score of 44 hits for phase 1 and

47 hits for phase 2 is suggested for DETECTS. Thus, technicians are advised to determine non-credible performance only when respondents achieve a number of hits equal or lower than the cut-off scores in both DETECTS phases. This very conservative criterion was adopted because, as stated by Greve and Bianchini (2004), maintaining very high levels of specificity (proportion of non-simulators correctly classified) and PP+ (probability of an evaluated subject being a simulator when evaluated as one) is essential. With the cut-off scores stated above, these parameters were maintained at 100%, avoiding false positive errors. Nonetheless, high levels of sensitivity (proportion of simulators correctly classified) and PP- (probability of an evaluated subject not being a simulator when classified as such) were also maintained, avoiding false negative errors.

As stated above, after establishing the cut-off scores, they were applied to the entire sample. With this procedure, only one participant, a simulator who presented a score of 50 hits in both DETECTS phases, was misclassified. Nonetheless, unlike other simulators, this participant obtained results similar to the normative group in all memory tasks. Therefore, in a neuropsychological or forensic assessment, he would be diagnosed as having no memory problem and his simulation attempt would be ineffective. Furthermore, none of the participants from the normative group was classified as a simulator and none of the simulators was classified as having real memory problems since none of the simulators presented high scores in DETECTS and low scores in the memory tasks.

Limitations and Future Directions

This study can be an important first step for studying DETECTS effectiveness. Given DETECTS' ability to discriminate between simulators and participants from the normative group

(100% efficacy), studies with clinical samples are now recommended. Although DETECTS is substantially based on TOMM (Tombaugh, 1996), which has proved to be effective with clinical samples, and although DETECTS was used with a very small clinical sample where all patients performed above the established cut-off scores (see Appendix 1), large clinical trials are now necessary. Lastly, DETECTS might have some minor disadvantages inherent to most computerized tests, such as patients' inability to change their response, or examiners' inability to pause the test if a participant looks away. Although these situations never occurred during our study, they might be more frequent, for instance, in patients with attention deficits.

Conclusions and Practical Implications

Psychological assessment tests can be vulnerable to memory malingering, a prevalent phenomenon in forensic and neuropsychological assessments. Thus, developing many effective Performance Validity Tests (PVT) can be crucial in this field. We developed a new PVT: DETECTS. When applied with three WMS-III tests, DETECTS was capable of perfectly differentiating our sample, even when simulators were warned and taught about memory problems, malingering behaviour and the different techniques used to detect memory malingering (e.g., PVTs). Lastly, DETECTS may have several advantages over other performance validity detection strategies and PVTs: (1) when used with the chosen memory tests, none of the simulators was classified as having legitimate memory problems; (2) DETECTS was not identified as a PVT by informed simulators; (3) DETECTS has a reduced application time and time constraints are somewhat frequent during forensic and neuropsychological evaluations (Fazio et al., 2017); and (4) DETECTS is a computerized test that is rigorous, easy to use, accurately measures response time, which may be important for confirming a diagnosis, and can

also be easily made available for open use (please contact the main author to obtain this PVT).

References

Allen, L. M., Conder, R. L., Green, P., & Cox, D. R. (1997). *CARB 97: Manual for the Computerized Assessment of Response Bias*. Durham, NC: CogniSyst.

American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders: DSM-5*. Washington, DC: American Psychiatric Publishing.

Armistead-Jehle, P., Lange, B. J., & Green, P. (2017). Comparison of neuropsychological and balance performance validity testing. *Applied Neuropsychology: Adult, 24*, 190-197. http://dx.doi.org/10.1080/23279095.2015.1132219

Batt, K., Shores, E. A., & Chekaluk, E. (2008). The effect of distraction on the word memory test and test of memory malingering performance in patients with a severe brain injury. *Journal of the International Neuropsychological Society: JINS, 14*, 1074-1080. doi:10.1017/S135561770808137X.

Bauer, L., O'Bryant, S. E., Lynch, J. K., McCaffrey, R. J., & Fisher, J. M. (2007). Examining the test of memory malingering trial 1 and word memory test immediate recognition as screening tools for insufficient effort. *Assessment, 14*, 215-222. doi:10.1177/1073191106297617

Bianchini, K., Mathias, C., & Greve, K. (2001). Symptom validity testing: A critical review. T*he Clinical Neuropsychologist*, *15*, 19-45. doi:10.1076/clin.15.1.19.1907

Bigler, E. D. (2012). Symptom validity testing, effort, and neuropsychological assessment. *Journal of the International Neuropsychological Society*, *18*, 632-640. https://doi.org/10.1017/S1355617712000252

Bigler, E. D. (2014). Effort, symptom validity testing, performance validity testing and traumatic brain injury. *Brain Injury, 28,* 1623-1638. http://dx.doi.org/10.3109/02699052.2014.947627

Blaskewitz, N., Merten, T., & Kathmann, N. (2008). Performance of children on symptom validity tests: TOMM, MSVT, and FIT. *Archives of Clinical Neuropsychology, 23,* 379-391. doi:10.1016/j.acn.2008.01.008

Boone, K. B. (2009). The need for continuous and comprehensive sampling of effort/response bias during neuropsychological examinations. *The Clinical Neuropsychologist, 23,* 729–741. doi: 10.1080/13854040802427803

Bush, S. S., Ruff, R. M., Tröster, A. I., Barth, J. T., Koffler, S. P., Pliskin, N. H., Reynolds, C. R., & Silver, C. H. (2005). Symptom validity assessment: practice issues and medical necessity NAN policy & planning committee. *Archives of Clinical Neuropsychology, 20,* 419-426. doi:10.1016/j.acn.2005.02.002

Chafetz, M. (2011). Reducing the probability of false positives in malingering detection of social security disability claimants. *The Clinical Neuropsychologist, 25,* 1239-1252. doi:10.1080/13854046.2011.586785

Duncan, A. (2005). The impact of cognitive and psychiatric impairment of psychotic disorders on the test of memory malingering (TOMM). *Assessment, 12,* 123-129. doi:10.1177/1073191105275512

Erdodi, L. A., Tyson, B. T., Abeare, C. A., Zuccato, B. G., Rai, J. K., Seke, K. R., Sagar, S., & Roth, R. M. (2017). Utility of critical items within the recognition memory test and word choice test. *Applied Neuropsychology: Adult.* http://dx.doi.org/10.1080/23279095.2017.1298600

Etherton, J. L., Bianchini, K. J., Greve, K. W., & Ciota, M. A. (2005). Test of memory malingering performance is unaffected by laboratory-induced pain: implications for clinical use. *Archives of Clinical Neuropsycholog, 20,* 375-384. doi: 10.1016/j.acn.2004.09.007

Fazio, R. L., Denning, J. H., & Denney, R. L. (2017). TOMM trial 1 as a performance validity indicator in a criminal forensic sample. *The Clinical Neuropsychologist, 31,* 251-267. http://dx.doi.org/10.1080/13854046.2016.1213316

Fife-Schaw, C. (2006). Levels of measurement. In G. M. Breakwell, S. Hammond, C. Fife-Schaw, & J. A. Smith (Eds.), *Research Methods in Psychology*: 3rd ed. (pp. 50-63). London: Sage.

Gast, J., & Hart, K. (2010). The performance of juvenile offenders on the test of memory malingering. *Journal of Forensic Psychology Practice, 10,* 53-68. http://dx.doi.org/10.1080/15228930903173062

Gervais, R. O., Rohling, M. L., Green, P., & Ford, W. (2004). A comparison of WMT, CARB, and TOMM failure rates in non-head-injury disability claimants. *Archives of Clinical Neuropsychology, 19,* 475-487. http://doi.org/10.1016/j.acn.2003.05.001

Green, P. (2003). *Word Memory Test for Windows: User's manual and program.* Canada: Green's Publishing.

Green, P. (2004). *Medical Symptom Validity Test (MSVT) for Microsoft Windows: User's Manual*. Canada: Green's Publishing.

Green, P. (2008). *Manual for the nonverbal medical symptom validity test for windows*. Canada: Green’s Publishing.

Green, P. (2011). Comparison between the test of memory malingering (TOMM) and the

nonverbal medical symptom validity test (NV-MSVT) in adults with disability claims. *Applied Neuropsychology, 18,* 18-26. doi:10.1080/09084282.2010.523365

Green, P., Rohling, M., Lees-Haley, P., & Allen, L. (2001). Effort has a greater effect on test scores than severe brain injury in compensation claimants. *Brain Injury, 15,* 1045-1060. doi:10.1080/02699050110088254

Greiffenstein, M. F., Greve, K. W., Bianchini, K. J., & Baker, W. J. (2008). Test of memory malingering and word memory test: a new comparison of failure concordance rates. *Archives of Clinical Neuropsychology, 23,* 801-807. http://doi.org/10.1016/j.acn.2008.07.005

Greve, K. W., & Bianchini, K. J. (2004). Setting empirical cut-offs on psychometric indicators of negative response bias: a methodological commentary with recommendations. *Archives of Clinical Neuropsychology, 19,* 533-541. doi: 10.1016/j.acn.2003.08.002

Greve, K. W., Bianchini, K. J., Black, F. W., Heinly, M. T., Love, J. M., Swift, D. A., & Ciota, M. (2006). Classification accuracy of the test of memory malingering in persons reporting exposure to environmental and industrial toxins: Results of a known-groups analysis. *Archives of Clinical Neuropsychology, 21,* 439-448. http://doi.org/10.1016/j.acn.2006.06.004

Haber, A. H., & Fichtenberg, N. L. (2006). Replication of the test of memory malingering (TOMM) in a traumatic brain injury and head trauma sample. *The Clinical Neuropsychologist, 20,* 524-532. doi:10.1080/13854040590967595

Haines, M. E., & Norris, M. P. (1995). Detecting the malingering of cognitive deficits: An update. *Neuropsychology Review, 5,* 125-148. doi:10.1007/BF02208438

Iverson, G. L. (2003). Detecting malingering in civil forensic evaluations. In A. M. Horton Jr. &

L. C. Hartlage (Eds.), *Handbook of Forensic Neuropsychology* (pp. 137-177). New York: Springer.

Iverson, G. L. (2007). Identifying exaggeration and malingering. *Pain Practice*, *7*, 94-102. doi:10.1111/j.1533-2500.2007.00116.x

Iverson, G. L., Le Page, J., Koehler, B. E., Shojania, K., & Badii, M. (2007). Test of memory malingering (TOMM) scores are not affected by chronic pain or depression in patients with fibromyalgia. *The Clinical Neuropsychologist*, *21*, 532-546. doi:10.1080/13854040600611392

Lange, R. T., Pancholi, S., Bhagwat, A., Anderson-Barnes, V., & French, L. M. (2012). Influence of poor effort on neuropsychological test performance in U.S. military personnel following mild traumatic brain injury. *Journal of Clinical and Experimental Neuropsychology*, *34*, 453–466. doi:10.1080/13803395.2011.648175

Larrabee, G. J. (2012). Performance Validity and Symptom Validity in Neuropsychological Assessment. *Journal of the International Neuropsychological Society*, *18*, 1-7. https://doi.org/10.1017/S1355617712000240

Leppma, M., Long, D., Smith, M., & Lassiter, C. (2017). Students seeking ADHD treatment: Performance validity assessment using the NV-MSVT and IVA-Plus. *Applied Neuropsychology: Adult*. http://dx.doi.org/10.1080/23279095.2016.1277723

Lippa, S. M., Lange, R. T., Bhagwat, A., & French, L. M. (2017). Clinical utility of embedded performance validity tests on the repeatable battery for the assessment of neuropsychological status (RBANS) following mild traumatic brain injury. *Applied Neuropsychology: Adult, 24*, 73-80. http://dx.doi.org/10.1080/23279095.2015.1100617

MacAllister, W. S., Nakhutina, L., Bender, H. A., Karantzoulis, S., & Carlson, C. (2009).

Assessing effort during neuropsychological evaluation with the TOMM in children and adolescents with epilepsy. *Child Neuropsychology: A Journal on Normal and Abnormal Development in Childhood and Adolescence*, *15,* 521-531. doi:10.1080/09297040902748226

Nishimoto, T., Miyawaki, K., Ueda, T., Une, Y., & Takahashi, M. (2005). Japanese normative set of 359 pictures. *Behavior Research Methods, 37,* 398-416. doi: 10.3758/BF03192709

O'Bryant, S. E., Engel, L. R., Kleiner, J. S., Vasterling, J. J., & Black, F. W. (2007). Test of memory malingering (TOMM) trial 1 as a screening measure for insufficient effort. *The Clinical Neuropsychologist, 21*, 511-521. http://dx.doi.org/10.1080/13854040600611368

O'Bryant, S. E., Finlay, C. G., & O'Jile, J. R. (2007). TOMM performances and self-reported symptoms of depression and anxiety. *Journal of Psychopathology and Behavioral Assessment, 29*, 111-114. doi:10.1007/s10862-006-9034-9

O'Bryant, S. E., & Lucas, J. A. (2006). Estimating the predictive value of the test of memory malingering: An illustrative example for clinicians. *The Clinical Neuropsychologist*, *20,* 533-540. doi:10.1080/13854040590967568

Oorsouw, K., & Merckelbach, H. (2010). Detecting malingered memory problems in the civil and criminal arena. *Legal and Criminological Psychology*, *15,* 97-114. doi: 10.1348/135532509X451304

Porter, S., & Woodworth, M. (2006). «I'm sorry I did it … but he started it»: A comparison of the official and self-reported homicide descriptions of psychopaths and non-psychopaths. *Law and Human Behavior*, *31,* 91-107. doi: 10.1007/s10979-006-9033-0

Powell, M. R., Gfeller, J. D., Hendricks, B. L., & Sharland, M. (2004). Detecting symptom- and test-coached simulators with the test of memory malingering. *Archives of Clinical*

Neuropsychology, *19*, 693-702.m http://doi.org/10.1016/j.acn.2004.04.001

Reslan, S., & Axelrod, B. N. (2017). Evaluating the medical symptom validity test (MSVT) in a sample of veterans between the ages of 18 to 64. *Applied Neuropsychology: Adult, 24*, 132-139. http://dx.doi.org/10.1080/23279095.2015.1107565

Rickards, T. A., Cranston, C. C., Touradji, P., & Bechtold, K. T. (2017). Embedded performance validity testing in neuropsychological assessment: Potential clinical tools. *Applied Neuropsychology: Adult*. http://dx.doi.org/10.1080/23279095.2017.1278602

Ryan, J. J., Glass, L. A., Hinds, R. M., & Brown, C. N. (2010). Administration order effects on the test of memory malingering. *Applied Neuropsychology*, *17*, 246-250. doi:10.1080/09084282.2010.499802

Simões, M. R. (2006). Testes de validade de sintomasnaavaliação de comportamentos de simulação. In A. C. Fonseca, M. R. Simões, M. C. T. Simões & M. S. Pinho, (Eds.), *Psicologia Forense* (pp. 280-309). Coimbra: Almedina.

Simon, M. J. (2007). Performance of mentally retarded forensic patients on the test of memory malingering. *Journal of Clinical Psychology*, *63*, 339-344. doi: 10.1002/jclp.20351

Slick, D. J., Tan, J. E., Strauss, E., & Hultsch, D. F. (2004). Detecting malingering: A survey of experts' practices. *Archives of Clinical Neuropsychology*, *19*, 465-473. doi:10.1016/j.acn.2003.04.001

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning & Memory, 6*, 174-215.

Superlab (4.5) [Computer Software]. San Pedro, Cedrus Corporation.

Tan, J. E., Slick, D. J., Strauss, E., & Hultsch, D. F. (2002). How'd they do it? Malingering

strategies on symptom validity tests. *The Clinical Neuropsychologist*, *16*, 495-505. doi:10.1076/clin.16.4.495.13909

Tombaugh, T. M. (1996). *Test of Memory Malingering (TOMM)*. Canada: Multi-Health Systems.

Vagnini, V. L., Berry, D., Clark, J. A., & Jiang, Y. (2008). New measures to detect malingered neurocognitive deficit: Applying reaction time and event-related potentials. *Journal of Clinical and Experimental Neuropsychology, 30,* 766-776. doi:10.1080/13803390701754746

Vanderslice-Barr, J. L., Miele, A. S., Jardin, B., & McCaffrey, R. J. (2011). Comparison of computerized versus booklet versions of the TOMM™. *Applied Neuropsychology*, *18*, 34-36. doi:10.1080/09084282.2010.523377

Van Dyke, S. A., Millis, S. R., Axelrod, B. N., Hanks, R. A. (2013). Assessing effort: Differentiating performance and symptom validity. *The Clinical Neuropsychologist, 27,* 1234-1246. http://dx.doi.org/10.1080/13854046.2013.835447

Wang, D. D., Proctor, R. W., Pick, D. F. (2009). Allocation of effort as a function of payoffs for individual tasks in a multitasking environment. *Behavior Research Methods, 41,* 705-716. doi:10.3758/BRM.41.3.705

Wechsler, D. (1997). *WMS-III: Wechsler memory scale administration and scoring manual* (3rd. ed.). Psychological Corp.

Weinborn, M., Woods, S. P., Nulsen, C., & Leighton, A. (2012). The effects of coaching on the verbal and nonverbal medical symptom validity tests. *The Clinical Neuropsychologist*, *26*, 832–849. doi:10.1080/13854046.2012.686630

Whiteside, D. M., Dunbar-Mayer, P., & Waters, D. P. (2009). Relationship between TOMM performance and PAI validity scales in a mixed clinical sample. *The Clinical*

Neuropsychologist, *23*, 523-533. doi:10.1080/13854040802389169

Willison, J., & Tombaugh, T. N. (2006). Detecting simulation of attention deficits using reaction time tests. *Archives of Clinical Neuropsychology*, *21*, 41-52. http://doi.org/10.1016/j.acn.2005.07.005

Table 1

Mean and standard deviation for hits in DETECTS according to phase and group of participants.

Group	Phase 1	Phase 2
Normative	48.78 (1.46)	49.68 (.57)
Simulator	30.63 (9.71)	30.33 (11.23)

Table 2

Mean and standard deviation for participants' response times (ms) according to DETECTS phase, group of participants, and type of response (hit or false alarm).

Group	Phase 1 Hit	Phase 2 Hit	Phase 1 False Alarm	Phase 2 False Alarm
Normative	*1374* (326)	*1107* (227)	*2031* (919)	*1376* (658)
Simulator	*2250* (769)	*1889* (748)	*2242* (609)	*2095* (950)

Table 3

Sensitivity, Specificity, Positive Predictive Power (PP+), and Negative Predictive Power (PP-) for possible cut-off scores in DETECTS phase 1

Classification	Hits	Sensitivity	Specificity	PP+	PP-
Non-Credible Performance	42	96%	100%	100%	98%
Non-Credible Performance	43	96%	100%	100%	98%
Non-Credible Performance	44	96%	100%	100%	98%
Credible Performance	45	96%	98%	96%	98%
Credible Performance	46	96%	90%	85%	97%

Table 4

Sensitivity, Specificity, Positive Predictive Power (PP+), and Negative Predictive Power (PP-) for possible cut-off scores in DETECTS phase 2

Classification	***Hits***	***Sensitivity***	***Specificity***	***PP+***	***PP-***
Non-Credible Performance	45	92%	100%	100%	95%
Non-Credible Performance	46	92%	100%	100%	95%
Non-Credible Performance	47	92%	100%	100%	98%
Credible Performance	48	92%	95%	92%	98%
Credible Performance	40	83%	100%	100%	91%