Performance Invalidity Base Rates Among Healthy Undergraduate Research Participants

Thomas P. Ross*, Ashley M. Poston, Patricia A. Rein, Andrew N. Salvatore, Nathan L. Wills, Taylor M. York

College of Charleston, Charleston, SC, USA

*Corresponding author at: Department of Psychology, College of Charleston, Charleston, SC 29424, USA. Tel.: +1-843-953-3339; fax: +1-843-953-7151.
E-mail address: rosttp@cofc.edu (T.P. Ross).

Accepted 1 October 2015

Abstract

Few studies have examined base rates of suboptimal effort among healthy, undergraduate students recruited for neuropsychological research. An and colleagues (2012, Conducting research with non-clinical healthy undergraduates: Does effort play a role in neuropsychological test performance? Archives of Clinical Neuropsychology, 27, 849–857) reported high rates of performance invalidity (30.8%–55.6%), calling into question the validity of findings generated from samples of college students. In contrast, subsequent studies have reported much lower base rates ranging from 2.6% to 12%. The present study replicated and extended previous work by examining the performance of 108 healthy undergraduates on the Dot Counting Test, Victoria Symptom Validity Test, Word Memory Test, and a brief battery of neuropsychological measures. During initial testing, 8.3% of the sample scored below cutoffs on at least one Performance Validity Test, while 3.7% were classified as invalid at Time 2 (M interval = 34.4 days). The present findings add to a growing number of studies that suggest performance invalidity base rates in samples of non-clinical, healthy college students are much lower than An and colleagues initial findings. Although suboptimal effort is much less problematic than suggested by An and colleagues, recent reports as high as 12% indicate including measures of effort may be of value when using college students as participants. Methodological issues and recommendations for future research are presented.

Keywords: Performance Validity Tests; Effort; Base rates; Undergraduate students

Introduction

Recruiting participants from introductory psychology courses is a common mechanism for enabling psychological research at universities and colleges in the United States and Canada (Wintre, North, & Sugar, 2001). This practice is also common to neuropsychological research as college samples have been employed in a variety of investigations such as analogue studies of malingering (e.g., Demakis, 1999; Suhr & Gunstad, 2000; Youngjohn, Lees-Hayley, & Binder, 1999), test development and standardization (e.g., Green, 2003; Tombaugh, 1996; Wechsler, 1997a), normative studies (see Mitrushina, Boone, Razani, & D’Elia, 2005; Strauss, Sherman, & Spreen, 2006), and numerous psychometric investigations (e.g., Abwender, Swan, Bowerman, & Connolly, 2001; Humes, Welsh, Retzlaff, & Cookson, 1997; Ross, 2014; Ross et al., 2007; Schatz & Ferris, 2013; Thiruselvam, Vogt, & Hoelzle, 2015; Troyer, Moscovitch, & Winocur, 1997).

An, Zakzanis, and Joordens (2012) cautioned against recruiting undergraduate participants for neuropsychological research. They noted participant pool mechanisms typically incentivize participation by awarding course credit or extra-credit; however, receiving credit is not contingent upon effort put forth during a study. An and colleagues hypothesized college students are not motivated to perform their best on testing relative to patients who are invested in neuropsychological evaluations for health reasons. An and colleagues (2012) examined the neuropsychological test performance of 36 healthy undergraduates who were recruited from an Introductory Psychology class; each student received course credit for participation. Participants were administered a small number of neuropsychological measures that included three stand-alone Performance Validity Tests (PVTs)
designed specifically to detect suboptimal effort. The PVTs included the Dot Counting Test (DCT; Boone, Lu, & Herzberg, 2002), the Test of Memory Malingering (TOMM; Tombaugh, 1996), and the Victoria Symptom Validity Test (VSVT; Slick, Hopp, Strauss, & Thompson, 1997). Participants were classified as demonstrating suboptimal effort (i.e., producing invalid test scores) if persons performed below cutoffs on any one of the PVT indices (see Table 1). The participants were examined serially using the same measures across two testing sessions (test–retest interval \( M = 79.7 \) days). Of 36 participants who completed the initial testing, only 13 of these individuals participated in Session 2. The authors hypothesized those who exhibit poor effort upon initial assessment would do so again when retested (suggesting suboptimal effort was due to stable, dispositional factors).

An and colleagues (2012) reported that 55.6% of participants (\( n = 20 \)) performed below cutoffs on at least one PVT at initial testing. The majority of persons exhibiting poor effort at Time 1 performed below cutoffs on the VSVT (detecting 65% of the invalid cases), followed by the DCT (detecting 45% of the invalid cases). Among those who participated in repeat testing, 30.8% (\( n = 4 \)) performed below cutoffs on at least one PVT at Time 2 (three cases detected by the DCT; one detected by the VSVT). No participant performed below cutoffs on the TOMM at Time 1 or Time 2. The authors inspected each case of performance invalidity to determine if those classified as invalid at Time 2 were also classified as invalid at Time 1. Of the four participants’ who failed at least one PVT indicator at Time 2, three of these individuals also failed at least one PVT at Time 1. Spearman rank-order correlations between PVT scores at Time 1 and Time 2 ranged from \( r = .61 \) to .77 (\( n = 13 \); \( ps < .02 \)). The authors interpreted these findings as support for the position that base rates of performance invalidity are high in undergraduate samples and suboptimal effort among participants is relatively stable and likely attributable to dispositional factors. Moreover, they raised the possibility that studies of college undergraduates may suffer from threats to validity if measures of effort are not utilized.

An and colleagues (2012) study has since been criticized for a number of methodological limitations (see DeRight & Jorgensen, 2015; Santos, Kazakov, Reamer, Park, & Osmon, 2014; Silk-Eglit et al., 2014). These limitations include a small sample size upon initial testing (\( n = 36 \)) and the use of a restricted sample that consisted of mostly Asian (77%) females (86%). Moreover, only half of the sample reported that English was their primary language (and all tests were administered in English). The number of participants at retesting was also very small (\( n = 13 \)) and the span of the test–retest interval varied widely, ranging from 29 to 115 days. In addition to sampling limitations, Silk-Eglit and colleagues (2014) criticized An and colleagues’ study for employing cutoffs that contradict PVT manual recommendations. For example, An and colleagues used a cutoff score of \( \leq 50 \) for the TOMM, which conflicts with the recommendation that \( \leq 17 \) should be used for this index when classifying suboptimal performance (see Slick et al., 1997). Moreover, An and colleagues used response latency scores for difficult items of the VSVT as a single indicator of suboptimal effort when the manual cautions against using these scores in isolation when making such decisions (see Silk-Eglit et al., 2014; Slick et al., 1997).

Although the use of college students as participants is not uncommon in neuropsychological research, there is a dearth of studies that directly examine the base rates of suboptimal test-taking effort in such samples. Following An and colleagues report, three studies have re-examined this issue and failed to replicate such high rates of suboptimal effort (i.e., \( 30\%–55\% \)). Silk-Eglit and colleagues (2014) examined the performance of 133 students who were administered PVTs included among a battery of common neuropsychological measures (e.g., Trial Making Test). Using archival data, participants were obtained from two independent samples. Participants in Sample 1 (\( n = 83 \)) were administered two cognitive tasks and three PVTs designed to be stand-alone measures of effort (e.g., Word Memory Test, WMT) using cutoffs outlined in Table 1. Participants in Sample 2 (\( n = 50 \)) were administered four neuropsychological tests and three measures of effort: first, \( \leq 85\% \) correct on Immediate

<table>
<thead>
<tr>
<th>Table 1. Stand-alone Performance Validity Test (PVT) cutoffs employed in previous studies of suboptimal effort among college undergraduates</th>
</tr>
</thead>
<tbody>
<tr>
<td>PVT</td>
</tr>
<tr>
<td>-----</td>
</tr>
<tr>
<td><strong>An and colleagues (2012)</strong></td>
</tr>
<tr>
<td>DCT</td>
</tr>
<tr>
<td>TOMM</td>
</tr>
<tr>
<td>VSVT</td>
</tr>
<tr>
<td><strong>Silk-Eglit and colleagues (2014)</strong>*</td>
</tr>
<tr>
<td>TOMM</td>
</tr>
<tr>
<td>VSVT</td>
</tr>
<tr>
<td>WMT</td>
</tr>
<tr>
<td><strong>Santos and colleagues (2014)</strong></td>
</tr>
<tr>
<td>WMT</td>
</tr>
</tbody>
</table>

Note: “Sample 1 from a multi-sample investigation.”
Recognition, Delayed Recognition or Consistency Index of the Medical Symptom Validity Test (Green, 2004); second, <7 on Reliable Digit Span (RDS; Greiffenstein, Baker, & Gola, 1994); and third, <15 correct on the Forced Choice Recognition trial of the of the California Verbal Learning Test-II (CVLT-II; Delis, Kaplan, Kramer, & Ober, 2000). Silk-Egilit and colleagues (2014) reported only 2.26% of the 133 undergraduates included across both samples exhibited performance invalidity. More specifically, only three persons in Sample 1 performed below cutoffs on any one PVT, while no persons were classified as exhibiting performance invalidity in Sample 2. Silk-Egilit and colleagues therefore concluded base rates of suboptimal effort among samples healthy college students is in all likelihood low.

Santos and colleagues (2014) examined 110 undergraduate research participants recruited from psychology courses using a small battery of neuropsychological measures that included one PVT—the WMT (see Table 1). Using criteria outlined in the WMT test manual (Green, 2003), seven individuals (or 6.4%) of the sample performed below cutoffs for profile invalidity. Santos and colleagues (2014) also examined whether positive, neutral, or negative demand characteristics (e.g., examiner encouragement and test administration order in terms of difficulty) related to obtained rates of performance invalidity, but they found no differences between conditions. The authors concluded the WMT was robust to such demand characteristic and were in agreement with Silk-Egilit and colleagues’ (2014) position that base rate of suboptimal effort in college samples are relatively low.

DeRight and Jorgensen (2015) examined the frequency of suboptimal effort in a sample of 77 non-clinical healthy college students using CNS Vital Signs (CNS-VS; Gualtieri & Johnson, 2006), a computerized test battery which includes embedded measures of effort while assessing several neurocognitive domains (e.g., verbal and visual memory, finger tapping, symbol-digit coding and attention). During a single session that lasted ~90 min, all participants first completed the CNS-VS battery then rated their own effort on these tasks using a Likert-type scale. Immediately afterwards, all participants were re-administered the CNS-VS, which included the same measures but were presented in a randomized order. DeRight and Jorgensen (2015) reported 12% of the 77 participants failed at least one CNS-VS embedded indicator upon first administration. Of the seven possible validity indicators used, failure rates varied widely from 0% (for verbal memory) to 7% (for the continuous performance task). On the repeat administration of the CNS-VS, 11% of the participants failed at least one validity indicator. Upon inspection of these cases, six individuals (8%) failed at least one indicator across both administrations of the CNS-VS battery. Two cases were classified as invalid on the repeat administration of the CNS-VS but not the baseline, while three persons failed validity indicators on the baseline but not the repeat administration. DeRight and Jorgensen also reported some of the simple (i.e., less effortful) tasks (e.g., Finger Tapping) were correlated with failure on validity indicators, while complex tasks (e.g., Stroop interference) were not. Failed validity indicators were correlated with participants’ self-ratings of test-taking effort at baseline ($r = -.32; p < .001$) and repeat administration ($r = -.39; p < .001$). The authors concluded base rates of performance invalidity are significantly high, especially so for samples of young adults without a history of neurological disease. Moreover, they asserted their findings were more consistent with An and colleagues (2012) and contradict reports of low base rates (i.e., Santos et al., 2014; Silk-Egilit et al., 2014).

Although few published studies on this topic exist, the estimated base rates of suboptimal effort among healthy college samples vary widely thus far (i.e., from 2.2% to 55.6%). The disparate findings are likely attributable to significant methodological differences across studies. The highest frequencies of performance invalidity among healthy undergraduates were found by An and colleagues (2014) who employed a sample that differed substantially in size and demographic composition relative to other studies reviewed. Additionally, the PVTs used to classify suboptimal effort varied in type and number across studies. For example, Santos and colleagues (2014) investigation employed only one PVT (i.e., the WMT), which was not included in the An and colleagues study. Similarly, the An and colleagues study employed the DCT, which was not included in the investigation by Silk-Egilit and colleagues (2014). Moreover, studies have employed different cutoffs in instances where the same PVTs (e.g., VSVT) were used (see Table 1) and the investigation by DeRight and Jorgensen (2015) used embedded indicators in isolation, which can result in lower sensitivity and specificity (Armistead-Jehle & Hansen, 2011; Miele, Gunner, Lynch, & McCaffrey, 2012). Finally, studies have differed in scope and design, especially with regard to examining the stability of suboptimal effort among participants. An and colleagues (2012) was the only study among those reviewed to address the temporal stability of invalidity classification across separate examination dates, so this element of their study has not been replicated. Unfortunately, these methodological differences limit the degree findings can be generalized across studies and therefore prevent general conclusions about the validity of using non-clinical samples of healthy college students in neuropsychological research.

To address the aforementioned limitations, the present study investigated performance invalidity base rates among healthy, college undergraduates using three stand-alone PVTs, and other neuropsychological measures common to previous work. To better replicate and extend the findings of previous studies, participants were examined serially to explore the temporal stability of any profile invalidity detected. In keeping with stated recommendations for future studies (e.g., An et al., 2012; Santos et al., 2014), the WMT was included along with other PVTs. Finally, the present study examined the base rates of impaired range test scores on a small battery of common neuropsychological measures.

$$r \approx p > .001$$
Method

Participants

Participants included an initial sample of 117 undergraduates who were enrolled in Introductory Psychology courses at a mid-sized, liberal arts, and sciences university in the southeastern, United States. Using a screening procedure for exclusionary criteria (described below), 108 healthy individuals ($M$ Age = 19.83, $SD$ = 2.01) were retained for data analysis. Most participants were female (76%) and right handed (84%) with a mean education level of 13.69 years ($SD$ = 0.91). The majority of persons self-identified as Caucasian (72.2%), followed by African American (15.7%), Asian (7.4%), Hispanic (2.8%), and 1.9% reported “other” ethnic identities. Participants’ estimated mean Full Scale IQ using the 61-item, North American Adult Reading Test (NAART; Blair & Spreen, 1989) was 104.50 ($SD$ = 8.1).

Materials and Procedure

After obtaining their informed consent, participants were first administered a brief, self-report questionnaire to obtain information about their demographic background and health history. Persons with a history of neurologic disease or trauma, attention deficit-hyperactivity disorder or learning disability, or other psychiatric conditions were excluded from analyses. Using these criteria, nine individuals were excluded (1 due to epilepsy, 1 due to head trauma, 3 due to learning disability and/or ADHD, and 4 due to psychiatric conditions that included unipolar, bipolar, or anxiety disorders). Participants were recruited and self-registered for this study using the online experimental management system (SONA Systems, Ltd, Version 2.72; Tallinn, Estonia). All persons received required “research credits” toward their psychology course in exchange for their participation. As customary at this institution (and others), students were awarded credit for any study for which they signed-up and participated (regardless of their actual performance level).

Participants completed three stand-alone PVTs that included the VSVT (Slick et al., 1997), the DCT (Boone et al., 2002), and the WMT (Green, 2003). The VSVT and DCT were chosen because these specific procedures were employed in the An and colleagues (2012) study. The WMT was chosen because of its known sensitivity to suboptimal effort (see Green, 2003; Larrabee, 2012) and for consistency, as this measure was employed in the two studies which have failed to replicate the findings of An and colleagues (2012).

Participants also completed the NAART, Letter-Number Sequencing (LNS) subtest of the WMS-III (Wechsler, 1997b), Color-Word Interference Test of the Delis–Kaplan Executive Function System (D-KEFS-CW; Delis, Kaplan, & Kramer, 2001), the Trail Making Test Parts A and B (TMT; Reitan, 1986), the Finger Tapping Test (FTT; Reitan & Wolfson, 1993), and the Grooved Pegboard Test (GPT; Matthews & Klove, 1964). The NAART was used to estimate the mean intellectual level of the sample in a time-efficient manner. The additional measures used (e.g., TMT and FTT) are among those frequently included in neuropsychological examinations by practitioners (see Butler, Retzlaff, & Vanderploeg, 1991; Rabin, Barr, & Burton, 2005) and represent a mix of relatively simple (e.g., FTT) and effortful (e.g., Stroop Interference) tasks. Additionally, these measures were included in one or more of the previous studies reviewed and therefore would better enable comparisons across investigations.

After obtaining informed consent, participants were first administered a brief questionnaire about their demographic background and health history. Next, all participants completed the WMT to ensure the 30-min period between the immediate and delayed recall subtests. Finally, the DCT, VSVT, and the aforementioned neuropsychological tests were administered to all participants. All three of the PVTs were administered at Time 1 and repeated at Time 2; however, the neuropsychological tests were not repeated in order to increase the breadth of measures included in the study while staying within time constraints. More specifically, the NAART, LNS, and GPT were administered as part of the Time 1 protocol, while the TMT, FTT, and D-KEFS Stroop subtests were administered at Time 2. For each session, the administration order of the DCT, VSVT, and the three neuropsychological tests included was determined randomly. Each testing appointment lasted for ~60 min and the mean number of days between each session was 34.4 (ranging from 28 to 56 days).

All tests were administered and scored in accordance with the aforementioned manuals by well-trained, advanced undergraduate students under the supervision of a PhD level, clinical psychologist. Examiners were blind to the exact nature of the study and had no prior knowledge of the An and colleagues (2012) report in particular. Prior to the stating specific instructions for any one procedure, all participants were told the following by examiners: “I’ll be asking you to complete several different tasks. During some of these tasks, I may ask you to work quickly while I time you using a stop watch. You may find that some of these tasks are relatively easy, while others may seem more difficult or challenging. All I ask is that you try your very best at all times.” Adhering to published manuals, the following cutoffs were used to classify performance invalidity: ≤ 17 correct on VSVT difficult items, or a
DCT E-score ≤ 14, or ≤ 82.5% correct on the immediate recognition (IR) or delayed recognition (DR) or consistency index (CNS) or ≤ 70% correct on the multiple choice (MC) or ≤ 60% correct on the paired associates (PA) indices of the WMT. Consistent with the conservative approach used in prior research in this area, performance invalidity was classified when persons performed below recommended cutoffs on any one PVT index.

Results

At initial testing, nine individuals (or 8.33% of the sample) scored below cutoffs on at least one PVT and lower percentages of suboptimal performance were observed across most validity indices (see Table 2). Upon inspection of cases performing below cutoffs, five of the nine instances of suboptimal effort were observed on only one PVT index. The mean number of failed validity indicators was 1.67 (primarily due to four cases failing more than one of the WMT indicators). Only two persons failed a validity indicator derived from more than one PVT procedure (e.g., the DCT and the WMT). The percentages of invalid cases identified by each PVT index are shown in Table 2. The highest number of cases classified as invalid at Time 1 were observed for the DCT E-score (n = 5), followed by the WMT indices which in combination detected the remaining cases. Interestingly, two individuals performed below cutoffs on the WMT Multiple Choice index despite having acceptable Immediate Recall, Delayed Recall, and Consistency scores. In contrast, no one in the present sample performed below a cutoff of ≤ 17 correct on VSVT difficult items at Time 1.

Four participants (3.7% of the sample) scored below cutoffs on at least one PVT during repeat testing, and two of these cases were also classified as invalid at Time 1. Similar to findings observed for initial testing, cases classified as invalid at Time 2 were most frequently identified by poor performance on the DCT E-score (see Table 2). The mean number of PVT indicators failed by those classified as invalid at Time 2 was 1.5. Only one of these cases of invalidity failed an indicator from more than one test procedure. The single case classified as invalid based on WMT scores at Time 2 was the result of poor performance on WMT-MC and PA scores (despite normal performance on WMT-IR, DR, and CNS scores).

Most participants performed within the normal ranges on neuropsychological measures administered and the mean and standard deviations for these measures were commensurate with published normative data for individuals of similar age and education level. The percentage of valid cases (n = 99) scoring < 2 SDs from the mean on each neuropsychological measure ranged from 2% (for LNS scores) to 6% (for K-DEF5-IS scores). Of the cases classified as invalid at Time 1 or Time 2, only three of these persons performed 2 SDs below the mean on at least one neuropsychological measure. The maximum number of impaired range scores on a neuropsychological measure was 2 (observed for one invalid case only). In contrast, 18% of those persons who passed all PVT indicators performed within an impaired range on at least one neuropsychological measure, 10% scored within impaired ranges on 2 measures, and 4% scored within impaired ranges on 3 measures. For the valid cases (n = 99), the mean number of neuropsychological test scores occurring < 2 SDs was 1.64; the mean number observed for the invalid cases was 1.67. In keeping with previous research (e.g., Silk-Eglit et al., 2014), the association between profile validity classification and neuropsychological test scores was not examined because of very low bases rates and the related statistical problems (e.g., small n, low power, and restriction of range).

Table 2. Means, SDs, and percent of sample below cutoffs on PVTs at Time 1 and Time 2 (n = 108)

<table>
<thead>
<tr>
<th>Test</th>
<th>Time 1</th>
<th>Time 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean (SD) N below</td>
<td>Mean (SD) N below</td>
</tr>
<tr>
<td>DCT-E</td>
<td>9.61 (2.20) 5 4.6</td>
<td>8.54 (2.15) 3 2.7</td>
</tr>
<tr>
<td>VSVT-D</td>
<td>23.56 (0.89) 0 0</td>
<td>23.28 (1.10) 1 0.9</td>
</tr>
<tr>
<td>WMT-IR</td>
<td>98.24 (4.07) 2 1.8</td>
<td>99.19 (1.95) 0 0</td>
</tr>
<tr>
<td>WMT-DR</td>
<td>98.65 (2.59) 1 0.9</td>
<td>98.37 (1.53) 0 0</td>
</tr>
<tr>
<td>WMT-CNS</td>
<td>97.47 (4.24) 2 1.8</td>
<td>98.70 (2.78) 0 0</td>
</tr>
<tr>
<td>WMT-MC</td>
<td>93.52 (10.28) 3 2.7</td>
<td>96.25 (5.89) 1 0.9</td>
</tr>
<tr>
<td>WMT-PA</td>
<td>92.73 (13.76) 2 1.8</td>
<td>97.60 (5.90) 1 0.9</td>
</tr>
</tbody>
</table>

Notes: DCT-E = Dot Counting Test E-Score; VSVT-D = Victoria Symptom Validity Test Difficult Items Correct; WMT-IR = Word Memory Test Immediate Recognition; WMT-DR = Word Memory Test Delayed Recognition; WMT-CNS = Word Memory Test Consistency Index; WMT-MC = Word Memory Test Multiple Choice Index; and WMT-PA = Word Memory Test Paired Associates Index.

*The following cutoffs were used to classify performance invalidity: ≤ 17 correct on VSVT difficult items, or a DCT E-score ≥ 14, or ≤ 82.5% correct on WMT-IR or DR or CNS or ≤ 70% correct on WMT-MC or ≤ 60% correct on WMT-PA.

*Mean testing interval was 34.4 days.
Discussion

This study sought to explore the base rates of suboptimal effort using a large sample of non-clinical, healthy college students. Additionally, the stability of PVT profiles was examined along with base rates of impaired range scores on common neuropsychological tests. The present study found considerably lower base rates for performance invalidity when compared with the high percentages reported by An and colleagues (2012). These dissimilar findings were observed upon initial and repeat testing sessions using three stand-alone PVTs sensitive to suboptimal effort. Although the low base rates precluded significance testing, careful inspection of each case did not support the temporal stability of poor effort as suggested by two previous reports (An et al., 2012; DeRight & Jorgensen, 2015). That is, those individuals who performed sub-optimally at Time 1 did not necessarily do so again at Time 2 and vice versa. The lack of observed stability among these cases across such a short time period (M = 34.4 days) is more consistent with the interpretation that situational (rather than dispositional) factors are influencing participants’ motivation to perform on testing. Although results addressing the stability (i.e., consistency) of suboptimal effort varied across studies, we believe the present study provided the most stringent examination to date. That is, the present study’s design utilized a larger sample and more uniform test—retest interval than used by An and colleagues (2012) and the repeated test administration in the DeRight and Jorgensen (2015) study did not involve testing sessions on separate occasions.

Although base rate of invalidity was low and did not allow for inferential statistics, the present study did not yield any indication that those participants classified as invalid performed more poorly on the neuropsychological measures relative to those participants with valid profiles. For example, the mean number of neuropsychological test scores occurring < 2 SDs for those classified as invalid was nearly identical to the mean for those who passed every validity indicator. The percentages of impaired range scores observed in the present study are commensurate with base rates of impaired scores (% ≤ 1 SD, % ≤ 2 SD, etc.) documented in “normal” samples when batteries of neuropsychological tests are administered (see Binder, Iverson, & Brooks, 2009). The present findings are consistent with recent investigations (e.g., Santos et al., 2014) that reported base rates within the 2% – 12% range. Taken together, the present study contributes to a small but growing body of evidence suggesting that the high percentages reported by An and colleagues (2012) are aberrant. More recent studies show a convergence of findings obtained using several different performance validity indicators that include stand-alone and embedded measures.

We agree with previous interpretations (e.g., Santos et al., 2014; Silk-Eglit et al., 2014) that attribute the high base rates reported by An and colleagues (2014) to methodological factors unique to their study. Understanding these methodological differences is important for elucidating the circumstances under which base rates of performance validity may increase or decrease among a population frequently included in research. As noted previously, An and colleagues used criteria that conflicted with the published recommended cutoffs. If criteria used by An and colleagues (e.g., ≤20 correct on VSVT difficult items or ≤23 correct on easy items) were applied to the present sample, only three additional cases (totaling 11%) would be classified as invalid at Time 1, while two additional cases (totaling 5.5%) would be classified as invalid at Time 2. Thus, totals derived from using more liberal criteria and additional VSVT indicators are still well below the 30.8%–55.6% reported by An and colleagues, providing further evidence their findings are anomalous and cannot be fully explained by the use of different cutoff scores to classify performance invalidity. The striking disparity between the results of An and colleagues (2012) and the present study are likely attributable to An and colleagues use of a sample severely restricted in size and demographic composition (e.g., 77.2% Asian). Moreover, only half of the participants in the An and colleagues study reported that English was their primary language. Accordingly, the characteristics unique to An and colleagues sample may have resulted in less accurate estimates of base rates (i.e., more false positives) relative to the use of larger and more diverse samples.

Although recent estimates of base rates are much lower than initially reported by An and colleagues, the values do vary significantly and estimates as high as 12% by DeRight and Jorgensen (2015) would suggest some degree of caution is warranted. The range in base rates observed across recent studies is likely due to the different PVTs used. For example, the present study included the MC and PA scores for the WMT and the DCT E-score which were not employed by Silk-Eglit and colleagues (2014). Similarly, Santos and colleagues (2014) included only one PVT in their study. Including these additional PVT scores may account for the slightly higher rates of performance invalidity observed in the present study. DeRight and Jorgensen (2015) used embedded measures of effort in isolation taken from a lesser known battery which may explain the higher base rates obtained when compared with the present study. We agree with Silk-Eglit and colleagues (2014) who argued invalidity base rates are relatively low among undergraduate samples and previous neuropsychological research using student participants should not be discounted. However, evidence from the present study and DeRight and Jorgensen (2015) suggest that suboptimal effort can be present in a significant amount (i.e., 8% to 12%) of healthy undergraduate research participants. Therefore, we recommend that researchers using undergraduate participants take precautions and include PVTs or consider the interpretation of poor effort when very low scores are observed on neuropsychological tests. Given most persons classified as invalid in the present study failed only one PVT index, we recommend that future investigations of invalidity base rates include multiple PVTs. Other types of neuropsychological studies using undergraduate samples would likely benefit from including >1 PVT indicator. Additionally, studies using the
WMT in isolation would benefit from including the MC and PA scores as these indices detected cases of questionable validity not identified by IR, RD, and CNS scores. Fortunately, there are several PVTs available for use (see Larrabee, 2012; Leighton, Weinborn, & Mayberry, 2014). Some of these measures (e.g., the DCT) require very little time to administer and embedded measures (if typically included in the research protocol) do not increase test administration time.

The present study had a number of strengths that include the use of a prospective design, large sample size, and examiners who were blind to the study’s nature. In addition, the use of multiple stand-alone PVTs and a serial testing procedure allowed for more direct comparisons with An and colleagues (2012) study relative to previous replication studies (DeRight & Jorgensen, 2015; Santos et al., 2014; Silk-Eglit et al., 2014). However, the present investigation was not without limitations which include the use of a sample that was primarily Caucasian (72%) and female (76%) in demographic composition and using a relatively short interval between assessments. Also, the present study did not include all of the stand-alone PVTs (e.g., TOMM) used in prior research (e.g., Silk-Eglit et al.) or include available embedded validity procedures (e.g., RDS or CVLT-II Recognition Hits). The present study also used a relatively small test battery (requiring only 60 min of administration time). Although other investigations of base rates employed longer batteries than used in the present study (e.g., Silk-Eglit et al.), the length of protocols used in this area of research are generally much shorter than what is typical of clinical and/or forensic examinations (Lezak, Howieson, Bigler, & Tranel, 2012). Therefore, it is possible that higher rates of suboptimal performance could be observed in lengthier test batteries with additional PVTs.

Given the paucity of research in this area, we believe additional investigations are needed to identify more precisely the base rates of suboptimal performance among healthy college students. A better understanding of the circumstances or methodological variables that influence base rates would be of great value given the frequent use of such “convenience samples” in psychological research and other disciplines. Ideally, future studies in this area would include samples with ethnic groups currently underrepresented (e.g., African Americans and Hispanic/Latino) and employ a variety of stand-alone and embedded validity indicators using a test–retest design. Interestingly, DeRight and Jorgensen (2015) reported that participants who failed at least one validity indicator were more often tested in the early morning (i.e., prior to noon). Therefore, future studies should examine other methodological variables that might influence performance invalidity base rates such as lengthier testing protocols, longer intervals between repeated assessments, time of day and other demand characteristics.

Conflict of interest

None declared.

References
