Aggregating Validity Indicators Embedded in Conners’ CPT-II Outperforms Individual Cutoffs at Separating Valid from Invalid Performance in Adults with Traumatic Brain Injury

Laszlo A. Erdodi1,2,3, Robert M. Roth3, Ned L. Kirsch4, Renee Lajiness-O’neill5, Brent Medoff4

1Department of Psychology, University of Windsor, Windsor, ON, Canada
2Clinical Neuropsychology Service, University Hospital, London Health Sciences Centre, London, ON, Canada
3Neuropsychology Service, Department of Psychiatry, Geisel School of Medicine at Dartmouth, Lebanon, NH, USA
4Department of Physical Medicine and Rehabilitation, Adult Neurorehabilitation Day Treatment Program, University of Michigan, Ann Arbor, MI, USA
5Department of Psychology, Eastern Michigan University, Ypsilanti, MI, USA

*Corresponding author at: Neuropsychology Service, Department of Psychiatry, Geisel School of Medicine at Dartmouth, One Medical Center Drive, Lebanon, NH 03756-0001, USA.
E-mail address: lerdodi@gmail.com.

Accepted 26 May 2014

Abstract

Continuous performance tests (CPT) provide a useful paradigm to assess vigilance and sustained attention. However, few established methods exist to assess the validity of a given response set. The present study examined embedded validity indicators (EVIs) previously found effective at dissociating valid from invalid performance in relation to well-established performance validity tests in 104 adults with TBI referred for neuropsychological testing. Findings suggest that aggregating EVIs increases their signal detection performance. While individual EVIs performed well at their optimal cutoffs, two specific combinations of these five indicators generally produced the best classification accuracy. A CVI-5A ≥3 had a specificity of .92–.95 and a sensitivity of .45–.54. At ≥4 the CVI-5B had a specificity of .94–.97 and sensitivity of .40–.50. The CVI-5s provide a single numerical summary of the cumulative evidence of invalid performance within the CPT-II. Results support the use of a flexible, multivariate approach to performance validity assessment.

Keywords: ADHD; Forensic neuropsychology; Performance validity testing; Assessment

There is a growing consensus that establishing the validity of a given neuropsychological profile before interpreting it clinically is an essential part of the assessment process (Heilbronner, Sweet, Morgan, Larrabee, & Millis, 2009). While free-standing performance validity tests (PVTs) are still considered the gold standard measures of cognitive effort, their limitations (time-consuming, do not measure clinically relevant cognitive constructs, easier to identify and hence more susceptible to coaching) prompted a search for alternative solutions. As a result, a variety of embedded validity indicators (EVIs) were developed that unobtrusively measure test taking effort and provide valuable information about the credibility of a given response set.

Employing EVIs instead of relying on external PVTs represents a methodological advancement, as it reduces the level of inference involved in determining the validity of a given test score. Analyzing performance patterns within an instrument to establish the credibility of the response set seems to be a logical approach in performance validity assessment. A number of EVIs have been well established (Greiffenstein, Baker, & Gola, 1994) and research continues to produce novel and clinically useful measures within tests to assess the veracity of a given performance pattern (Johnson, Silverberg, Millis, & Hanks, 2012).

Continuous performance tests (CPTs) are designed to quantify various aspects of an individual’s performance during vigilance/sustained attention tasks and have been shown to be sensitive to ADHD (Balint, Czobor, Meszaros, Simon, & Bitter, 2008) as well as acquired attention deficit resulting from traumatic brain injury (TBI; Larrabee, 2012a). However, recent investigations suggest that CPTs are vulnerable to distortions caused by suboptimal effort. Leark, Dixon, Hoffman, and Huynh (2002) found that simulators produced excessive omission (OMI) and commission (COM) errors as well as having slower responding and higher response...
time variability than those who received standard instructions on the test of variables of attention (TOVA). Similarly, in a sample of 50 individuals assessed for personal injury and disability status, Henry (2005) found that the group that failed at least one PVT obtained significantly worse scores on all TOVA variables compared with those who produced a valid neuropsychological profile. Suhr, Hammers, Dobbins-Buckland, Zimak, and Hughes (2008) produced similar findings, with non-credible performance contributing to both self-report and objective data. They found that individuals who failed the Word Memory Test (WMT; Green, Allen, & Astner, 1996) were more impaired on the Conners’ CPT, second edition (CPT-II), highlighting the importance of independently measuring performance validity before clinically interpreting the results. Likewise, Marshall and colleagues (2010) reported that in a sample of 268 adults referred for assessment of ADHD, individual TOVA cutoffs of >25 OMI errors, >30 COM errors and an RT variability >180 were associated with a specificity >.90 and sensitivity ranging from 0.26 to 0.63 in reference to established PV Ts. On the CPT-II, OMI and COM T > 80 individually produced a specificity of 0.87 and 1.00, with a sensitivity of 0.57 and 0.04, respectively.

More recently, the use of EVIs within CPTs has been extended to the assessment of attention deficits secondary to TBI. Ord, Boetttcher, Greve, and Bianchini (2010) studied EVIs in the CPT-II using a sample of 82 patients with TBI. Invalid responding was defined as failing two or more of the following indicators: Test of Memory Malingering and WMT (using standard cutoffs), Portland Digit Recognition Test (East <23, Hard <18; Total <45), reliable digit span (RDS) <7 and MMPI-II (F > 80, FBS > 27, Meyers > 6). They identified OMI and hit reaction time standard error (HRT SE) as the most effective at separating invalid response sets in mild TBI (mTBI) with sensitivity of 0.30 and 0.41, respectively, at a specificity of 0.95. Combining the two indicators improved sensitivity to 0.44.

A similar study by Lange and colleagues (2013), involving 158 military service members undergoing neuropsychological assessment following TBI, found that OMI, COM and perseverations (PER) were the highest performing EVIs. Poor effort was defined as failing the WMT at the standard cutoffs. The single best cutoff was COM T > 68.7, with a sensitivity of 0.26 and specificity of 0.99 within the mTBI subsample. When these measures were combined, there was a small increase in sensitivity (0.29).

Together, these findings suggest that scores on CPTs are influenced by diagnostically non-credible response styles. The technical manual for the CPT-II states that three of the 12 subscales also function as EVIs: Response style (β), OMI and PER. As an index of the speed-accuracy trade-off, extreme values of β are conceptualized as a bias in response style. Conners (2004) suggested T > 60 as the cutoff for an overly cautious (OMI prone) response style and T < 40 as the cutoff for a hypervigilant (COM prone) response style. Likewise, scores of T > 100 on either the OMI or PER subscales are presented as possible signs of invalid response as they may reflect disengagement from the task or random responses, respectively. However, these cutoffs are presented as tentative EVIs as they may in fact reflect serious attention or neurological problems. Moreover, the manual does not offer any data on the signal detection property of these cutoffs in detecting invalid profiles.

Given the need for empirically based detection of exaggerated attention deficit and the lack of consensus on what the best markers of non-credible CPT-II profiles are, the present study was designed to further investigate the potential of EVIs within the CPT-II to detect invalid response sets in adults referred for neuropsychological testing following TBI. The project had two main goals: (i) to cross-validate previous cutoffs reported in the literature and (ii) to advance the methodology of aggregating EVIs to improve the compound signal detection performance pioneered by Ord and colleagues (2010) and Lange and colleagues (2013).

**Method**

**Participants**

Neuropsychological test data from 104 adults (55.8% males, 44.2% females, 90.4% right handed, 9.6% left handed) clinically referred for neuropsychological assessment following a TBI were collected at the outpatient neurorehabilitation service of a Midwestern academic medical center. Mean age of the sample was 38.8 years (SD = 16.7, range: 17–74). Mean level of education was 13.7 years (SD = 2.6, range: 7–20). Mean Full Scale IQ (FSIQ) was 92.6 (SD = 15.9, range: 60–130). Age, education, and FSIQ were normally distributed with skew within ±0.50 and kurtosis within ±1.0. The age distribution was slightly bimodal, with one peak ~20 and another ~50.

The majority of the sample (n = 78) was classified as mTBI based on available injury parameters [duration of loss of consciousness, self-reported peri-traumatic amnesia, and Glasgow Coma Scale (GCS) score]. The rest of the sample (n = 26) was classified as moderate to severe (mod/sev)TBI. The mTBI group had significantly higher GCS scores (M = 14.7, SD = 0.6) than the mod/sevTBI group (M = 7.0, SD = 3.3). Likewise, the proportion of the sample with positive neuroradiological findings was significantly lower [z(102) = 2.66, p < .01] in the mTBI group (60%) than the mod/sevTBI group (88%). The mTBI group scored slightly higher on the Beck Depression Inventory-II (M = 16.5, SD = 11.3) than the mod/sevTBI group (M = 11.8, SD = 11.5), but the contrast was not significant: t(96) = 1.77, p = .08.
Procedure

A fixed battery approach was employed using standardized measures of general intelligence, memory, attention, executive functions, and language and motor skills. Level of emotional distress was assessed with self-report measures. Cognitive effort was measured with a combination of stand-alone and embedded PVTs.

Materials

Each participant was administered the CPT-II. The main stand-alone PVT was the WMT. EVIs from a variety of standard neuropsychological instruments were also used as alternative measures of cognitive effort in an attempt to establish a more comprehensive evaluation of the credibility of each neurocognitive profile.

The Effort Index (EI-7)

Although the WMT has become a de facto gold standard for the detection of suboptimal cognitive effort (Bauer, O'Bryant, Lynch, McCaffrey, & Fisher, 2007; Gervais, Rohling, Green, & Ford, 2004; Hartman, 2002), it has several potential limitations. First, there are reports that it may be prone to false positives (Batt, Shores, & Chekaluk, 2008; Frederick, 2009; Greve, Ord, Curtis, Bianchini, & Brennan, 2008), although this criticism has since been challenged (Green, Montijo, & Brockhaus, 2011). Second, it generates a dichotomous classification system (Pass/Fail). While this is quite practical in settings where a categorical decision is expected, it lacks the detail required for a more nuanced assessment of performance validity. In effect, viewing effort as a dimensional construct could add important information to refine group-based classification models that may be missed using a pass/fail classification system (Feinstein, 1977). Third, there is a growing consensus about the need for a comprehensive assessment of effort in neuropsychological testing instead of relying on a single PVT (Boone, 2009).

To overcome the potential limitations of relying on a single PVT producing a dichotomous classification, a composite based on several independent effort indicators was developed for the purpose of this study (EI-7). Seven empirically supported PVTs were selected, and together labeled as the EI-7: the WMT, the RDS (Greiffenstein, Baker, & Gola, 1994; Heinly, Greve, Bianchini, Love, & Brennan, 2005; Jasinski, Berry, Shandera, & Clark, 2011; Larrabee, 2003; Mathias, Greve, Bianchini, Houston, & Crouch, 2002), the Word Choice Test (WCT; Pearson, 2008), the California Verbal Learning Test, Second Edition, forced choice recognition (CVLT-II FCR; Moore & Donders, 2004); Finger Tapping Test raw score of the dominant hand (FTT-DH; Arnold et al., 2005); and the Trail Making Test (TMT A & B raw scores; Iverson, Lange, Green, & Franzen, 2002). The FTT-DH and TMT A and B were chosen because together they tap domains that are closely related to the target constructs of the CPT-II (simple reaction time, visuomotor speed, inhibition, simultaneous processing of two classes of stimuli).

The resulting aggregate of independent PVTs has the advantage of broadly sampling test taking behavior over time while monitoring cognitive effort over a wide range of tests. Also, by rescaling several dichotomous PVTs into a composite resulting in interval scale, the underlying continuity in cognitive effort can be recaptured.

Recognizing that on most PVTs there are levels of failures, and that using a single cutoff (Pass/Fail approach) may produce an artificial dichotomy that results in a loss of clinical data about the examinee’s degree of engagement, each of the widely accepted cutoffs was assigned a separate value to better capture the effort gradient inherent in cognitive testing. A performance that passed the most liberal cutoff (i.e., very low probability of poor effort) was assigned the value of zero. The next most liberal cutoff (high sensitivity, low specificity) available in the literature was assigned the value of one, the next available cutoff was assigned a value of two, and finally, the most conservative (low sensitivity, high specificity) cutoff was assigned a value of three, reflecting a gradient of increasing confidence in the presence of poor effort.

Table 1 provides the key for transforming dichotomous cutoffs into a continuous scale as well as the observed base rate (BR_{observed}) at each cutoff in the sample. The WMT classified the largest proportion of the sample as invalid (39.4%), while the TMT-B had the lowest BR_{observed} of failure (5.8%). Table 2 provides a frequency, percentage, and cumulative percentage as well as a classification label for the first nine levels of the EI-7. Since the frequency at more extreme EI-7 values decreases rapidly, the table was truncated at 8.

An EI-7 value of zero means that the individual did not have any score on these seven PVTs that reached the most liberal of the cutoffs. Therefore, this subsample consistently demonstrated good effort and is considered an unequivocal PASS. Having obtained an EI-7 value of one is also considered a Pass, as failing one of the seven PVTs at the most liberal cutoff can happen for reasons other than poor effort. An EI-7 value of two or three starts to become problematic, as it implies either multiple PVT failures at the most liberal cutoff or failing a single PVT at a more conservative cutoff. Thus, this range is labeled Borderline. An EI-7 value of 4 or above is considered a Fail, with increasing confidence in the classification accuracy as the EI-7 value increases. These gradations of degree of certainty associated with each level of PVT failure are similar to the Slick, Sherman, and Iverson (1999) criteria for
malingered neurocognitive dysfunction (MND). However, given the lack of data on external incentive status, the original MND model could not be applied to the current sample.

Although the EI-7 was used as a primary index to establish the validity of a given neuropsychological profile, given its experimental nature, analyses based on it were repeated using the WMT to provide cross-validation with an established instrument. To maximize the purity of the reference groups used to establish cutoffs on the CPT-II subtests, borderline cases (EI-7 values 2–3) were excluded from these analyses, following recommendations by Greve and Bianchini (2004). Thus, participants with EI-7 scores of \( \leq 3 \) were classified as \textit{Pass}, while those with EI-7 scores of \( \geq 4 \) were classified as \textit{Fail}.

\[ \geq 4 \] cutoff protects against incidental PVT failures that are commonly considered to be insufficient evidence to label an entire neurocognitive profile as invalid. For example, performing very poorly on a single component of the EI-7 could result in a score of 3, even though technically it represents a single PVT failure, which is classified as a \textit{Pass} according to broadly accepted forensic standards. However, an EI-7 score of \( \geq 4 \) means at least two independent PVT failures, and hence, provides the minimum acceptable level of evidence for invalid responding.

\textit{Data Analysis}

Descriptive statistics (frequency, percentage, and cumulative percentage; mean, median, standard deviation, skew, kurtosis, and range) were computed for the key variables. Significance testing was performed using the \( F \) - and \( t \)-tests as well as \( \chi^2 \). Effect size was reported using partial \( \eta^2 \), Cohen’s \( d \), and \( \Phi^2 \). Confidence intervals (CIs) around point estimates were computed using the \( t \) distribution.

\[ \text{Base rate} \]

\[ \text{WMT failures}^a \]

\[ \text{Immediate recall, delayed recall, and consistency} \leq 82.5\% \]

\[ \text{Note: ACS-WCT = Advanced Clinical Solutions—Word Choice Test; CVLT-II FCR = California Verbal Learning Test, Second Edition forced choice recognition; FTT = finger tapping test; DH = dominant hand; RDS = reliable digit span; TMT = Trail Making Test; WMT = Green’s Word Memory Test.} \]

\[ \text{Note: The shading provides an additional visual cue for the increasing probability that the given response has been accurately classified as invalid.} \]

\[ \text{Table 1. The components of the Effort Index (EI-7), base rates for failing each cutoff and cumulative failure rates} \]

\[ \text{Table 2. Frequencies, percentages, cumulative percentages, and descriptive labels for the first nine levels of the EI-7} \]

\[ \text{Note: ACS-WCT = Advanced Clinical Solutions—Word Choice Test; CVLT-II FCR = California Verbal Learning Test, Second Edition forced choice recognition; FTT = finger tapping test; DH = dominant hand; RDS = reliable digit span; TMT = Trail Making Test; WMT = Green’s Word Memory Test.} \]

\[ \text{Note: Immediate recall, delayed recall, and consistency} \leq 82.5\% \]
Signal detection properties (sensitivity, specificity, positive [PPP], and negative predictive power [NPP]) were computed using the formula provided by Baldessarini, Finklestein, and Arana (1983) in addition to the overall area under the curve (AUC) statistic. AUC is the overall classification accuracy of the model. Sensitivity (or true positive [TP] rate) is the ratio of TP and the sum of TP and false negatives (FN). Specificity (or true negative [TN] rate) is the ratio of TN and the sum of TN and false positives (FP). In the present study, sensitivity is the probability that an instrument will correctly detect an invalid profile, while specificity is the probability that valid profiles will be identified as such. PPP is the ratio of TP and the sum of TP and FP, whereas NPP is the ratio of TN and the sum of TN and FN. In the context of performance validity assessment, PPP is the probability that an individual’s performance is invalid given that he or she failed the PVT. Conversely, NPP is the probability that an individual’s performance was valid given that he or she passed the PVT.

Likelihood ratios (LRs) were also reported given their utility in interpreting individual scores in clinical settings. The value of +LR expresses how much more likely an individual with a given condition (i.e., invalid profile) will obtain a positive result (i.e., Fail) compared with an individual without the condition (i.e., valid profile). Essentially, +LR is TP rate over FP rate. Conversely, –LR captures how much less likely an individual with a given condition (i.e., invalid profile) will obtain a negative result (i.e., Pass) compared with an individual without the condition (i.e., valid profile). In other words, –LR is the FN rate over TN rate. It follows from these definitions that higher values for +LR, and lower values for –LR reflect better discriminative power (Grimes & Schultz, 2005).

**Results**

A series of *t*-tests were performed using *Pass/Fail* status on the EI-7 and WMT as IV, and the 12 CPT-II subscales as DVs. All contrasts were significant except for β and Hit RT ISI Change. One-sample *t*-tests against the population mean of $T = 50$ revealed that the *Fail* group had significant elevations on all but the same two scales. The *Fail* group also produced more variable distributions on all CPT-II scales except for COM and $d'$. The largest effects ($d > 1.0$) were observed on the subscales that were previously identified in the literature as sensitive to poor effort: OMI, COM, HRT SE, VAR, and PER. A similar pattern, although with smaller effects, was observed using the WMT as IV. A summary of the analyses are provided in Table 3.

Next, the signal detection properties of various cutoffs on these five scales were examined in reference to the EI-7 and WMT. On OMI, >60 produced an AUC of 0.74–0.73, with good sensitivity (0.61–0.58), and specificity (0.88–0.89). At >65, AUC was essentially the same, with minimal decreases in sensitivity (0.57–0.55) and increases in specificity (0.91–0.92). At >80, AUC deteriorated (0.70–0.71), with notable losses in sensitivity (0.46–0.48) and modest gains in specificity (0.93–0.95). Raising the cutoff to >100 lowered AUC (0.69), achieving excellent specificity (0.98–1.00) at the cost of a substantial loss in sensitivity (0.39–0.38). Further increasing the cutoff to >110 produced no net gain. At the optimal cutoff of >65, +LR was 5.8, whereas –LR was 0.75.

**Table 3.** Independent *t*-tests comparing valid and invalid neuropsychological profiles on each of the CPT-II subscales with associated effect size measures, one-sample *t*-tests comparing each mean to the mean of $T = 50$ of the normative sample, and *F*-tests comparing sample variances

<table>
<thead>
<tr>
<th>CPT-II subtests (T-scores)</th>
<th>PASS (EI-7 ≤ 1)</th>
<th>FAIL (EI-7 ≥ 4)</th>
<th>$p$</th>
<th>$d_{EI-7}$</th>
<th>$d_{WMT}$</th>
<th>EI-7</th>
<th>WMT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><em>n</em> = 59</td>
<td><em>n</em> = 28</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$M$</td>
<td>$SD$</td>
<td>$M$</td>
<td>$SD$</td>
<td>$σ_1$ versus $σ_2$</td>
<td>$σ_1$ versus $σ_2$</td>
<td>$P$</td>
</tr>
<tr>
<td>OMI</td>
<td>50.8</td>
<td>18.2</td>
<td>114.5$^b$</td>
<td>84.0</td>
<td>&lt;.001</td>
<td>1.05</td>
<td>1.04</td>
</tr>
<tr>
<td>COM</td>
<td>49.9</td>
<td>9.4</td>
<td>60.8$^b$</td>
<td>11.4</td>
<td>&lt;.001</td>
<td>1.04</td>
<td>0.61</td>
</tr>
<tr>
<td>Hit RT</td>
<td>48.8</td>
<td>10.4</td>
<td>57.3$^b$</td>
<td>17.7</td>
<td>&lt;.01</td>
<td>.59</td>
<td>.59</td>
</tr>
<tr>
<td>Hit RT SE</td>
<td>50.9</td>
<td>11.6</td>
<td>70.4$^b$</td>
<td>20.9</td>
<td>&lt;.001</td>
<td>1.15</td>
<td>.97</td>
</tr>
<tr>
<td>Variability SE</td>
<td>50.7</td>
<td>9.3</td>
<td>68.0$^b$</td>
<td>17.9</td>
<td>&lt;.001</td>
<td>1.21</td>
<td>1.03</td>
</tr>
<tr>
<td>$d'$</td>
<td>50.1</td>
<td>9.8</td>
<td>56.5$^b$</td>
<td>7.9</td>
<td>&lt;.01</td>
<td>.72</td>
<td>.30</td>
</tr>
<tr>
<td>Response style (β)</td>
<td>49.1</td>
<td>12.0</td>
<td>50.9$^b$</td>
<td>8.2</td>
<td>.48</td>
<td>.18</td>
<td>.39</td>
</tr>
<tr>
<td>Perseverations</td>
<td>52.6</td>
<td>14.7</td>
<td>84.6$^b$</td>
<td>39.6</td>
<td>&lt;.001</td>
<td>1.07</td>
<td>.73</td>
</tr>
<tr>
<td>Hit RT BC</td>
<td>48.7</td>
<td>8.9</td>
<td>59.1$^b$</td>
<td>20.1</td>
<td>&lt;.05</td>
<td>.67</td>
<td>.49</td>
</tr>
<tr>
<td>Hit RT BC SE</td>
<td>52.1</td>
<td>9.5</td>
<td>63.3$^b$</td>
<td>19.8</td>
<td>&lt;.01</td>
<td>.72</td>
<td>.57</td>
</tr>
<tr>
<td>Hit RT ISI change</td>
<td>53.3$^*$</td>
<td>11.6</td>
<td>54.1</td>
<td>18.3</td>
<td>.80</td>
<td>.05</td>
<td>.04</td>
</tr>
<tr>
<td>Hit RT ISI change SE</td>
<td>49.6</td>
<td>10.2</td>
<td>57.7$^*$</td>
<td>18.7</td>
<td>&lt;.05</td>
<td>.54</td>
<td>.43</td>
</tr>
</tbody>
</table>

*Note:* *$^a$One-sample t*-test against the mean of $T = 50$, two tailed, $p < 0.05$.

*One-sample t*-test against the mean of $T = 50$, two tailed, $p < 0.01$. 

Archives of Clinical Neuropsychology 29 (2014) 456–466
On COM, a cutoff of >60 produced an AUC of 0.67–0.68, with both sensitivity (0.50–0.43) and specificity (0.88–0.89) in the acceptable range. Raising it to >65 resulted in a decrease in AUC (0.62–0.61), with an asymmetric trade-off: substantial loss in sensitivity (0.29–0.25) and modest increase in specificity (0.95–0.97). Increasing the cutoff to >70 further deflated the AUC (0.56), shrinking sensitivity (0.18–0.15) with no improvement in specificity. At the optimal cutoff of >60, +LR was 4.2, whereas –LR was 0.57.

A score of >60 on HRT SE resulted in AUC 0.75–0.68, with robust sensitivity (0.68–0.60), but unacceptably low specificity (0.81–0.77). While the overall cost of raising the cutoff to >65 was a negligible decrease in AUC (0.73–0.67) and a notable loss in sensitivity (0.54–0.48), the specificity improved significantly (0.92–0.86). At >70, AUC was restored (0.75–0.68), with both sensitivity (0.54–0.45) and specificity (0.97–0.91) in the acceptable range. This cutoff proved to be the point of diminishing returns: at >80 and >90 the AUC (0.68–0.63 & 0.55) and especially the sensitivity (0.36–0.30 and 0.11–0.10) declined precipitously. At the optimal cutoff of >70, +LR was 18.0, whereas –LR was 0.47.

On VAR, a cutoff of >60 produced the highest AUCs (0.75–0.68), trading specificity (0.86–0.84) for sensitivity (0.64–0.53). At >65, the benchmark specificity was reached (0.90–0.89), with a minimal sacrifice in AUC (0.72–0.68) and sensitivity (0.54–0.48). Further increasing the cutoffs (>70 and >75) consolidated the specificity (0.92 and 0.98, respectively) at the expense of both AUC (0.69–0.67 & 0.63–0.64) and sensitivity (0.46–0.43 & 0.29–0.30). At the optimal cutoff of >65, +LR was 5.4, whereas –LR was 0.51.

Finally, on PER, a cutoff of >60 had unacceptably low specificity (0.81–0.77) with good sensitivity (0.61–0.55) and acceptable overall AUC (0.71–0.66). Increasing the cutoff to >70 improved the AUC against the EI-7 (0.73), but not the WMT (0.65) with a loss in sensitivity (0.54–0.45) but a gain in specificity (0.91–0.84). Adjusting the cutoff to >80 resulted in little change in AUC (0.72–0.66), but consolidated the specificity (0.94–0.89) without sacrificing too much sensitivity (0.50–0.43). Further increases to >90 and >100 disproportionately favored specificity (0.98–0.97 in both cases) at the expense of sensitivity (0.36–0.38 and 0.32–0.33, respectively). At these highly conservative cutoffs, the AUCs trended downwards: 0.67 and 0.65, respectively, for both EI-7 and WMT. At the optimal cutoff of >70, +LR was 6.0, whereas –LR was 0.51. Table 4 and Table 6 provide a visual summary of these values.

Although β failed to show promise as an EVI based on the initial contrasts, given that it has been identified as a potential validity scale in the test manual, it was included in the signal detection analyses. Following Conners’ (2004) guidelines, scores between 40

Table 4. AUC, SENS, and SPEC at various cutoffs of embedded individual and composite CPT-II validity indicators against the EI-7 and WMT

<table>
<thead>
<tr>
<th></th>
<th>OMl-WMT</th>
<th>OMl-WMT</th>
<th>HRT SE</th>
<th>VAR</th>
<th>VAR</th>
<th>VAR</th>
<th>PER</th>
<th>CVI-5A</th>
<th>CVI-5B</th>
</tr>
</thead>
<tbody>
<tr>
<td>EI-7 &gt;60</td>
<td>.74</td>
<td>.73</td>
<td>.69</td>
<td>.68</td>
<td>.75</td>
<td>.68</td>
<td>.71</td>
<td>.66</td>
<td>.75</td>
</tr>
<tr>
<td>SENS</td>
<td>.61</td>
<td>.58</td>
<td>.50</td>
<td>.43</td>
<td>.66</td>
<td>.60</td>
<td>.64</td>
<td>.53</td>
<td>.61</td>
</tr>
<tr>
<td>SPEC</td>
<td>.88</td>
<td>.89</td>
<td>.88</td>
<td>.89</td>
<td>.81</td>
<td>.77</td>
<td>.86</td>
<td>.84</td>
<td>.81</td>
</tr>
<tr>
<td>&gt;65</td>
<td>.74</td>
<td>.74</td>
<td>.62</td>
<td>.61</td>
<td>.73</td>
<td>.67</td>
<td>.72</td>
<td>.68</td>
<td>.73</td>
</tr>
<tr>
<td>AUC</td>
<td>.57</td>
<td>.55</td>
<td>.29</td>
<td>.25</td>
<td>.54</td>
<td>.48</td>
<td>.54</td>
<td>.45</td>
<td>.54</td>
</tr>
<tr>
<td>SPEC</td>
<td>.91</td>
<td>.92</td>
<td>.95</td>
<td>.97</td>
<td>.92</td>
<td>.86</td>
<td>.90</td>
<td>.89</td>
<td>.91</td>
</tr>
<tr>
<td>&gt;80</td>
<td>.70</td>
<td>.71</td>
<td>.56</td>
<td>.56</td>
<td>.75</td>
<td>.68</td>
<td>.69</td>
<td>.67</td>
<td>.72</td>
</tr>
<tr>
<td>SENS</td>
<td>.46</td>
<td>.48</td>
<td>.18</td>
<td>.15</td>
<td>.54</td>
<td>.45</td>
<td>.46</td>
<td>.43</td>
<td>.50</td>
</tr>
<tr>
<td>SPEC</td>
<td>.93</td>
<td>.95</td>
<td>.95</td>
<td>.97</td>
<td>.97</td>
<td>.91</td>
<td>.92</td>
<td>.92</td>
<td>.94</td>
</tr>
<tr>
<td>&gt;100</td>
<td>.69</td>
<td>.69</td>
<td>.68</td>
<td>.63</td>
<td>.63</td>
<td>.64</td>
<td>.67</td>
<td>.67</td>
<td>.68</td>
</tr>
<tr>
<td>SENS</td>
<td>.39</td>
<td>.38</td>
<td>.36</td>
<td>.30</td>
<td>.29</td>
<td>.30</td>
<td>.36</td>
<td>.38</td>
<td>.40</td>
</tr>
<tr>
<td>SPEC</td>
<td>.98</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>&gt;110</td>
<td>.67</td>
<td>.68</td>
<td>.55</td>
<td>.55</td>
<td>.55</td>
<td>.55</td>
<td>.65</td>
<td>.65</td>
<td>.65</td>
</tr>
<tr>
<td>SENS</td>
<td>.36</td>
<td>.35</td>
<td>.11</td>
<td>.10</td>
<td>.11</td>
<td>.10</td>
<td>.32</td>
<td>.33</td>
<td>.33</td>
</tr>
<tr>
<td>SPEC</td>
<td>.98</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Note: AUC = area under the curve (overall classification accuracy); SENS = sensitivity; SPEC = specificity; EI-7 = Pass ≤1 and Fail = ≥4; WMT = Green’s Word Memory Test at Conventional Pass/Fail Classification; OMl = CPT-II omission errors (T-score); COM = CPT-II commission errors (T-score); HRT SE = CPT-II hit reaction time standard error (T-score); VAR = CPT-II Variability Standard Error (T-score); PER = CPT-II Perseverative Errors (T-score); CVI-5 = CPT-II Validity Indicator.

aOMI >65, COM >65, HRT SE >65, VAR >65, and PER >70.
bOMI >60, COM >60, HRT SE >60, VAR >60, and PER >60.
and 60 were defined as \textit{Pass}, and scores outside that range as \textit{Fail}. The scale performed poorly against both the EI-7 and the WMT (AUC: 0.49–0.54; sensitivity: 0.11–0.18; specificity: 0.88–0.91; +LR: 0.9–1.9; –LR: 1.0–0.9).

Replicating the methodology used to develop the EI-7, a composite of these five EVIs was created and was labeled “CPT-II Validity Index-5” (CVI-5). It ranges from 0 (none of the five EVIs reached the cutoff) to 5 (all EVIs were above the cutoff), and provides an index of cumulative EVI failures. Given the inescapable trade-off between high sensitivity and low specificity of liberal cutoffs and the opposite pattern associated with conservative ones, two different indices were developed, capitalizing on the strength of each of these strategies.

The first one, labeled CVI-5A represents a conservative approach. As such, each of its components reaches or exceeds the benchmark specificity of 0.90 against at least one of the reference PVTs: OMI \( > 65 \), COM \( > 65 \), HRT SE \( > 65 \), VAR \( > 65 \), and PER \( > 70 \). The purpose of this aggregate measure was to exploit the potential for improving signal detection performance of a multivariate model of validity assessment. Specifically, this index was developed to examine whether combining multiple indicators would increase the confidence in classifying a profile as invalid.

A CVI-5A value of 0 is an unequivocal \textit{PASS}. A value of 1 is considered \textit{Borderline}, given that any of the five EVI in the composite could individually raise legitimate concerns about the validity of the overall response set. A value of 2 is considered \textit{Fail}, with a specificity of 0.90 against the EI-7 and 0.87 against the WMT. Values of \( \geq 3 \) are considered to be unequivocal \textit{FAIL}, with increasing confidence in the classification accuracy with any additional CVI-5 component in the failing range (specificity: 0.98–0.92, and 1.00–0.98, respectively).

The second one was labeled CVI-5B and is based on the opposite approach: each constituent EVI is calibrated to maximize sensitivity. Therefore, the lowest reasonable cutoff of \( \geq 60 \) was used to define failure. As a result, three of the components did not reach the benchmark specificity of 0.90. The goal behind the CVI-5B was to explore the potential in aggregating sub-threshold EVI failures into a single composite that would have better signal detection properties than any of its individual component.

A CVI-5B value of 0 is an unequivocal \textit{PASS}. A value of 1 is still considered a \textit{Pass}, since three of its five components do not meet the minimum requirements to function as independent PVTs (sensitivity < 0.90). A CVI-5B value of 2 is considered \textit{Borderline} for the same reason, but it starts to raise questions about the credibility of the profile. A value of 3 is the first level of \textit{Fail}, given the cumulative evidence of questionable test performance. At \( \geq 4 \), the composite provides strong evidence of invalid response pattern (specificity: 97–0.94). Therefore, this level is labeled \textit{FAIL}. A CVI-5B value of 5 had a perfect specificity, meaning that all those who failed all of the individual EVIs also failed the reference PVTs.

Univariate ANOVAs found a comparably strong association between the levels of both CVI-5A and CVI-5B and the EI-7 (partial \( \eta^2 = 0.49 \), very large effect). Mean EI-7 values increased more slowly and leveled off sooner across levels of CVI-5B compared with its counterpart, which is to be expected given the more liberal cutoffs compared with the CVI-5A. The only significant difference between the two versions of the CVI-5 was observed at the value of 1. This pattern is predicted by the inner logic of the composites described above. The mean EI-7 values and the surrounding 95% CIs at each level provide further justification for the labeling: at 0, the upper limit is \( < 1.5 \) units, whereas at 5, the lower limit of the 95% CI is 2.9 and 5.0, respectively.

A similar pattern was found against the WMT using non-parametric significance tests. A gradual overall increase in failure rate was observed as CVI-5 values increased, with a very large effect (95% CI: 0.29) for both versions. As observed against the EI-7, at the highest level (all five indicators failed) the CVI-5 produced a perfect sensitivity. While the failure rate at the same CVI-5 value varied between the two versions, the discrepancy never reached statistical significance. Table 6 provides a visual summary of these analyses.

Although sensitivity and specificity are important features of an instrument, they are nested in group level analyses. In routine clinical practice, while evaluating the credibility of an individual response set, the most relevant test parameters are PPP and NPP (i.e., what a test score predicts in a given examinee). Therefore, those values were computed at the conventional hypothetical population base rates of 10%, 20%, 30%, 40%, and 50%. Results are displayed in Table 6. The expected changes in the confidence in classification accuracy as a function of cutoffs and base rates were observed.

Besides having better signal detection properties than all but one (HRT SE \( > 70 \)) of their individual components, the CVI-5s allow the assessor to harvest the compound discriminant power of multiple imperfect EVIs and transform them into a superior single index. Moreover, the two versions have been developed to detect different manifestations of invalid response pattern: the combined effect of multiple failed EVIs that provide evidence of non-credible responding (CVI-5A) or consistent sub-threshold level EVI failures that individually contain insufficient evidence to establish the invalidity of a given profile, but repeated failures allow the examiner to correctly identify non-credible responding (CVI-5B).

For example, if all of the constituent EVIs fall between a \( T \)-score of 61 and 65, they produce a CVI-5A value of 0, which is an unequivocal \textit{PASS}. The same pattern of performance produces a CVI-5B value of 5, which is an unequivocal \textit{FAIL} with a perfect specificity. Therefore, the clinician can confidently label the response set as invalid. As the example above suggests, the two versions of the CVI-5 complement each other and while they overlap significantly, they can provide non-redundant and clinically relevant information about the veracity of given response sets.
Finally, Table 7 illustrates the well-documented paradox that individuals with mTBI fail PVTs at a higher rate than those with mod/sev TBI. Goodness-of-fit tests produced medium effects with CVI-5s, EI-7, and WMT. Data on the EI-7 shows that the most extreme form of invalid responding was concentrated in the mTBI group.

Discussion

The present study explored the potential of aggregate EVIs within the CPT-II to improve the assessor’s ability to discriminate between valid and invalid response sets in adults referred for neuropsychological assessment following TBI. The results converge in a number of clinically relevant findings. (i) Two of the EVIs at the cutoffs originally proposed by Conners’ (OMI and PER T > 100) performed reasonably well in the current sample against other PVTs, but tended to sacrifice sensitivity for specificity. In contrast, β had unacceptably poor overall signal detection properties. (ii) In addition, other CPT-II scales previously identified in the literature as potential EVIs at various cutoffs (OMI, COM, PER, and HRT SE) also performed well at discriminating between valid
and invalid response sets in the present sample. The single best individual EVI was HRT SE > 70. (iii) The practice of aggregating EVIs advocated by previous research (Lange et al., 2013; Larrabee 2003, 2012a, 2012b; Ord et al., 2010) also proved effective in our sample. The CVI-5s not only increased the overall classification accuracy but also improved the ecological validity of the final decision, as it sampled multiple facets of sustained attention to determine the validity of a given response set. (iv) The CPT-II appears to be more sensitive to invalid responding than acquired attention deficit when used in a TBI population. This finding is consistent with previous reports (Brenner et al., 2009; Skandsen et al., 2010) and has two main implications. On the one hand, the CPT-II could be co-opted as a PVT. On the other hand, given the remarkably intact CPT-II profiles of those who produced valid response sets, it could be useful in ruling out residual deficits in processing speed, vigilance, inhibition and sustained attention following TBI. (v) As commonly reported, individuals with mTBI failed reference PVTs at significantly higher rates than those with more severe injuries, as well as failing the CVI-5s at higher rates.

An in-depth analysis of the 12 CPT-II scales sheds some light on the success of the OMI and PER scales at identifying invalid response sets. These are the only two scales on which extreme error rates can be easily achieved. Producing a severely impaired score (i.e., high T-scores) on OMI simply requires failing to respond to a target, of which there are many. Likewise, one can obtain an extreme elevation on PER through incessant/random responding, pressing the button often, disregarding the actual stimulus. Again, there are ample opportunities to “score” on this scale during the 14-min administration time. Moreover, what constitutes a “failure” on these two parameters is clear. The other 10 scales, however, have built-in physiological or statistical ceilings that are virtually impossible to penetrate. In addition, several of these scales are derived measures that may be difficult to conceptualize even for trained professionals who are well versed in the signal detection paradigm. For example, it is quite easy to inflate OMI errors by disengaging from the test, but it is more challenging to directly manipulate standard deviations, standard errors or discriminability indices. While extreme response patterns on the OMI and PER scales can affect the derived signal detection parameters, they are exceedingly difficult to influence deliberately.

Given the inherent trade-off among combinations of cutoffs, a flexible application of the signal detection properties of a list of cutoffs appears to be the most sensible approach in a clinical setting. While individual CPT-II scales at various cutoffs continue to be useful at assessing the validity of the overall profile, there are clear rational and empirical advantages of aggregating EVIs and making the final decision with respect to validity based on a composite score. First, an aggregate measure relies on input from multiple different sources, reducing the influence of random findings and generating a more stable estimate of performance validity. Thus, it offers a more representative index by sampling multiple facets of the target constructs. Second, classification accuracy data show that the CVI-5s performed better than most of the individual cutoffs.

Also, given that the two versions of the CVI-5 were specifically developed to detect different manifestations of invalid responding, they offer matching alternative strategies to detect them. The CVI-5A was designed to aggregate the cumulative evidence of more extreme impairment in fewer facets of sustained visual attention. In contrast, the CVI-5B was purposefully calibrated to be overly sensitive. As such, it monitors sub-threshold level (i.e., suspect) performance patterns that are not sufficiently extreme to be classified as invalid on their own, but clearly deviate from the centroid of valid responders. Signal detection analyses suggest that even mild impairment (i.e., barely reaching the clinical cutoff) on a certain combination of the five EVIs provides evidence of invalid responding. Although the rationale behind this mechanism may not be apparent, a string of T-scores between 61 and 65 on

<table>
<thead>
<tr>
<th>CVI-5</th>
<th>EI-7</th>
<th>WMT</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td></td>
</tr>
<tr>
<td>Mild</td>
<td>Mod/sev</td>
<td>Mild</td>
</tr>
<tr>
<td>0</td>
<td>62.3&lt;sup&gt;a&lt;/sup&gt;</td>
<td>37.7&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>1</td>
<td>80.0</td>
<td>20.0</td>
</tr>
<tr>
<td>2</td>
<td>100.0</td>
<td>0.0</td>
</tr>
<tr>
<td>3</td>
<td>80.0</td>
<td>20.0</td>
</tr>
<tr>
<td>4</td>
<td>100.0</td>
<td>0.0</td>
</tr>
<tr>
<td>5</td>
<td>74.8</td>
<td>25.2</td>
</tr>
<tr>
<td>x²</td>
<td>11.6</td>
<td>9.0</td>
</tr>
<tr>
<td>p</td>
<td>.04</td>
<td>.11</td>
</tr>
<tr>
<td>ϕ²</td>
<td>0.11</td>
<td>0.09</td>
</tr>
</tbody>
</table>

<sup>a</sup>x²(1) goodness of fit against the proportion of mTBI and mod/sevTBI in the sample, p < .05.
these scales represents a clinically implausible pattern of performance: equally elevated OMI and COM errors, overall and block-by-block response time variability in conjunction with sub-threshold level perseverative/random responding.

The side-by-side comparison between the EI-7 and WMT provides further evidence for the enhanced utility of composites versus stand-alone PVTs. Although the former incorporates the information contained by the latter, it parcels out the data provided by its monolithic approach and adds data from other independent PVTs. As a result, the difference between the valid and invalid subsamples on the CPT-II subtests is generally more pronounced when the criterion PVT is the EI-7 as opposed to the WMT, as captured by the differences in effect size estimates (Table 3), as well as AUCs, sensitivity, and specificity (Table 4).

Another advantage of the multivariate approach to effort assessment from a practitioner’s perspective is the inherent flexibility of the model. Namely, it allows the user to choose a more liberal or conservative cutoff to define poor effort as well as from five different hypothetical base rates of invalid responding, to match the expectations and decision-making criteria of unique assessment settings. This level of customization has the potential to improve the classification accuracy at the individual level by applying a more nuanced assessment of performance validity.

The present study has several strengths. It used a large clinical sample with a representative FSIQ, age, education and handedness distribution, and standard neuropsychological instruments. It empirically validated cutoffs published in the technical manual and subsequent research studies, introduced alternative combination of cutoffs, and reported detailed signal detection properties of both individual EVIs and two new composites (CVI-5 A and B) in the CPT-II.

The study also has a number of limitations that should be taken into consideration. The participants were referred for neuropsychological assessment in the context of TBI. Hence, the results may not generalize to individuals with other conditions such as ADHD. The sample only contained adults, even though the CPT-II was normed on children as young as 6 years of age and continues to be a widely used instrument in pediatric populations. The effect of administration sequence was not modeled systematically despite previous reports that time-related changes can influence CPT-II scores (Erdodi & Lajiness-O’Neill, 2014; Erdodi, Lajiness-O’Neill & Saules, 2010). Thus, further research replicating the present methodology in samples with different clinical diagnoses, age ranges and PVTs to cross-validate the EVIs in the CPT-II is necessary to establish the generalizability of the findings.

In summary, aggregating EVIs in the CPT-II improves the overall efficiency of the validity assessment on both theoretical and empirical grounds. In order to maximize the classification accuracy of a given response set, a flexible approach is recommended, which takes advantage of the opportunity to refine the signal detection accuracy of these EVIs using different levels and combinations of cutoffs. Future research exploring different combination of EVIs and cutoff as well as reference PVTs in TBI and other clinical populations is needed to increase the confidence in the multivariate approach to calibrating EVIs.

References


