

Multivariate examination of embedded indicators of performance validity for ADHD evaluations: A targeted approach

John-Christopher A. Finley, Julia M. Brooks, Amanda N. Nili, Alison Oh, Hannah B. VanLandingham, Gabriel P. Ovsiew, Devin M. Ulrich, Zachary J. Resch & Jason R. Soble

To cite this article: John-Christopher A. Finley, Julia M. Brooks, Amanda N. Nili, Alison Oh, Hannah B. VanLandingham, Gabriel P. Ovsiew, Devin M. Ulrich, Zachary J. Resch & Jason R. Soble (13 Sep 2023): Multivariate examination of embedded indicators of performance validity for ADHD evaluations: A targeted approach, *Applied Neuropsychology: Adult*, DOI: [10.1080/23279095.2023.2256440](https://doi.org/10.1080/23279095.2023.2256440)

To link to this article: <https://doi.org/10.1080/23279095.2023.2256440>



Published online: 13 Sep 2023.



Submit your article to this journal [↗](#)



Article views: 204



View related articles [↗](#)





View Crossmark data [↗](#)



Citing articles: 8 View citing articles [↗](#)



Multivariate examination of embedded indicators of performance validity for ADHD evaluations: A targeted approach

John-Christopher A. Finley^a , Julia M. Brooks^{b,c}, Amanda N. Nili^{b,d}, Alison Oh^{b,e}, Hannah B. VanLandingham^{b,e}, Gabriel P. Ovsiew^b, Devin M. Ulrich^b, Zachary J. Resch^b, and Jason R. Soble^{b,f} 

^aDepartment of Psychiatry and Behavioral Sciences, Northwestern University Feinberg School, Chicago, IL, USA; ^bDepartment of Psychiatry, University of Illinois College of Medicine, Chicago, IL, USA; ^cDepartment of Psychology, University of Illinois at Chicago, Chicago, IL, USA; ^dDepartment of Medical Social Sciences, Northwestern University Feinberg School, Chicago, IL, USA; ^eDepartment of Psychology, Illinois Institute of Technology Chicago, IL, USA; ^fDepartment of Neurology, University of Illinois College of Medicine, Chicago, IL, USA

ABSTRACT

This study investigated the individual and combined utility of 10 embedded validity indicators (EVIs) within executive functioning, attention/working memory, and processing speed measures in 585 adults referred for an attention-deficit/hyperactivity disorder (ADHD) evaluation. Participants were categorized into invalid and valid performance groups as determined by scores from empirical performance validity indicators. Analyses revealed that all of the EVIs could meaningfully discriminate invalid from valid performers (AUCs = .69-.78), with high specificity ($\geq 90\%$) but low sensitivity (19%-51%). However, none of them explained more than 20% of the variance in validity status. Combining any of these 10 EVIs into a multivariate model significantly improved classification accuracy, explaining up to 36% of the variance in validity status. Integrating six EVIs from the Stroop Color and Word Test, Trail Making Test, Verbal Fluency Test, and Wechsler Adult Intelligence Scale-Fourth Edition was as efficacious (AUC = .86) as using all 10 EVIs together. Failing any two of these six EVIs or any three of the 10 EVIs yielded clinically acceptable specificity ($\geq 90\%$) with moderate sensitivity (60%). Findings support the use of multivariate models to improve the identification of performance invalidity in ADHD evaluations, but chaining multiple EVIs may only be helpful to an extent.

KEYWORDS

ADHD; embedded validity indicators; neuropsychology; performance validity; symptom validity testing

Introduction

Fabrication and exaggeration of attention-deficit/hyperactivity disorder (ADHD) symptoms have become a serious issue in healthcare given the high rate of misdiagnosis and inaccurate treatment (Cook et al., 2018; Rabiner, 2013). Prior estimates have suggested that 10%–48% of adults undergoing evaluation for known or suspected ADHD perform invalidly on neurocognitive testing (e.g., Abramson et al., 2023; Hirsch et al., 2022; Martin & Schroeder, 2020; Musso & Gouvier, 2014; Ovsiew et al., 2023; White et al., 2022). Methodological differences contribute to this variability in performance invalidity base rates, with newer studies using contemporary standards for classifying invalid performance generally yielding base rates of 10%–20% (see Ovsiew et al., 2023). Examinees are often incentivized by medications and academic accommodations to feign impairment in a sophisticated manner (Lovett & Harrison, 2021), and simulation studies have demonstrated that coached examinees will perform poorly on select measures of cognitive ability (Booksh et al., 2010). Learning how to successfully feign a diagnosis of ADHD may also be easier with more access to online information. It is therefore incumbent

on researchers to cultivate better validity assessment methods for ADHD evaluations to reduce undue healthcare costs and iatrogenic effects from inappropriate treatment (Lakhan & Kirchgessner, 2012).

Clinicians are recommended to employ several types of performance validity tests (PVTs), including embedded validity indicators (EVIs; i.e., cut-scores within tests assessing various cognitive functions) and freestanding validity tests (i.e., tests solely dedicated to indexing performance validity), among other measures and information (e.g., medical history and behavioral presentation) to assess the validity of performance data (Suhr & Berry, 2017; White et al., 2022). Despite the broad availability of PVTs, it remains challenging to determine when impaired performance is genuine or feigned among individuals undergoing ADHD evaluations (Marshall et al., 2016). For these reasons, it is increasingly important to identify PVTs that can accurately discriminate between genuine and feigned impairment that is commonly observed in ADHD evaluations. ADHD has been associated with genuine deficits in executive functioning, attention/working memory, and processing speed (Theiling & Petermann, 2016; Willcutt et al., 2005; Woods et al., 2002). So, individuals may be inclined to feign impairment on

these types of measures (e.g., Fuermaier et al., 2017; Quinn, 2003; Scimeca et al., 2021). There are some existing validity indicators that are embedded within measures of executive functioning, attention/working memory, and processing speed, but few of these EVIs have been validated in ADHD populations (Ausloos-Lozano et al., 2022; Bing-Canar et al., 2022; Khan et al., 2022; Kosky et al., 2022; Scimeca et al., 2021; Wallace et al., 2019; Williamson et al., 2014). Complicating matters further, many of these EVIs inadequately detect invalid performance (Whiteside et al., 2019). These EVIs may have poor classification accuracy because they are confounded by the effects of genuine cognitive and neuropsychiatric dysfunction associated with ADHD (Willcutt et al., 2005). Similar problems have been observed among EVIs in other populations where mild cognitive and neuropsychiatric impairment is pervasive (e.g., Davis, 2018; Fazio et al., 2019; Finley et al., 2023; Glassmire et al., 2019).

Using multiple EVIs in an evaluation may increase the sensitivity of detecting invalid performance (Larrabee, 2003a, 2008, 2014; Victor et al., 2009), although the EVIs must be selected and interpreted systematically (Soble et al., 2020). It is important to first identify the most sensitive EVIs that can be used together (Berthelson et al., 2013; Bilder et al., 2014; Boone, 2021; Erdodi, 2023). The next step is to determine whether the EVIs are capturing unique information. Integrating multiple EVIs that index redundant information can bias the overall interpretation of validity status because it artificially inflates the number of positive or negative findings (Silk-Eglit et al., 2015). This is a common issue in performance validity assessment since most PVTs are moderately correlated (Berthelson et al., 2013).

Boone (2013) recommended using validity indicators from measures of different cognitive domains to avoid redundancy rather than relying on multiple measures within one domain. While it is important to use multiple validity indicators throughout an assessment and across different domains, it may also be helpful to assess the validity of performance within a single domain using several EVIs (Erdodi, 2019) given the concern that patients may feign select cognitive deficits. Some studies have identified non-redundant validity measures that can be used together within specific cognitive domains, including memory (Schutte et al., 2011), language (Whiteside et al., 2015), and visuospatial abilities (Whiteside et al., 2011). Only two studies have identified multiple EVIs within measures of executive functioning, attention/working memory, and processing speed that can be used together. Whiteside et al. (2019) demonstrated that using EVIs within the Stroop Color and Word Test, Wisconsin Card Sorting Test, and Trail Making Test-B in combination versus in isolation improved classification accuracy in persons with traumatic brain injury. Moreover, White et al. (2020) showed that EVIs within the Trail Making Test-A and -B, Verbal Fluency, and Stroop Color and Word Test improved invalidity detection in a mixed neuropsychiatric sample. These studies provide evidence that it may be beneficial to use multiple EVIs within measures of executive functioning, attention/working memory, and processing speed together rather than individually,

though neither of these studies focused on ADHD populations.

It is also important to consider how many EVIs should be used together to maximize discrimination of valid and invalid performance. Indiscriminately including more EVIs can increase the chance of false-positive findings (Erdodi, 2019; Sherman et al., 2020), whereas including more free-standing validity tests can increase examination length and cost. To date, the ideal number of EVIs to administer in an ADHD evaluation has not been empirically determined. Depending on how many EVIs are used together, the number of failures among such EVIs may have different implications regarding the classification accuracy of invalid performance (Sherman et al., 2020). As a heuristic in performance validity methodology, failing two or more PVTs should generally result in identification of invalid performance (Sweet et al., 2021). However, failing three or more PVTs may be a better criterion for identifying invalid performance in cases when there are several validity tests administered together, such as 10 or more PVTs (Larrabee et al., 2019). While still up for debate, there may also be instances where it is most important to consider the number of failures relative to the number of tests administered (Sherman et al., 2020). For instance, failing two out of 15 EVIs may be less indicative of invalid performance than failing two out of four EVIs (i.e., 13% compared to a 50% failure rate). It has been proposed that failing two or more out of four to nine PVTs is an appropriate proportion of failures to determine invalid performance, with the most parsimonious model suggesting two out of five (Meyers & Volbrecht, 2003; Sherman et al., 2020; Victor et al., 2009). Few studies, much less with ADHD samples, have examined the accuracy of a multivariate model of EVIs with regard to the proportion of failures.

Overall, the assessment of performance validity in ADHD evaluations is improving but, as indicated above, several gaps remain in the literature. This study first sought to identify whether cut-scores for 10 EVIs within measures of executive functioning, attention/working memory, and processing speed had adequate classification accuracy for ADHD evaluations. We hypothesized that all of the EVIs included in this study would demonstrate acceptable classification accuracy since they have previously been described in the literature as adequate indicators of performance validity for ADHD populations. The second aim was to identify which combination of EVIs could optimize the performance validity classification. We hypothesized that integrating multiple EVIs would result in better classification accuracy relative to any EVI used in isolation. More specifically, we hypothesized that the best combination of EVIs would only include a select number of the 10 EVIs with the highest individual classification accuracy statistics. We also hypothesized that the best combination of EVIs would be a select number of EVIs derived from measures that have diverse and non-redundant testing paradigms. For instance, EVIs within speeded visuomotor scanning and cognitive flexibility measures (e.g., Trail Making Test-B) may work well with EVIs within non-speeded auditory attention and working memory

measures (e.g., Letter-Number Sequencing) because their paradigms are discrete and likely do not index the same aspects of performance validity. In addressing our second aim, we were particularly interested in understanding which number and proportion of validity indicator failures would be required to optimize sensitivity and specificity when using a combination of multiple EVIs.

Method

Participants and procedures

This study examined cross-sectional data from 599 adult outpatients who were consecutively referred for neuropsychological evaluation at a Midwestern academic medical center for diagnostic classification and treatment planning related to ADHD from 2018–2023. The evaluation included a review of medical and academic records, semi-structured clinical interview, a validity-controlled assessment of self-reported ADHD symptoms (Clinical Assessment of Attention Deficit-Adult; Bracken & Boatwright, 2005) and self-reported psychopathology (Minnesota Multiphasic Personality Inventory-2-Restructured Form; Ben-Porath & Tellegen, 2008), and assessment of neurocognitive functioning using a comprehensive battery of standardized performance tests. ADHD and other mental health diagnoses were classified according to the *Diagnostic and Statistical Manual of Mental Health Disorders-Fifth Edition* (American Psychiatric Association, 2013), which was based on information gleaned from the clinical interview and self-report questionnaires. All patients provided written informed consent to include their data as part of an ongoing study that has been IRB-approved. Some data from this sample has been used in other publications (i.e., Abramson et al., 2023; Ausloos-Lozano et al., 2022; Bing-Canar et al., 2022; Khan et al., 2022; Ovsiew et al., 2023; Phillips et al., 2023; Scimeca et al., 2021; Skymba et al., 2023; Ka Yin Tse et al., 2023); however, due to ongoing and continuous data collection for the database, it is not possible to determine the extent of overlap with prior studies.

Patients with missing data ($n = 14$) were excluded from the study, which resulted in a final sample of 585. Approximately 30% of the sample had a diagnosis of ADHD ($n = 176$), 46% had a diagnosis of ADHD and a psychiatric disorder ($n = 268$), 19% had a diagnosis of a psychiatric disorder ($n = 109$; primary diagnoses included depression [$n = 43$], posttraumatic stress disorder [$n = 26$], anxiety [$n = 26$], somatic symptom disorder [$n = 11$], and bipolar II [$n = 3$]), and 5% did not have any mental health diagnosis ($n = 32$). No patients were diagnosed with an intellectual disability, major neurocognitive disorder, or severe mental illness. One patient in the sample was pursuing disability during the time of their evaluation, but that patient did not demonstrate invalid performance on any of the PVTs used in this study. See Table 1 for sample demographics.

Based on current practice standards (e.g., Jennette et al., 2022; Rhoads et al., 2021; Sweet et al., 2021), patients with 0–1 criterion PVT failures were classified as valid performers and those with two or more criterion PVT failures were

classified as invalid performers. Using a reference standard consisting of two freestanding and three embedded criterion PVTs validated among ADHD populations, 515 (88%) patients were categorized into the valid performance group and 70 (12%) into the invalid performance group. See Table 2 for more details regarding the criterion PVTs.

Measures

This study examined the utility and viability of the following measures as potential validity indicators: Coding, Symbol Search, and Letter-Number Sequencing from the Wechsler Adult Intelligence Scale-Fourth Edition (WAIS-IV; Wechsler, 2008); Lexical Fluency (FAS; Heaton et al., 2004); Trail Making Test-A and -B (TMT-A and TMT-B; Heaton et al., 2004); Word Reading, Color Naming, and Color-Word from the Stroop Color and Word Test (SCWT; Golden, 1978); and the number of Omissions from the Conner's Continuous Performance Test-Third Edition (CPT-3; Conners, 2014). These measures were selected because prior work has identified cut-scores within them that can be used as EVIs for ADHD populations (Ausloos-Lozano et al., 2022; Bing-Canar et al., 2022; Khan et al., 2022; Scimeca et al., 2021). Additionally, they are known to index aspects of executive functioning, attention/working memory, and processing speed, including basic and complex visuomotor processing speed (WAIS-IV Symbol Search and Coding, TMT-A, SCWT Word Reading and Color Naming), lexical fluency (FAS), working memory (WAIS-IV Letter-Number Sequencing), visuomotor sequencing, cognitive flexibility, and complex attention (TMT-B), response inhibition (SCWT Color-Word), and selective and sustained attention (CPT-3 Omissions). Demographically adjusted standardized scores were used for each measure since these types of scores were reported in prior studies as viable metrics for PVTs (see Table 2).

Given the number of participants with psychiatric comorbidities, we compared symptoms of internalizing psychopathology between the validity groups using the Beck Depression Inventory-2nd Edition (Beck et al., 1996) and the Beck Anxiety Inventory (Beck et al., 1993). The Beck Depression Inventory-2nd Edition is a 21-item questionnaire assessing symptoms of depression over the past two weeks. The Beck Anxiety Inventory is a 21-item questionnaire assessing symptoms of anxiety over the past week.

Statistical analyses

All statistical assumptions were met, and post-hoc power analyses indicated our findings had an observed power greater than 80%. Posthoc power analyses were conducted through a statistical software package in R with consideration of the significance level set at .05. The effect size measures were determined based on the magnitude of the observed effects from the independent sample *t*-tests, chi-square tests, correlation analyses, and logistic regression analyses that are described below. Preliminary analyses included independent sample *t*-tests and chi-square tests to

Table 1. Paired Sample t-tests between participants with valid and invalid neurocognitive performance.

Measures	Valid Performers (<i>n</i> = 515)	Invalid Performers (<i>n</i> = 70)	Effect Sizes (<i>d</i> or <i>V</i>)
	Mean (SD; Range)	Mean (SD; Range)	
Age	<i>M</i> = 28.12 (6.85; 18–60)	<i>M</i> = 27.54 (7.82; 18–55)	.07
Education	<i>M</i> = 15.01 (2.01; 11–20)	<i>M</i> = 14.81 (2.42; 8–18)	.33**
Female Sex	319 (62%)	33 (47%)	.09*
Racial Identity			
White	240 (47%)	27 (39%)	.05
Black	79 (15%)	13 (19%)	.02
Hispanic	111 (22%)	21 (30%)	.06
Asian	53 (10%)	7 (10%)	.00
Multiracial	32 (6%)	2 (2%)	.04
Psychiatric Symptoms			
BDI-II Scores	<i>M</i> = 18.96 (11.73; 0–56)	<i>M</i> = 19.27 (13.24; 0–50)	.01
BAI Scores	<i>M</i> = 12.63 (9.23; 0–46)	<i>M</i> = 13.85 (12.69; 0–47)	.13
Tests with Embedded Validity Indicators			
WAIS-IV Symbol Search	<i>M</i> = 11.27 (3.05; 4–19)	<i>M</i> = 7.06 (3.54; 2–16)	.64***
WAIS-IV Coding	<i>M</i> = 10.20 (2.60; 3–19)	<i>M</i> = 7.92 (3.18; 1–14)	.58***
WAIS-IV Letter-Number Sequencing	<i>M</i> = 9.95 (2.39; 5–19)	<i>M</i> = 7.79 (2.44; 1–17)	.63***
Lexical Fluency FAS	<i>M</i> = 47.24 (10.43; 11–79)	<i>M</i> = 41.55 (11.98; 23–86)	.41***
TMT-A	<i>M</i> = 48.61 (11.98; 3–83)	<i>M</i> = 40.04 (11.46; 11–71)	.54***
TMT-B	<i>M</i> = 46.96 (10.64; 7–82)	<i>M</i> = 36.41 (10.96; 13–59)	.64***
SCWT Word Reading	<i>M</i> = 41.01 (11.48; 14–73)	<i>M</i> = 30.43 (13.31; 14–60)	.54***
SCWT Color Naming	<i>M</i> = 43.24 (10.38; 17–80)	<i>M</i> = 33.51 (10.44; 7–56)	.64***
SCWT Color-Word	<i>M</i> = 49.43 (10.73; 20–84)	<i>M</i> = 39.71 (12.61; 18–88)	.56***
CPT-3 Omissions	<i>M</i> = 49.53 (9.67; 41–90)	<i>M</i> = 54.82 (13.62; 43–90)	.33***
Criterion Performance Validity Tests			
BVMT-R Recognition Discrimination	<i>M</i> = 5.92 (0.30; 4–6)	<i>M</i> = 5.03 (1.22; 0–6)	.59***
Dot Counting Test E-Score	<i>M</i> = 9.00 (2.38; 4–23)	<i>M</i> = 15.67 (6.71; 8–49)	.63***
Reliable Digit Span	<i>M</i> = 10.38 (2.06; 6–17)	<i>M</i> = 7.91 (1.86; 5–12)	.73***
Rey 15-Item Test Recall + Recognition	<i>M</i> = 28.80 (1.51; 17–30)	<i>M</i> = 25.66 (3.91; 14–30)	.68***
RAVLT Effort Score	<i>M</i> = 17.64 (2.70; 3–20)	<i>M</i> = 12.18 (4.39; 0–20)	.76***

Note: *M*: Mean; *SD*: Standard deviation; Effect Sizes: Cramér's *V* or Cohen's *d*; BDI-II: Beck Depression Inventory-2nd Edition; BAI: Beck Anxiety Inventory; WAIS-IV: Wechsler Adult Intelligence Scale-4th Edition; TMT: Trail Making Test; SCWT: Stroop Color and Word Test; CPT-3: Conner's Continuous Performance Test-Third Edition; BVMT-R: Brief Visuospatial Memory Test-Revised; RAVLT: Rey Auditory Verbal Learning Test.

p* < .05; *p* < .01; ****p* < .001.

Table 2. Criterion performance validity tests.

Performance Validity Test	Cutoff	SN	SP	Failure Cutoff Reference(s)	Sample Failure Rate
BVMT-R Recognition Discrimination	≤5 (raw)	.35	.90	Phillips et al. (2023)	77/585 (13%)
Dot Counting Test E-Score	≥14 (raw)	.54	.92	Abramson et al. (2023)	60/585 (10%)
Reliable Digit Span (Forward/Backward)	≤7 (raw)	.22–.35	≥.93	Bing-Canar et al. (2022); Marshall et al. (2010)	59/585 (10%)
Rey 15-Item Test Recall + Recognition	≤23 (raw)	.50	.91	Ashendorf et al. (2021)	22/585 (4%)
RAVLT Effort Score	≤13 (raw)	.41	.90	Tse et al. (2023)	80/585 (14%)

Note: BVMT-R: Brief Visuospatial Memory Test-Revised; RAVLT: Rey Auditory Verbal Learning Test.

examine demographic and clinical characteristics between the valid and invalid performers, and bivariate correlations to examine relationships among the predictor and criterion validity tests.

For the primary analyses, receiver operating characteristic curve analyses were performed to derive cut-scores at which each EVI could optimally discriminate invalid from valid performers while maintaining ≥.90 specificity. It was predetermined that cutoffs had to yield statistically significant areas under the curve (AUCs) to be retained for follow-up analyses. Simple logistic regression analyses were then conducted to further examine the predictive ability of each EVI cutoff and indicate how much variance in criterion performance validity status each EVI explained individually. For the logistic regressions, we used the binary cutoffs from the EVIs that were gleaned from the above analyses.

Multiple logistic regression analyses were then conducted to examine how much variance in criterion performance validity status the EVI cutoffs explained in combination. We

examined two combinations of the EVIs. The first combination of EVIs was based on the *best subset selection method* (for review, see Kassambara, 2018, or James et al., 2013). This method uses a variable selection approach to derive the most parsimonious combination of EVIs required to minimize redundancy and maximize sensitivity while maintaining ≥.90 specificity. Specifically, it considers all possible models with a select number of variables (EVI) and their mean square prediction error to balance variable count and prediction error. The fitted models are then evaluated based on a number of criterion values, such as prediction error and Akaike information criterion. This statistical analysis was conducted using the *Leaps* package from R (Lumley & Lumley, 2013). This best subset selection method was appropriate for our study and negated the common pitfalls of stepwise approaches because our model examined relatively few predictor EVIs among a large sample (Snyder & Lawson, 1993; Thompson, 1995). The second combination was a “kitchen sink” model comprising all 10 of the EVIs.

Table 3. Embedded validity indicators intercorrelations in the valid performing group ($n = 515$).

	Coding	Symbol Search	LNS	Lexical Fluency FAS	TMT-A	TMT-B	SCWT Word	SCWT Color	SCWT Color-Word	CPT-3 Omissions
Coding	–	.52***	.12**	.13**	.40***	.44***	.35***	.38***	.38***	–.22***
Symbol Search		–	.15***	.16***	.37***	.42***	.38***	.42***	.38***	.24***
LNS			–	.23***	.14***	.24***	.16***	.22***	.24***	–.14*
Lexical Fluency FAS				–	.23***	.24***	.24***	.27***	.25***	–.03
TMT-A					–	.48***	.33***	.31***	.32***	–.11*
TMT-B						–	.30***	.36***	.36***	–.17***
SCWT Word							–	.67***	.52***	–.12**
SCWT Color								–	.70***	–.17***
SCWT Color-Word									–	–.19***
CPT-3 Omissions										–

Note: LNS: Letter-Number Sequencing; TMT: Trail Making Test; SCWT: Stroop Color and Word Test; CPT-3: Conner's Continuous Performance Test-Third Edition.

* $p < .05$; ** $p < .01$; *** $p < .001$.

Both multivariate models were examined with the binary cutoffs from the EVIs. In sum, these statistical procedures were conducted to identify a multivariate model that was based on quality (best subset selection model) versus a model based on quantity (kitchen sink model).

We then compared the efficacy of the univariate and multivariate models based on Akaike information criterion as well as examination of model fit statistics, including maximum likelihood estimation, (pseudo) variance explained, and classification accuracy statistics. Models with lower Akaike information criterion values were considered to have a better balance between goodness of fit and parsimony. Given the uncertainty in the number and proportion of cut-off failures that should be used in the clinical assessment of performance validity for multiple EVIs (Larrabee et al., 2019; Sherman et al., 2020), we also compared the classification accuracy statistics based on the number and proportion of failures within and between the multivariate models. AUC values of .70–.79 were considered acceptable classification accuracy, values of .80–.89 were considered excellent, and values of $\geq .90$ were considered outstanding (Hosmer et al., 2013). Because these multivariate models may include overly optimistic measures, we used bootstrap resampling to create 1,000 different samples that yielded a C index, which was compared to the receiver operating characteristic curve of the two multivariate models, to assess reliability.

This study's design and analysis plan were not preregistered. Neither the data nor the materials have been made available on a permanent third-party archive.

Results

Criterion performance validity tests and validity status

Approximately 12% of the sample ($n = 70$) demonstrated invalid performance based on performing below the cut-score on ≥ 2 criterion PVTs. Specifically, 60% of patients within the invalid group ($n = 42/70$) performed below two criterion PVT cut-scores, 26% ($n = 18/70$) performed below three cut-scores, 10% ($n = 7/70$) performed below four cut-scores, and 4% ($n = 3/70$) performed below all five cut-scores. In the valid group, 77% ($n = 397/515$) of the patients did not perform below any criterion PVT cut-scores, and 23% ($n = 118/515$) performed below one cut-score. The criterion PVTs were modestly correlated with each other (r s ranged from $-.03$ to $-.20$) and all correlations were in the expected direction.

Correlations between the Dot Counting Test and Reliable Digit Span ($r = -.20$) and Rey 15-Item test ($r = -.14$) were the highest among the criterion PVTs. Most of the criterion PVTs were significantly, but modestly correlated with the predictor EVIs. The Dot Counting Test was also most correlated with the predictor EVIs, though modestly so (r s ranged from $-.14$ to $-.36$). The criterion PVTs with forced choice memory paradigms, including the Rey Auditory Verbal Learning Test Effort Score and Brief Visuospatial Memory Test-Revised Recognition Discrimination, were least associated with the predictor EVIs. These findings are available upon request.

Group differences

As shown in Table 1, the invalid group performed significantly worse than the valid group on all criterion PVTs, with medium effects. The invalid group also performed significantly worse than the valid group across all the other cognitive tests, with small to medium effects. The invalid group had about one year less of education and fewer female participants than the valid group, resulting in statistically significant differences with small effect sizes. Because these differences were modest in their effect, follow-up analyses were not adjusted for demographics. There were no significant differences in symptoms of internalizing psychopathology between groups.

Correlations among embedded validity indicators

As shown in Table 3, the EVIs generally demonstrated small to medium correlations. Unsurprisingly, the strongest correlations were found among EVIs within similar cognitive measures. For instance, Symbol Search and Coding, which are both processing speed subtests from the WAIS-IV, showed a correlation of .52, and the three SCWT subtests showed large intercorrelations. Similar patterns were seen between the TMT subtests. All correlations were in the expected direction.

Embedded validity indicator cutoffs and classification accuracy statistics

Receiver operating characteristic curve analyses were statistically significant for all the EVIs, with AUCs ranging from .69–.78. Thus, the majority of the EVIs demonstrated acceptable AUC values (Hosmer et al., 2013). Optimal cut-scores for each EVI and their associated classification

Table 4. Accuracy and optimal cut-scores for each embedded validity indicator for detecting invalidity.

Embedded Validity Indicator (Cutoff Metric)	AUC (95% CI)	Cutoff	SN	SP	10% Base Rate		20% Base Rate		30% Base Rate	
					PPV	NPV	PPV	NPV	PPV	NPV
WAIS-IV Symbol Search (ACSS)	.76*** (.69, .83)	≤6	.51	.90	.36	.94	.56	.88	.69	.81
WAIS-IV Coding (ACSS)	.78*** (.72, .85)	≤6	.35	.93	.36	.93	.56	.85	.68	.77
WAIS-IV LNS (ACSS)	.78*** (.72, .85)	≤7	.49	.92	.40	.94	.60	.88	.72	.81
		≤6	.25	.97	.48	.92	.68	.84	.78	.75
SCWT Word Reading (T-score)	.75*** (.68, .82)	≤25	.43	.90	.32	.93	.52	.86	.65	.79
SCWT Color Naming (T-score)	.75*** (.69, .81)	≤28	.37	.91	.31	.93	.51	.85	.64	.77
SCWT Color-Word (T-score)	.73*** (.67, .80)	≤35	.39	.90	.30	.93	.49	.86	.63	.77
TMT-A (T-score)	.70*** (.63, .78)	≤33	.26	.91	.24	.92	.42	.83	.55	.74
TMT-B (T-score)	.77*** (.71, .83)	≤34	.38	.91	.32	.93	.51	.85	.64	.77
CPT-3 Omissions (T-score)	.69*** (.62, .76)	≥60	.19	.90	.17	.91	.32	.82	.45	.72
Lexical Fluency FAS (T-score)	.69*** (.61, .77)	≤34	.30	.90	.25	.92	.43	.84	.56	.75

Note: $N = 585$; AUC: Area under the curve; SN: Sensitivity; SP: Specificity; PPV: Positive predictive power; NPV: Negative predictive power; WAIS-IV: Wechsler Adult Intelligence Scale-4th Edition; LNS: Letter-Number Sequencing; SCWT: Stroop Color and Word Test; TMT: Trail Making Test; CPT-3: Conner's Continuous Performance Test-Third Edition; ACSS: Age-corrected scaled scores; Stroop T-scores were adjusted for age and educational attainment; Lexical Fluency FAS and Trail Making Test T-scores were adjusted for age, sex, educational attainment, and race; CPT-3 Omissions T-scores were adjusted for age and sex.

*** $p < .001$.

Table 5. Univariate logistic regression analyses of embedded validity indicators predicting validity group membership.

Embedded Validity Indicator	Model X^2	Nagelkerke's R^2	AIC	B	Standard Error (B)	Wald
WAIS-IV Symbol Search	55.76***	.200	370.23	2.80	.32	7.89
WAIS-IV Coding	34.25***	.146	393.75	2.00	.31	6.34
WAIS-IV LNS	31.93***	.103	396.07	2.30	.39	5.96
SCWT Word Reading	36.16***	.116	391.83	1.83	.29	6.33
SCWT Color Naming	29.68***	.095	399.31	1.49	.26	5.26
SCWT Color-Word	31.91***	.103	396.08	1.72	.29	5.96
TMT-A	15.59***	.061	400.39	1.23	.31	3.95
TMT-B	28.12***	.098	399.56	1.62	.29	5.52
Lexical Fluency FAS	19.49***	.077	401.59	1.55	.30	4.45
CPT-3 Omissions	3.08***	.017	515.91	0.68	.39	1.97

Note: $N = 585$; B : Unstandardized beta coefficient; AIC: Akaike information criterion; WAIS-IV: Wechsler Adult Intelligence Scale-4th Edition; LNS: Letter-Number Sequencing; SCWT: Stroop Color and Word Test; TMT: Trail Making Test; CPT-3: Conner's Continuous Performance Test-Third Edition.

*** $p < .001$.

accuracy statistics are included in Table 4. Sensitivities ranged from .19–.51 while maintaining $\geq .90$ specificity. Interestingly, the optimal cut-score for Letter-Number Sequencing subtest was an age-corrected scaled score of ≤ 7 (yielding a sensitivity of .49 and specificity of .92), which is considered “within normal limits” when referenced against the normative sample. Because this cut-score has the potential to deem someone's performance as “invalid before impaired” (Erdodi & Lichtenstein, 2017), we also included in the table below the sensitivity and specificity associated with an age-corrected scaled score of ≤ 6 , which falls below the normal limits range. Compared to the optimal cut-score of ≤ 7 , a cut-score of ≤ 6 resulted in lower sensitivity (.25) with slightly higher specificity (.97). Only the cut-score of ≤ 7 was retained for the follow-up analyses (see Discussion section for our reasoning). The Symbol Search age-corrected scaled score was the most sensitive EVI (.51) with a cut-score of ≤ 6 . The Symbol Search and Letter-Number Sequencing EVIs were the only indicators that hovered the “Larrabee limit” (i.e., $\geq .50$ sensitivity and $\geq .90$ specificity), indicating sufficient psychometric properties for clinical use.

Univariate predictive models

Table 5 displays the results of simple logistic regressions examining the predictive ability of each EVI using the cut-scores that were established in the above analyses. As

expected, these results showed that the EVI scores significantly predicted validity group membership. Again, the Symbol Search EVI performed better than all other EVIs when each was examined independently, explaining approximately 20% of the variance. Conversely, the CPT-3 Omissions EVI explained the least amount of variance ($< 2\%$) in validity status.

Multivariate predictive models

Combinations of the EVIs were then examined in two multivariate logistic regression models. Despite their wide range of intercorrelations, variance inflation factor and tolerance values among the EVIs indicated that multicollinearity did not significantly affect either of the multivariate models (Menard, 2000).

As seen in Table 6, the best subset selection method revealed that the most effective parsimonious model for classifying validity status comprised the WAIS-IV Symbol Search age-corrected scaled score (≤ 6), WAIS-IV Coding age-corrected scaled score (≤ 6), WAIS-IV Letter-Number Sequencing age-corrected scaled score (≤ 7), SCWT Word Reading T-score (≤ 25), TMT-B T-score (≤ 34), and Lexical Fluency FAS T-score (≤ 34). The best subset multivariate model demonstrated excellent overall classification accuracy (AUC = .86) and explained significantly more variance (total Nagelkerke's $R^2 = 36.00\%$) in validity status than any

Table 6. Multivariate logistic regression analyses of embedded validity indicators predicting validity group membership.

Embedded Validity Indicator	Model χ^2	Nagelkerke's R^2	AIC	B	Standard Error (B)	Wald	Odds Ratio
Kitchen Sink Multivariate Model							
	110.37***	.363	323.32				
WAIS-IV Symbol Search				1.16***	.17	3.56	5.39
WAIS-IV Coding				0.36*	.09	1.41	2.01
WAIS-IV LNS				1.04***	.10	3.34	4.47
SCWT Word Reading				0.03	.02	0.80	0.90
SCWT Color Naming				0.02	.02	0.61	0.99
SCWT Color-Word				0.10	.02	0.53	1.03
TMT-A				0.09	.02	0.30	1.01
TMT-B				0.20	.03	1.21	1.26
Lexical Fluency FAS				0.10	.02	0.83	1.02
CPT-3 Omissions				−0.01	.01	−0.08	0.31
Area Under the Curve = .86		Sensitivity = .70		Specificity = .90			
Best Subset Selection Multivariate Model							
	109.34***	.360	316.34				
WAIS-IV Symbol Search				1.12**	.15	3.39	5.32
WAIS-IV Coding				0.34*	.09	1.39	2.00
WAIS-IV LNS				1.51**	.11	4.15	5.63
SCWT Word Reading				0.37*	.03	1.66	1.90
TMT-B				0.39*	.04	1.47	1.93
Lexical Fluency FAS				0.44*	.02	1.09	2.02
Area Under the Curve = .86		Sensitivity = .69		Specificity = .91			

Note: $N = 585$; B : Unstandardized beta coefficient; AIC: Akaike information criterion; WAIS-IV: Wechsler Adult Intelligence Scale-4th Edition; LNS: Letter-Number Sequencing; SCWT: Stroop Color and Word Test; TMT: Trail Making Test; CPT-3: Conner's Continuous Performance Test-Third Edition.

* $p < .05$; ** $p < .01$; *** $p < .001$.

Table 7. Classification accuracy based on the number and proportion of failures within the multivariate models.

Number of EVI Failures (Proportion of Failures)	Base Rate Failures	Sensitivity	Specificity	False Positives	Overall Classification
Best Subset Selection Multivariate Model					
1 (17%)	21%	.80	.71	29%	72%
2 (33%)	5%	.60	.91	9%	87%
3 (50%)	5%	.48	.96	4%	90%
4 (67%)	1%	.19	.98	2%	88%
5 (83%)	1%	.11	.99	<1%	88%
6 (100%)	<1%	.01	1.00	0%	88%
Kitchen Sink Multivariate Model					
1 (10%)	23%	.89	.53	47%	57%
2 (20%)	14%	.72	.76	24%	75%
3 (30%)	5%	.60	.90	10%	86%
4 (40%)	3%	.51	.94	6%	88%
5 (50%)	3%	.37	.96	4%	88%
6 (60%)	2%	.24	.98	2%	89%
7 (70%)	2%	.15	.99	1%	89%
8 (80%)	<1%	.07	.99	<1%	88%
9 (90%)	<1%	.04	.99	<1%	88%
10 (100%)	<1%	.02	1.00	0%	88%

Note: $N = 585$; Base rate failures is the proportion of participants who scored below the number of cutoff(s) listed in the left-hand column; this is the sum of true and false positives divided by the sample size.

single EVI. The overall statistically derived sensitivity of the model increased to .69 while maintaining $\geq .90$ specificity. Furthermore, the Akaike information criteria statistics indicated that the best subset selection multivariate model was a better overall fit than the models for any single EVI.

The kitchen sink model comprising all 10 of the EVIs also explained significantly more variance than any independent EVI. The kitchen sink model explained nearly the same amount of variance (36.30%) as the more parsimonious, best subset selection model despite containing four more EVIs as predictors. Classification accuracy statistics ($AUC = .86$; sensitivity = .70; specificity = .90) were equivalent to the best subset selection model. However, the Akaike information criteria statistics indicated that the best subset selection model was a better fit than the kitchen sink model regarding the number of parameters. Interestingly, there were fewer EVIs within the kitchen sink model that

remained significantly predictive of classification accuracy relative to the best subset selection model.

Table 7 illustrates the classification accuracy statistics from the kitchen sink and best subset selection models based on their number and proportion of failures. Failing any two EVIs from the kitchen sink model resulted in a 24% false-positive rate, whereas failing any two EVIs from the best subset selection model resulted in a 9% false-positive rate. However, failing any three EVIs from the kitchen sink model resulted in a 10% false-positive rate. Thus, both of the models upheld an appropriate false-positive rate of 10% or less when requiring three failures from the kitchen sink model and two failures from the best subset selection model. Furthermore, these number of failures relative to the total number of EVIs within their respective model are proportionally similar (30-33%), and they yielded an optimal and similar balance between sensitivity (.60) and specificity

(.91-.90). Compared to the original models' receiver operating characteristic curve index of .86, the bootstrap analyses revealed a validated *C* index of .79 and .82 for the kitchen sink and best subset selection models, respectively.

Discussion

Using a targeted validity assessment approach in a neuropsychological evaluation may improve the identification of invalid performance (Erdodi, 2019). The current study investigated the individual and combined utility of 10 EVIs derived from executive functioning, attention/working memory, and processing speed measures in a large sample of adult patients undergoing an ADHD diagnostic evaluation.

All of the measures could independently discriminate invalid from valid performance, consistent with prior research (Ord et al., 2021; Robinson et al., 2023; Scimeca et al., 2021) and our hypothesis. The cut-scores derived from these measures and their associated classification accuracy statistics are also consistent with those from other performance validity studies in ADHD samples (e.g., Ausloos-Lozano et al., 2022; Khan et al., 2022; Scimeca et al., 2021). Specifically, the cut-scores for each measure yielded a range of low to moderate sensitivity values (.19-.51) when the specificity was maintained at $\geq .90$. None of these cut-scores explained more than 20% of the variance in validity status by themselves, which is consistent with most EVIs used in ADHD samples (Wallace et al., 2019). Research in other clinical populations has also shown that validity indicators embedded within measures of executive functioning, attention/working memory, and processing speed are relatively insensitive, yielding similar classification accuracy values as those in the current study (e.g., Whiteside et al., 2019). It is possible that using standardized rather than raw scores amplified the insensitivity of our indicators. This may have been especially problematic for the measures with scaled score cutoffs because they are less granular (i.e., providing a less detailed representation of an individual's performance relative to the mean and standard deviation of the normative sample) than raw scores. However, standardized scores were used in the current study because they have previously been shown to be optimal indicators of performance validity within these measures (e.g., Ausloos-Lozano et al., 2022; Bing-Canar et al., 2022; Khan et al., 2022; Scimeca et al., 2021).

Another potential issue with using standardized cut-scores is that they are based on normative data that may not have accounted for noncredible performance, thereby increasing the chance of identifying cut-scores within the "normal" range (Erdodi & Lichtenstein, 2017). Indeed, we found that the optimal cut-score for Letter-Number Sequencing (age-corrected scaled score of ≤ 7) was within normal limits. Using this age-corrected scaled score cutoff of ≤ 7 would presume that very few individuals with bona fide ADHD should perform below the 16th percentile on the Letter-Number Sequencing subtest, which is not empirically supported. This cut-score ostensibly raises the likelihood for high false positives, yet two prior studies also found that a

cut-score of ≤ 7 was ideal and yielded high specificity (Erdodi & Abeare, 2020; Shura et al., 2016). Furthermore, lowering the cut-score to a scaled score of ≤ 6 significantly decreases the sensitivity (-0.24) and marginally increases the specificity (+0.05). Given the substantial reduction in sensitivity and consistent findings with prior research, we used the scaled score cutoff of ≤ 7 in the multivariate model. This cut-score worked well in combination with other EVIs, but may be inappropriate to use in isolation, especially among individuals with lower cognitive functioning than those in the current study sample. Many of the other EVIs had cut-scores within the "mildly impaired" range according to descriptors recommended by the American Academy of Neuropsychology (Guilmette et al., 2020). Again, the majority of the current study sample demonstrated normal cognitive performance, but it may not be uncommon for persons with ADHD to demonstrate valid and mildly impaired performance (Willcutt et al., 2005). Thus, most of the 10 measures in this study may be somewhat useful validity indicators when used individually, but should be interpreted with caution given their potential issues with sensitivity and specificity.

The variability in sensitivity among these EVIs also raises the question of how important it is to assess performance validity in a targeted manner. It has been demonstrated that people will feign impairment on select measures of cognition (Boone, 2009). So, in this study, we included measures that are likely to elicit both feigned and genuine impairment during an ADHD evaluation. Of these measures, examinees may be most inclined to feign impairment on those that appear to assess attention and working memory (Quinn, 2003) given that these are widely publicized deficits associated with ADHD. Yet, our study found that the most sensitive EVI (Symbol Search) was derived from a processing speed measure, while the least sensitive EVI (CPT-3 Omissions) was derived from a measure of sustained attention. All patients in this study were aware that they were undergoing an ADHD evaluation and could have underperformed on select measures, but no clear pattern of feigned impairment emerged across the entire sample. Thus, there is no evidence from this study to suggest that practitioners can better detect performance invalidity by using EVIs that appear to measure attention versus processing speed or executive functioning. In other words, collectively employing various EVIs derived from executive functioning, attention/working memory, and processing speed measures may be a sufficient and targeted strategy to increase the signal detection of performance invalidity, but utilizing a more targeted approach (i.e., hand-picking attention versus processing speed EVIs) does not necessarily add value in an ADHD evaluation.

Indeed, we found that combining the EVIs significantly improved the classification accuracy and explained more variance in performance validity status than using any EVI in isolation, which is consistent with our hypothesis and prior research (Whiteside et al., 2019). When compared to the best performing EVI (Symbol Search), combining the EVIs increased the sensitivity by 18-19% while maintaining

adequate specificity ($\geq .90$). A 18-19% increase in sensitivity can make a meaningful difference in a clinical evaluation for ADHD where performance invalidity is common, yet difficult to detect (Musso & Gouvier, 2014). However, this relative increase in sensitivity also suggests that sensitivity and classification accuracy does not necessarily rise linearly when adding more validity indicators to an evaluation. The multivariate analyses further indicated that chaining multiple validity indicators is only helpful to an extent (Sweet et al., 2021).

The efficacy of integrating multiple PVTs relies upon using sensitive and nonredundant indicators (Bilder et al., 2014). We used a variable selection method to statistically derive the most effective subset of EVIs that minimized redundancy, maximized sensitivity, and maintained high specificity. This model resulted in a combination of six EVIs (WAIS-IV Symbol Search, Coding, and Letter-Number Sequencing, SCWT Word Reading, TMT-B, and Lexical Fluency) that performed as well as using all 10 of the EVIs together. Inconsistent with our hypothesis, this model showed that some of the measures that are known to assess similar cognitive abilities (e.g., WASI-IV Symbol Search and Coding) indexed nonredundant aspects of performance validity. The similarity between the best subset selection and kitchen sink model in terms of classification accuracy illustrates the tradeoff between parsimony and quality (best subset selection) versus quantity (kitchen sink).

Although the overall variance and classification accuracy values were similar across the kitchen sink and best subset selection models, the lengthier, kitchen sink model contained fewer significantly predictive EVIs. Paired with its poor fit relative to the best subset selection model, this indicates that information provided by some of the EVIs was likely redundant within the kitchen sink model. Bivariate correlations showed that the EVIs were modestly to moderately redundant, consistent with most prior research (Berthelson et al., 2013), which underscores the importance of being selective about which EVIs to include in the multivariate models. Although the multicollinearity assumption (based on variance inflation factor and tolerance values) was not violated in any of the multivariate models, the null washout effects observed in the kitchen sink model suggest that the predictive power of the model was limited by the redundant EVIs. If practitioners used all the EVIs from the kitchen sink model in an evaluation, scores falling above or below the cut-scores on the EVIs that were nonsignificant (i.e., the EVIs that captured redundant information in the model) have the potential to artificially inflate the overall number of above or below cut-scores and thereby may bias the interpretation of validity status. Three validity indicators within this model that may introduce the most bias in the interpretation of validity status are the SCWT EVIs. These EVIs demonstrated relatively high correlations and did not remain significantly predictive of validity status when used together in the multivariate model.

Furthermore, if practitioners relied on two failures from any combination of EVIs in the kitchen sink model, there would be an unacceptably high false-positive rate (24%).

Conversely, two failures were associated with an acceptable misclassification rate (9%) in the best subset selection model. Two PVT failures has historically been a benchmark criterion for discerning invalid from valid performance (Slick et al., 1999). However, some research suggests that three or more failures on PVTs may be more appropriate for determining validity status when using ten or more PVTs in a clinical evaluation (Larrabee et al., 2019; Sherman et al., 2020). Indeed, our findings revealed that failing any three EVIs from the kitchen sink model resulted in very similar sensitivity (.60) and specificity (.90-.91) as failing any two EVIs from the best subset selection model. This was further supported by the similarity in overall classification accuracy for the kitchen sink model at three failures (86%) compared to the best subset selection model at two failures (87%). Failing two EVIs from the best subset selection model is also a similar proportion of failures (33% versus 30%) as failing three EVIs from the kitchen sink model. However, the proportion of failures may not be as robust of an indicator of invalidity as the total number of failures given the skewed nature in the number of validity failures among our sample. So, even though the best subset selection model may be appealing in that it derived the “best” subset of the fewest number of EVIs, it did not actually yield better classification accuracy than the kitchen sink model, depending on the number of PVT failures used to determine invalidity.

Taken together, these findings suggest that three or more failures across 10 or more EVIs within measures of executive functioning, attention/working memory, and processing speed can help determine the validity of cognitive performance among adults undergoing an ADHD evaluation. Alternatively, if practitioners have access to a smaller number of PVTs (e.g., between four to nine measures), they may achieve similar classification accuracy when relying on two failures. These number and proportion of failures are consistent with research indicating that failure on two out of four to nine PVTs and three out of 10 or more PVTs should yield ideal classification accuracy (Meyers & Volbrecht, 2003; Sherman et al., 2020; Victor et al., 2009). If practitioners wish to increase the specificity of identifying invalid performance while maintaining acceptable sensitivity (i.e., hovering around $\geq .50$ sensitivity), they may utilize ≥ 4 failures for the kitchen sink model and ≥ 3 for the best subset selection model.

Limitations and Future directions

The primary limitation of this study concerns the interpretation of the best subset selection model. While it was suitable for this study, variable selection methods should generally be interpreted with caution due to their tendency to make sample-specific decisions. This best subset model was conducted in a single sample of individuals without genuinely moderate or severe cognitive impairment. Although genuinely moderate to severe cognitive impairment is not common in persons with ADHD, such individuals may still present with a concomitant neurocognitive disorder. Thus, it is difficult to know how well the exact

combination of EVIs in the best subset selection model will generalize to other study samples. We encourage researchers to investigate whether these findings replicate in a clinically diverse sample of individuals with ADHD who exhibit significant cognitive impairment. The *C* indices derived from the bootstrap resampling indicated that the best subset selection model was robust and the kitchen sink model was slightly overly optimistic. While this internal validation method may suggest that the model findings will replicate in other research samples, it is impractical and idealistic to assume that practitioners should perform a careful and systematic evaluation of each available PVT, as done in this study.

Both multivariate models demonstrated relatively good classification accuracy for the assessment of performance validity, but no combination of EVIs explained more than half of the variance in validity status. It is possible that some of the unexplained variance and limited sensitivity is due to using embedded rather than freestanding measures. Embedded indicators are generally more susceptible to the effects of bona fide cognitive and neuropsychiatric impairment that is common in ADHD (Harrison et al., 2023). Using standardized instead of raw scores may have exacerbated this issue for the reasons described above. Thus, future researchers may wish to examine if these EVIs show incremental value when used in combination with freestanding PVTs. It is important to examine a battery of both EVIs and freestanding PVTs because this is common practice in neuropsychological evaluation. There may also be value in combining these EVIs with symptom validity tests embedded within ADHD self-report questionnaires (e.g., Connor's Adult ADHD Rating Scales [Conners, et al., 1999; Suhr et al., 2011] or Clinical Assessment of Attention Deficit-Adult [Bracken & Boatwright, 2005]) to evaluate the validity of obtained test data during an ADHD evaluation (e.g., Marshall et al., 2010). In order to examine multivariate models with more performance and symptom validity tests, prospective researchers may need to use a larger subset of invalid performers.

It is also important to acknowledge that the criterion PVTs used in this study are imperfect predictors of validity status. While we chose independent criterion PVTs that have been cross-validated in ADHD populations and assessed (or appeared to assess) different cognitive skills at various times throughout the evaluation, their modest degree of intercorrelation suggests they may have captured some redundant aspects of performance validity. Future research may benefit from utilizing more advanced exploratory statistical methods to identify an ideal combination of non-overlapping criterion PVTs that can be used to determine validity status. The congruency between criterion and predictor PVTs should also be considered. There was limited variability in the correlation between the criterion PVTs and predictor EVIs used in this study, but the predictor EVIs were least associated with criterion PVTs using forced choice memory paradigms. Some research has also indicated that memory and non-memory PVTs index different aspects of performance validity (e.g., Webber et al., 2020). Further

exploration of the congruency between these measures may be helpful.

Various methods can be used in research to develop the construct of validity status, but it is important to compare the rate of failure based on the criterion construct to prior research, assuming prior research has yielded accurate base rates. In this study, the base rate of failure for each criterion PVT generally aligned with more contemporary research (Abramson et al., 2023; Hirsch et al., 2022; Marshall et al., 2016; Martin & Schroeder, 2020; Ovsiew et al., 2023; Phillips et al., 2023), except for the Rey 15-Item Test. Only 4% of the sample failed this PVT, while 10–14% of the sample failed the other criterion PVTs. The Rey 15-item Test has demonstrated high specificity across mixed clinical populations, providing a better ability to rule-in rather than rule-out invalid performance (Nitch & Glassmire, 2007). It is well-suited for medical populations with genuine and severe cognitive impairment (Nitch & Glassmire, 2007), which might explain why it was insensitive to performance invalidity in this sample of young, highly educated patients with ADHD. To address these concerns, it would be helpful to replicate these findings using different criterion PVTs and consider using PVTs that are robust to the effects of education and other factors thought to influence performance. We encourage researchers to expand upon recent studies investigating the congruency between criterion and predictor PVTs as well as the issue of classifying validity status for patients with one criterion PVT failure (e.g., Erdodi, 2023).

A strength of this study was examining EVIs within measures of executive functioning, attention/working memory, and processing speed which may be genuinely impaired in persons undergoing an ADHD evaluation (Quinn, 2003; Willcutt et al., 2005). However, individuals may have underperformed on other neurocognitive tests that were not examined in this study. Using EVIs derived from measures within other cognitive domains (e.g., memory) may explain more of the variance in validity status than the types of measures used in this study. It would be helpful for future researchers to compare these EVIs to other domain-specific indicators to discern the relative value in EVIs derived from tests of executive functioning, attention/working memory, and processing speed.

Lastly, data regarding whether patients were seeking academic accommodations or medication were not collected. These data can help discern why patients feign impairment in ADHD evaluations. Future studies should consider this information in the context of their findings as well as their sample's base rate of failure.

Conclusions

This is the first study to identify EVIs within measures of executive functioning, attention/working memory, and processing speed that can be used together to optimally identify invalid performance in adults undergoing an ADHD evaluation. Findings showed that chaining multiple EVIs together improved classification accuracy, but only to an extent. The most parsimonious model comprising six EVIs (cut-scores

derived from the WAIS-IV Symbol Search, Coding, and Letter-Number Sequencing, SCWT Word Reading, TMT-B, and Lexical Fluency FAS) performed as well as the kitchen sink model comprising all 10 of the EVIs (AUC = .86). However, we recommend relying on two or more EVI failures from the parsimonious model and three or more failures from the kitchen sink model to deem performance as invalid. The predictive power rate of these PVT models was also strong given the high base rate of performance invalidity in ADHD evaluations (Musso & Gouvier, 2014). Nonetheless, we still encourage practitioners to primarily rely upon freestanding validity tests until these findings are replicated with more clinically diverse ADHD populations. We hope that these findings will encourage researchers to identify other combinations of EVIs that target specific types of cognitive performance. Accumulating a multivariate battery of validity tests that target executive functioning, attention/working memory, and processing speed performance may allow clinicians to assess the validity of performance data more thoroughly throughout an ADHD evaluation.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The author(s) reported there is no funding associated with the work featured in this article.

ORCID

John-Christopher A. Finley  <http://orcid.org/0000-0002-1854-8360>
Jason R. Soble  <http://orcid.org/0000-0003-3348-8762>

References

- Abramson, D. A., White, D. J., Rhoads, T., Carter, D. A., Hansen, N. D., Resch, Z. J., Jennette, K. J., Ovsiew, G. P., & Soble, J. R. (2023). Cross-validating the dot counting test among an adult ADHD clinical sample and analyzing the effect of ADHD subtype and comorbid psychopathology. *Assessment*, 30(2), 264–273. <https://doi.org/10.1177/10731911211050895>
- American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders*. (5th ed.). <https://doi.org/10.1176/appi.books.9780890425596>
- Ashendorf, L., Clark, E. L., & Humphreys, C. T. (2021). The Rey 15-item memory test in US veterans. *Journal of Clinical and Experimental Neuropsychology*, 43(3), 324–331. <https://doi.org/10.1080/13803395.2021.1932761>
- Ausloos-Lozano, J. E., Bing-Canar, H., Khan, H., Singh, P. G., Wisinger, A. M., Rauch, A. A., Ogram Buckley, C. M., Petry, L. G., Jennette, K. J., Soble, J. R., & Resch, Z. J. (2022). Assessing performance validity during attention-deficit/hyperactivity disorder evaluations: Cross-validation of non-memory embedded validity indicators. *Developmental Neuropsychology*, 47(5), 247–257. <https://doi.org/10.1080/87565641.2022.2096889>
- Beck, A. T., Epstein, N., Brown, G., & Steer, R. (1993). Beck anxiety inventory. *Journal of Consulting and Clinical Psychology*, 61(6), 1096–1099.
- Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Manual for the Beck Depression Inventory-II*. The Psychological Corporation.
- Ben-Porath, Y. S., & Tellegen, A. (2008). *Minnesota Multiphasic Inventory-2-Restructured Form (MMPI-2-RF): Manual for administration, scoring, and interpretation*. University of Minnesota Press.
- Berthelson, L., Mulchan, S. S., Odland, A. P., Miller, L. J., & Mittenberg, W. (2013). False positive diagnosis of malingering due to the use of multiple effort tests. *Brain Injury*, 27(7–8), 909–916. <https://doi.org/10.3109/02699052.2013.793400>
- Bilder, R. M., Sugar, C. A., & Helleman, G. S. (2014). Cumulative false positive rates given multiple performance validity tests: Commentary on Davis and Millis (2014) and Larrabee (2014). *The Clinical Neuropsychologist*, 28(8), 1212–1223. <https://doi.org/10.1080/13854046.2014.969774>
- Bing-Canar, H., Phillips, M. S., Shields, A. N., Ogram Buckley, C. M., Chang, F., Khan, H., Skymba, H. V., Ovsiew, G. P., Resch, Z. J., Jennette, K. J., & Soble, J. R. (2022). Cross-validation of multiple WAIS-IV digit span embedded performance validity indices among a large sample of adult attention deficit/hyperactivity disorder clinical referrals. *Journal of Psychoeducational Assessment*, 40(5), 678–688. <https://doi.org/10.1177/07342829221081921>
- Boone, K. B. (2009). The need for continuous and comprehensive sampling of effort/response bias during neuropsychological examinations. *The Clinical Neuropsychologist*, 23(4), 729–741. <https://doi.org/10.1080/13854040802427803>
- Boone, K. B. (2013). *Clinical practice of forensic neuropsychology: An evidence-based approach*. The Guilford Press.
- Boone, K. B. (2021). *Assessment of feigned cognitive impairment*. (2nd ed.) Guilford Publications.
- Bracken, B. A., & Boatwright, B. S. (2005). *Examiner's manual: Clinical Assessment of Attention Deficit-Child and Adult*. Psychological Assessment Resources.
- Conners, C. K. (2014). *Conners continuous performance test 3rd edition (Conners CPT 3) & connors continuous auditory test of attention (Conners CATA): Technical manual*. MHS.
- Cook, C., Buelow, M. T., Lee, E., Howell, A., Morgan, B., Patel, K., Bryant, A. M., Menatti, A., & Suhr, J. (2018). Malingered attention deficit/hyperactivity disorder on the Conners' adult ADHD rating scales: Do reasons for malingering matter? *Journal of Psychoeducational Assessment*, 36(6), 552–561. <https://doi.org/10.1177/0734282917696934>
- Conners, C. K., Erhardt, D., Epstein, J. N., Parker, J. D. A., Sitarenios, G., & Sparrow, E. (1999). Self-ratings of ADHD symptoms in adults: I. Factor structure and normative data. *Journal of Attention Disorders*, 3(3), 141–151. <https://doi.org/10.1177/108705479900300303>
- Davis, J. J. (2018). Performance validity in older adults: Observed versus predicted false positive rates in relation to number of tests administered. *Journal of Clinical and Experimental Neuropsychology*, 40(10), 1013–1021. <https://doi.org/10.1080/13803395.2018.1472221>
- Erdodi, L. A. (2019). Aggregating validity indicators: The salience of domain specificity and the indeterminate range in multivariate models of performance validity assessment. *Applied Neuropsychology. Adult*, 26(2), 155–172. <https://doi.org/10.1080/23279095.2017.1384925>
- Erdodi, L. A. (2023). Multivariate models of performance validity: The Erdodi index captures the dual nature of non-credible responding (continuous and categorical). *Assessment*, 30(5), 1467–1485. <https://doi.org/10.1177/10731911221101910>
- Erdodi, L. A., & Abeare, C. A. (2020). Stronger together: The Wechsler adult intelligence scale—Fourth edition as a multivariate performance validity test in patients with traumatic brain injury. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists*, 35(2), 188–204. <https://doi.org/10.1093/arclin/acz032>
- Erdodi, L. A., & Lichtenstein, J. D. (2017). Invalid before impaired: An emerging paradox of embedded validity indicators. *The Clinical Neuropsychologist*, 31(6–7), 1029–1046. <https://doi.org/10.1080/13854046.2017.1323119>
- Fazio, R. L., Faris, A. N., & Yamout, K. Z. (2019). Use of the Rey 15-item test as a performance validity test in an elderly population. *Applied Neuropsychology. Adult*, 26(1), 28–35. <https://doi.org/10.1080/23279095.2017.1353994>

- Finley, J. C. A., Brook, M., Kern, D. M., Reilly, J. L., & Hanlon, R. E. (2023). Profile of embedded validity indicators in criminal defendants with verified valid neuropsychological test performance. *Archives of Clinical Neuropsychology : The Official Journal of the National Academy of Neuropsychologists*, 38(4), 513–524. <https://doi.org/10.1093/arclin/acac073>
- Fuermaier, A., Tucha, O., Koerts, J., Lange, K. W., Weisbrod, M., Aschenbrenner, S., & Tucha, L. (2017). Noncredible cognitive performance at clinical evaluation of adult ADHD: An embedded validity indicator in a visuospatial working memory test. *Psychological Assessment*, 29(12), 1466–1479. <https://doi.org/10.1037/pas0000534>
- Glassmire, D. M., Wood, M. E., Ta, M. T., Kinney, D. I., & Nitch, S. R. (2019). Examining false-positive rates of Wechsler adult intelligence scale (WAIS-IV) processing speed-based embedded validity indicators among individuals with schizophrenia spectrum disorders. *Psychological Assessment*, 31(1), 120–125. <https://doi.org/10.1037/pas0000650>
- Golden, C. J. (1978). *The Stroop Color and Word Test: A manual for clinical and experimental uses*. Stoelting.
- Guilmette, T. J., Sweet, J. J., Hebben, N., Koltai, D., Mahone, E. M., Spiegler, B. J., Stucky, K., & Westerveld, M. Conference Participants. (2020). American Academy of clinical neuropsychology consensus conference statement on uniform labeling of performance test scores. *The Clinical Neuropsychologist*, 34(3), 437–453. <https://doi.org/10.1080/13854046.2020.1722244>
- Harrison, A. G., Beal, A. L., & Armstrong, I. T. (2023). Predictive value of performance validity testing and symptom validity testing in psychoeducational assessment. *Applied Neuropsychology. Adult*, 30(3), 315–329. <https://doi.org/10.1080/23279095.2021.1943396>
- Heaton, R. K., Miller, S. W., Taylor, M. J., & Grant, I. (2004). *Revised comprehensive norms for an expanded Halstead-Reitan battery: Demographically adjusted neuropsychological norms for African American and Caucasian adults*. Psychological Assessment Resources.
- Hirsch, O., Fuermaier, A. B., Tucha, O., Albrecht, B., Chavanon, M. L., & Christiansen, H. (2022). Symptom and performance validity in samples of adults at clinical evaluation of ADHD: A replication study using machine learning algorithms. *Journal of Clinical and Experimental Neuropsychology*, 44(3), 171–184. <https://doi.org/10.1080/13803395.2022.2105821>
- Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. (3rd ed.) John Wiley & Sons.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer.
- Jennette, K. J., Williams, C. P., Resch, Z. J., Ovsiew, G. P., Durkin, N. M., O'Rourke, J. J. F., Marceaux, J. C., Critchfield, E. A., & Soble, J. R. (2022). Assessment of differential neurocognitive performance based on the number of performance validity tests failures: A cross-validation study across multiple mixed clinical samples. *The Clinical Neuropsychologist*, 36(7), 1915–1932. <https://doi.org/10.1080/13854046.2021.1900398>
- Ka Yin Tse, P., Finley, J.-C. A., Frick, L., Guilfoyle, J., Brooks, J., Khalid, E., Charara, R., Resch, Z., Ulrich, D., Ovsiew, G. P., & Soble, J. (2023). Cross-validating the embedded performance validity indicators in the Rey auditory verbal learning test in mixed neuropsychiatric and attention deficit/hyperactivity disorder clinical samples. *Psychology & Neuroscience*, 16(2), 125–137. <https://doi.org/10.1037/pne0000302>
- Kassambara, A. (2018). *Machine learning essentials: Practical guide in R. Statistical Tools for High-Throughput Data Analysis*.
- Khan, H., Rauch, A. A., Obolsky, M. A., Skymba, H., Barwegen, K. C., Wisinger, A. M., Ovsiew, G. P., Jennette, K. J., Soble, J. R., & Resch, Z. J. (2022). A comparison of embedded validity indicators from the Stroop color and word test among adults referred for clinical evaluation of suspected or confirmed attention-deficit/hyperactivity disorder. *Psychological Assessment*, 34(7), 697–703. <https://doi.org/10.1037/pas0001137>
- Kosky, K. M., Lace, J. W., Austin, T. A., Seitz, D. J., & Clark, B. (2022). The utility of the Wisconsin card sorting test, 64-card version to detect noncredible attention-deficit/hyperactivity disorder. *Applied Neuropsychology. Adult*, 29(5), 1231–1241. <https://doi.org/10.1080/23279095.2020.1864633>
- Lakhan, S. E., & Kirchgessner, A. (2012). Prescription stimulants in individuals with and without attention deficit hyperactivity disorder: Misuse, cognitive impact, and adverse effects. *Brain and Behavior*, 2(5), 661–677. <https://doi.org/10.1002/brb3.78>
- Larrabee, G. J. (2003). Detection of malingering using atypical performance patterns on standard neuropsychological tests. *The Clinical Neuropsychologist*, 17(3), 410–425. <https://doi.org/10.1076/clin.17.3.410.18089>
- Larrabee, G. J. (2008). Aggregation across multiple indicators improves the detection of malingering: Relationship to likelihood ratios. *The Clinical Neuropsychologist*, 22(4), 666–679. <https://doi.org/10.1080/13854040701494987>
- Larrabee, G. J. (2014). False-positive rates associated with the use of multiple performance and symptom validity tests. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists*, 29(4), 364–373. <https://doi.org/10.1093/arclin/acu019>
- Larrabee, G. J., Rohling, M. L., & Meyers, J. E. (2019). Use of multiple performance and symptom validity measures: Determining the optimal per test cutoff for determination of invalidity, analysis of skew, and inter-test correlations in valid and invalid performance groups. *The Clinical Neuropsychologist*, 33(8), 1354–1372. <https://doi.org/10.1080/13854046.2019.1614227>
- Lee Booksh, R., Pella, R. D., Singh, A. N., & Drew Gouvier, W. (2010). Ability of college students to simulate ADHD on objective measures of attention. *Journal of Attention Disorders*, 13(4), 325–338. <https://doi.org/10.1177/1087054708329927>
- Lumley, T., & Lumley, M. T. (2013). Package 'leaps'. *Regression subset selection. Thomas Lumley Based on Fortran Code by Alan Miller*. Available online: <http://CRAN.R-project.org/package=leaps>.
- Lovett, B. J., & Harrison, A. G. (2021). Assessing adult ADHD: New research and perspectives. *Journal of Clinical and Experimental Neuropsychology*, 43(4), 333–339. <https://doi.org/10.1080/13803395.2021.1950640>
- Marshall, P., Schroeder, R., O'Brien, J., Fischer, R., Ries, A., Blesi, B., & Barker, J. (2010). Effectiveness of symptom validity measures in identifying cognitive and behavioral symptom exaggeration in adult attention deficit hyperactivity disorder. *The Clinical Neuropsychologist*, 24(7), 1204–1237. <https://doi.org/10.1080/13854046.2010.514290>
- Marshall, P. S., Hoelzle, J. B., Heyerdahl, D., & Nelson, N. W. (2016). The impact of failing to identify suspect effort in patients undergoing adult attention-deficit/hyperactivity disorder (ADHD) assessment. *Psychological Assessment*, 28(10), 1290–1302. <https://doi.org/10.1037/pas0000247>
- Martin, P. K., & Schroeder, R. W. (2020). Base rates of invalid test performance across clinical non-forensic contexts and settings. *Archives of Clinical Neuropsychology : The Official Journal of the National Academy of Neuropsychologists*, 35(6), 717–725. <https://doi.org/10.1093/arclin/aaa017>
- Menard, S. (2000). Coefficients of determination for multiple logistic regression analysis. *The American Statistician*, 54(1), 17–24. <https://doi.org/10.1080/00031305.2000.10474502>
- Meyers, J. E., & Volbrecht, M. E. (2003). A validation of multiple malingering detection methods in a large clinical sample. *Archives of Clinical Neuropsychology*, 18(3), 261–276. <https://doi.org/10.1093/arclin/18.3.261>
- Musso, M. W., & Gouvier, W. D. (2014). "Why is this so hard?" A review of detection of malingered ADHD in college students. *Journal of Attention Disorders*, 18(3), 186–201. <https://doi.org/10.1177/1087054712441970>
- Nitch, S. R., & Glassmire, D. M. (2007). Non-forced choice measures to detect noncredible cognitive performance. In K. B. Boone (Ed.), *Assessment of feigned cognitive impairment: A neuropsychological perspective*. (pp. 29–49). Guilford Publications.
- Ord, A. S., Miskey, H. M., Lad, S., Richter, B., Nagy, K., & Shura, R. D. (2021). Examining embedded validity indicators in Connors continuous performance test-3 (CPT-3). *The Clinical Neuropsychologist*, 35(8), 1426–1441. <https://doi.org/10.1080/13854046.2020.1751301>

- Ovsiew, G. P., Cerny, B. M., Boer, A. B. D., Petry, L. G., Resch, Z. J., Durkin, N. M., & Soble, J. R. (2023). Performance and symptom validity assessment in attention deficit/hyperactivity disorder: Base rates of invalidity, concordance, and relative impact on cognitive performance. *The Clinical Neuropsychologist*. <https://doi.org/10.1080/13854046.2022.2162440>
- Phillips, M. S., Wisinger, A. M., Lapitan-Moore, F. T., Ausloos-Lozano, J. E., Bing-Canar, H., Durkin, N. M., Ovsiew, G. P., Resch, Z. J., Jennette, K. J., & Soble, J. R. (2023). Cross-validation of multiple embedded performance validity indices in the Rey auditory verbal learning test and brief visuospatial memory test-revised in an adult attention deficit/hyperactivity disorder clinical sample. *Psychological Injury and Law*, 16(1), 27–35. <https://doi.org/10.1007/s12207-022-09443-3>
- Quinn, C. A. (2003). Detection of malingering in assessment of adult ADHD. *Archives of Clinical Neuropsychology*, 18(4), 379–395. <https://doi.org/10.1093/arclin/18.4.379>
- Rabiner, D. L. (2013). Stimulant prescription cautions: addressing misuse, diversion and malingering. *Current Psychiatry Reports*, 15(7), 375. <https://doi.org/10.1007/s11920-013-0375-2>
- Rhoads, T., Neale, A. C., Resch, Z. J., Cohen, C. D., Keezer, R. D., Cerny, B. M., Jennette, K. J., Ovsiew, G. P., & Soble, J. R. (2021). Psychometric implications of failure on one performance validity test: A cross-validation study to inform criterion group definition. *Journal of Clinical and Experimental Neuropsychology*, 43(5), 437–448. <https://doi.org/10.1080/13803395.2021.1945540>
- Robinson, A., Reed, C., Davis, K., Divers, R., Miller, L., Erdodi, L. A., & Calamia, M. (2023). Settling the score: Can CPT-3 embedded validity indicators distinguish between credible and Non-credible responders referred for ADHD and/or SLD? *Journal of Attention Disorders*, 27(1), 80–88. <https://doi.org/10.1177/10870547221121781>
- Schutte, C., Millis, S., Axelrod, B., & VanDyke, S. (2011). Derivation of a composite measure of embedded symptom validity indices. *The Clinical Neuropsychologist*, 25(3), 454–462. <https://doi.org/10.1080/13854046.2010.550635>
- Scimeca, L. M., Holbrook, L., Rhoads, T., Cerny, B. M., Jennette, K. J., Resch, Z. J., Obolsky, M. A., Ovsiew, G. P., & Soble, J. R. (2021). Examining Conners continuous performance test-3 (CPT-3) embedded performance validity indicators in an adult clinical sample referred for ADHD evaluation. *Developmental Neuropsychology*, 46(5), 347–359. <https://doi.org/10.1080/87565641.2021.1951270>
- Sherman, E. M., Slick, D. J., & Iverson, G. L. (2020). Multidimensional malingering criteria for neuropsychological assessment: A 20-year update of the malingered neuropsychological dysfunction criteria. *Archives of Clinical Neuropsychology : The Official Journal of the National Academy of Neuropsychologists*, 35(6), 735–764. <https://doi.org/10.1093/arclin/aca019>
- Shura, R. D., Miskey, H. M., Rowland, J. A., Yoash-Gantz, R. E., & Denning, J. H. (2016). Embedded performance validity measures with postdeployment veterans: Cross-validation and efficiency with multiple measures. *Applied Neuropsychology. Adult*, 23(2), 94–104. <https://doi.org/10.1080/23279095.2015.1014556>
- Silk-Eglit, G. M., Stencik, J. H., Miele, A. S., Lynch, J. K., & McCaffrey, R. J. (2015). Rates of false-positive classification resulting from the analysis of additional embedded performance validity measures. *Applied Neuropsychology. Adult*, 22(5), 335–347. <https://doi.org/10.1080/23279095.2014.938809>
- Skymba, H. V., Shields, A. N., Rauch, A. A., Phillips, M. S., Bing-Canar, H., Finley, A., Khan, H., Ovsiew, G. P., Durkin, N. M., Jennette, K. J., Resch, Z. J., & Soble, J. R. (2023). Does comorbid depression impact executive functioning (EF) in adults diagnosed with ADHD?: a comparison of EF across diagnoses in clinically-referred individuals. *Journal of Clinical and Experimental Neuropsychology*, 45(1), 1–11. <https://doi.org/10.1080/13803395.2023.2203464>
- Slick, D. J., Sherman, E. M., & Iverson, G. L. (1999). Diagnostic criteria for malingered neurocognitive dysfunction: Proposed standards for clinical practice and research. *The Clinical Neuropsychologist*, 13(4), 545–561. [https://doi.org/10.1076/1385-4046\(199911\)13:04;1-Y;FT545](https://doi.org/10.1076/1385-4046(199911)13:04;1-Y;FT545)
- Snyder, P., & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. *The Journal of Experimental Education*, 61(4), 334–349. <https://doi.org/10.1080/00220973.1993.10806594>
- Soble, J. R., Alverson, W. A., Phillips, J. I., Critchfield, E. A., Fullen, C., O'Rourke, J. J. F., Messerly, J., Highsmith, J. M., Bailey, K. C., Webber, T. A., & Marceaux, J. C. (2020). Strength in numbers or quality over quantity? Examining the importance of criterion measure selection to define validity groups in performance validity test (PVT) research. *Psychological Injury and Law*, 13(1), 44–56. <https://doi.org/10.1007/s12207-019-09370-w>
- Suhr, J. A., & Berry, D. T. (2017). The importance of assessing for validity of symptom report and performance in attention deficit/hyperactivity disorder (ADHD): Introduction to the special section on noncredible presentation in ADHD. *Psychological Assessment*, 29(12), 1427–1428. <https://doi.org/10.1037/pas0000535>
- Suhr, J. A., Buelow, M., & Riddle, T. (2011). Development of an infrequency index for the CAARS. *Journal of Psychoeducational Assessment*, 29(2), 160–170. <https://doi.org/10.1177/0734282910380190>
- Sweet, J. J., Heilbronner, R. L., Morgan, J. E., Larrabee, G. J., Rohling, M. L., Boone, K. B., Kirkwood, M. W., Schroeder, R. W., ... Suhr, J. A. (2021). American Academy of Clinical Neuropsychology (AACN) 2021 consensus statement on validity assessment: Update of the 2009 AACN consensus conference statement on neuropsychological assessment of effort, response bias, and malingering. *The Clinical Neuropsychologist*, 35(6), 1053–1106. <https://doi.org/10.1080/13854046.2021.1896036>
- Theiling, J., & Petermann, F. (2016). Neuropsychological profiles on the WAIS-IV of adults with ADHD. *Journal of Attention Disorders*, 20(11), 913–924. <https://doi.org/10.1177/1087054713518241>
- Thompson, B. (1995). Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial. *Educational and Psychological Measurement*, 55(4), 525–534. <https://doi.org/10.1177/001316449505500400>
- Victor, T. L., Boone, K. B., Serpa, J. G., Buehler, J., & Ziegler, E. A. (2009). Interpreting the meaning of multiple symptom validity test failure. *The Clinical Neuropsychologist*, 23(2), 297–313. <https://doi.org/10.1080/13854040802232682>
- Wallace, E. R., Garcia-Willingham, N. E., Walls, B. D., Bosch, C. M., Balthrop, K. C., & Berry, D. T. (2019). A meta-analysis of malingering detection measures for attention-deficit/hyperactivity disorder. *Psychological Assessment*, 31(2), 265–270. <https://doi.org/10.1037/pas0000659>
- Webber, T. A., Critchfield, E. A., & Soble, J. R. (2020). Convergent, discriminant, and concurrent validity of nonmemory-based performance validity tests. *Assessment*, 27(7), 1399–1415. <https://doi.org/10.1177/1073191118804874>
- Wechsler, D. (2008). *WAIS-IV: Technical and interpretive manual*. Pearson.
- White, D. J., Korinek, D., Bernstein, M. T., Ovsiew, G. P., Resch, Z. J., & Soble, J. R. (2020). Cross-validation of non-memory-based embedded performance validity tests for detecting invalid performance among patients with and without neurocognitive impairment. *Journal of Clinical and Experimental Neuropsychology*, 42(5), 459–472. <https://doi.org/10.1080/13803395.2020.1758634>
- White, D. J., Ovsiew, G. P., Rhoads, T., Resch, Z. J., Lee, M., Oh, A. J., & Soble, J. R. (2022). The divergent roles of symptom and performance validity in the assessment of ADHD. *Journal of Attention Disorders*, 26(1), 101–108. <https://doi.org/10.1177/1087054720964575>
- Whiteside, D. M., Caraher, K., Hahn-Ketter, A., Gaasedelen, O., & Basso, M. R. (2019). Classification accuracy of individual and combined executive functioning embedded performance validity measures in mild traumatic brain injury. *Applied Neuropsychology. Adult*, 26(5), 472–481. <https://doi.org/10.1080/23279095.2018.1443935>
- Whiteside, D. M., Kogan, J., Wardin, L., Phillips, D., Franzwa, M. G., Rice, L., Basso, M., & Roper, B. (2015). Language based embedded performance validity measures in traumatic brain injury. *Journal of Clinical and Experimental Neuropsychology*, 37(2), 220–227. <https://doi.org/10.1080/13803395.2014.1002758>
- Whiteside, D. M., Wald, D., & Busse, M. (2011). Classification accuracy of multiple visual spatial measures in the detection of suspect effort. *The Clinical Neuropsychologist*, 25(2), 287–301. <https://doi.org/10.1080/13854046.2010.538436>

- Willcutt, E. G., Doyle, A. E., Nigg, J. T., Faraone, S. V., & Pennington, B. F. (2005). Validity of the executive function theory of attention-deficit/hyperactivity disorder: A meta-analytic review. *Biological Psychiatry*, 57(11), 1336–1346. <https://doi.org/10.1016/j.biopsych.2005.02.006>
- Williamson, K. D., Combs, H. L., Berry, D. T., Harp, J. P., Mason, L. H., & Edmundson, M. (2014). Discriminating among ADHD alone, ADHD with a comorbid psychological disorder, and feigned ADHD in a college sample. *The Clinical Neuropsychologist*, 28(7), 1182–1196. <https://doi.org/10.1080/13854046.2014.956674>
- Woods, S. P., Lovejoy, D. W., & Ball, J. D. (2002). Neuropsychological characteristics of adults with ADHD: A comprehensive review of initial studies. *The Clinical Neuropsychologist*, 16(1), 12–34. <https://doi.org/10.1076/clin.16.1.12.8336>