Journal of the International Neuropsychological Society (2023), **29**, 789–797
doi:10.1017/S1355617722000893

Research Article

Development and application of novel performance validity metrics for computerized neurocognitive batteries

J. Cobb Scott[1,2], Tyler M. Moore[1], David R. Roalf[1], Theodore D. Satterthwaite[1], Daniel H. Wolf[1], Allison M. Port[1], Ellyn R. Butler[1], Kosha Ruparel[1], Caroline M. Nievergelt[4,5], Victoria B. Risbrough[4,5], Dewleen G. Baker[4,5], Raquel E. Gur[1,3] and Ruben C. Gur[1,2,3]

[1]Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA, [2]VISN4 Mental Illness Research, Education, and Clinical Center at the Corporal Michael J. Crescenz VA Medical Center, Philadelphia, PA, USA, [3]Lifespan Brain Institute, Department of Child and Adolescent Psychiatry and Behavioral Sciences, Children's Hospital of Philadelphia, Philadelphia, PA, USA, [4]Center for Excellent in Stress and Mental Health, VA San Diego Healthcare System, San Diego, CA, USA and [5]Department of Psychiatry, University of California (UCSD), San Diego, CA, USA

Abstract

Objectives: Data from neurocognitive assessments may not be accurate in the context of factors impacting validity, such as disengagement, unmotivated responding, or intentional underperformance. Performance validity tests (PVTs) were developed to address these phenomena and assess underperformance on neurocognitive tests. However, PVTs can be burdensome, rely on cutoff scores that reduce information, do not examine potential variations in task engagement across a battery, and are typically not well-suited to acquisition of large cognitive datasets. Here we describe the development of novel performance validity measures that could address some of these limitations by leveraging psychometric concepts using data embedded within the Penn Computerized Neurocognitive Battery (PennCNB). **Methods:** We first developed these validity measures using simulations of invalid response patterns with parameters drawn from real data. Next, we examined their application in two large, independent samples: 1) children and adolescents from the Philadelphia Neurodevelopmental Cohort ($n = 9498$); and 2) adult servicemembers from the Marine Resiliency Study-II ($n = 1444$). **Results:** Our performance validity metrics detected patterns of invalid responding in simulated data, even at subtle levels. Furthermore, a combination of these metrics significantly predicted previously established validity rules for these tests in both developmental and adult datasets. Moreover, most clinical diagnostic groups did not show reduced validity estimates. **Conclusions:** These results provide proof-of-concept evidence for multivariate, data-driven performance validity metrics. These metrics offer a novel method for determining the performance validity for individual neurocognitive tests that is scalable, applicable across different tests, less burdensome, and dimensional. However, more research is needed into their application.

Keywords: validity; psychometrics; item response theory; neuropsychological testing; malingering; learning and memory tests

(Received 30 March 2022; final revision 12 September 2022; accepted 6 October 2022; First Published online 12 December 2022)

Neuropsychological testing assesses brain-behavior functioning, typically to examine whether an injury or illness has negatively affected specific cognitive functions. However, several factors – including fatigue, disengagement, distraction, unmotivated responding, and intentional underperformance – can interfere with the reliable and valid measurement of cognitive functioning. Gauging whether such factors have affected the measurement of cognitive functioning is a significant challenge. Over the past three decades, research to formally discern valid from invalid test performance has rapidly increased, leading to assessment tools termed performance validity tests (PVTs; Larrabee, 2012). PVTs have provided substantial benefit to neuropsychological evaluations by providing evidence for or against credible cognitive performance.

Despite their clear utility, most PVT measures have limitations. Standalone PVTs typically require 10–15 min to administer, and many are still administered by paper and pencil. Such measures may not be suitable for "big data" acquisition, even though data integrity is a substantial concern in data collected at scale. Though several embedded PVTs exist, as described in Bilder & Reise (2019), neurocognitive assessments would benefit from having measures of performance validity built into every test, which may be facilitated by examining aberrant patterns of responses or speed. PVTs also rely on cutoff scores, which prevent exploration of potential gradients in performance validity (Walters et al., 2009) and impede flexibility for individuals utilizing the tests. For example, a clinician or researcher may value minimizing either false positives or false negatives, depending on population base rates of invalid performance or the evaluation context (McCormick et al., 2013). Cutoff measures also preclude examination of potential variations in task engagement *across* a

Corresponding author: J. Cobb Scott, email: scott1@pennmedicine.upenn.edu

J. Cobb Scott and Tyler M. Moore contributed equally to this work.

Cite this article: Scott J.C., Moore T.M., Roalf D.R., Satterthwaite T.D., Wolf D.H., Port A.M., Butler E.R., Ruparel K., Nievergelt C.M., Risbrough V.B., Baker D.G., Gur R.E., & Gur R.C. (2023) Development and application of novel performance validity metrics for computerized neurocognitive batteries. *Journal of the International Neuropsychological Society*, **29**: 789–797, https://doi.org/10.1017/S1355617722000893

Copyright © INS. Published by Cambridge University Press, 2022. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution, and reproduction in any medium, provided the original work is properly cited.

test battery. Relatedly, PVTs are typically used to predict insufficient engagement on separate, non-PVT measures, occasionally leading to challenging discrepancies (Loring & Goldstein, 2019) or false positives (Lippa, 2018). In addition, PVTs usually do not factor in speed of responding, though using speed in addition to accuracy may increase the sensitivity of PVTs (Kanser et al., 2019; Lupu et al., 2018). Innovative methods for performance validity that leverage item-level psychometrics (e.g., response consistency) may be beneficial for acquisition of large samples over the internet or for tests administered through telehealth, which has become especially relevant during the COVID-19 pandemic.

Examining item-level responses has the potential to increase sensitivity to insufficient effort and task disengagement. Person-fit metrics are one valuable tool to detect problematic responding at the item level. Person-fit (Reise, 1990; Tatsuoka & Linn, 1983; Tellegen, 1988) is a general term for psychometric methods designed to assess the consistency of a response pattern using some prior information (e.g., relative item difficulty). These methods examine whether an individual's pattern of responding on a test fits with the individual's overall ability level or the overall patterns of performance across all individuals. As a simple example, consider a 10-item test on which the first 5 items are of low difficulty and the last 5 items are of high difficulty. If an examinee answers the first 5 items correctly and last 5 items incorrectly, that pattern would fit the normative data well (high person-fit score) and suggest an ability level in between the difficulties of the easy items and the difficulties of the hard items. However, if an examinee answered the first 5 items incorrectly and last 5 items correctly, that pattern would not fit the normative data well – correct responses on hard items suggest that the examinee is of very high ability, whereas incorrect responses on easy items suggest that the examinee is of very low ability. This examinee would receive a low person-fit score (indicating person misfit). We propose that this lack of fit may be one useful indicator of invalid neurocognitive performance (e.g., reflecting haphazard responses or insufficient effort).

Here, we describe and apply novel methods for developing data-driven, embedded performance validity measures from item-level responses in the Penn Computerized Neurocognitive Battery (PennCNB; Gur et al., 2001, 2012, 2010), a neurocognitive battery that has been applied worldwide in numerous studies. First, we describe and present data from simulations that model invalid responding using a multivariate combination of metrics. Second, we examine associations of these simulation-derived metrics with established validity rules in "real-world" data from two large, independent samples. Finally, we examine associations of these performance validity metrics with psychopathology in these samples to examine whether individuals with elevated psychopathology are likely to be identified as false positives on these metrics.

Method

Neurocognitive battery

The PennCNB is a publicly available collection of tests assessing a range of neurocognitive functions (Gur et al., 2010; Moore et al., 2015). The 14 core PennCNB tests take approximately one hour to complete (see eTable 1). Numerous studies support the PennCNB's validity and reliability, including sensitivity to individual differences (Gur et al., 2012) and expected neurobehavioral profiles in clinical conditions (e.g., Aliyu et al., 2006; Hartung et al., 2016; Roalf et al., 2013). The battery has been applied in large-scale genomic studies (Greenwood et al., 2019; Gulsuner et al., 2020; Scott et al., 2021), intervention studies (Bhatia et al., 2017; Scott et al., 2021), and even in space flight (Basner et al., 2015; Garrett-Bakelman et al., 2019).

To create performance validity measures, we used six PennCNB tests conducive for psychometric characterization of item-level responses. Specifically, for these methods, tests need to be self-paced (no rapid signal-detection tests), include core measures of accuracy, and comprise fixed sets of administered items. Based on these criteria, we selected the following tests. Episodic Memory: The Penn Word Memory Test (CPW), Penn Face Memory Test (CPF) (Thomas et al., 2013) and Visual Object Learning Test (VOLT) (Glahn et al., 1997) have similar formats for presentation and recall. The CPW/CPF presents 20 words or faces to remember, respectively, and the recall portion shows these targets mixed with 20 non-targets. The VOLT shows participants a series of 10 three-dimensional Euclidean shapes, and the recall portion shows these targets mixed with 10 new shapes. Verbal Reasoning: The Penn Verbal Reasoning Test (PVRT) measures language-mediated reasoning ability using a series of analogy problems patterned after the Educational Testing Service factor-referenced test kit. Nonverbal Reasoning: The Penn Matrix Analysis Test (PMAT) measures nonverbal reasoning ability using problems similar to those used in the Raven's Progressive Matrices Test (Raven, 1989) and WAIS-IV Matrix Reasoning (Wechsler, 2008). Emotion Identification: The Penn Emotion Recognition Test (ER40) measures the social cognition domain of emotion identification. Participants are shown 40 individual faces and must determine whether the emotion expressed by the actor's face is happiness, sadness, anger, fear, or none at all. There are 4 female and 4 male faces for each emotion ($4 \times 2 \times 5 = 40$).

Participants

To generate *parameters* for the simulations described below, we used item-level accuracy and response time data from the Philadelphia Neurodevelopmental Cohort (PNC), a large ($n = 9498$), community-based sample of youth between ages 8 and 22. Please see Calkins et al., (2015) and Supplementary Methods for more extensive information on PNC recruitment, enrollment, and procedures. The Institutional Review Boards at the University of Pennsylvania (Penn) and CHOP approved this study.

To examine the utility of validity metrics derived from these simulations, we applied these metrics to PNC and Marine Resiliency Study-II (MRS-II) data for external validation. MRS-II participants ($n = 1444$) have been described in detail previously (Acheson et al., 2015), with more information available in Supplementary Methods. The institutional review boards of the University of California San Diego, Penn, VA San Diego Research Service, and Naval Health Research Center approved this study.

Procedures

The PennCNB was administered by assessors who undergo extensive training by experienced assessors, including hands-on instruction, observation of mock sessions, feedback after practice assessments, and standardized certification procedures. For both samples, participants were administered the PennCNB on laptops by proctors trained by experienced Penn staff. However, the PNC used a one-to-one proctor-examinee setup, and the MRS-II administered tests with multiple service members and one proctor.

After test administration, assessors assigned a code to rate the quality of data and took detailed notes to indicate when an examinee exhibited disengagement from a task. In addition, algorithmic

https://doi.org/10.1017/S1355617722000893 Published online by Cambridge University Press

validity rules for PennCNB tests have been developed specific to each test to detect certain subject-related problems (e.g., extended periods of inattention, random or habitual responding). These rules include flagging of repeated instances of impossible response times (e.g., < 200 msec), excessive outliers in performance, and unusual response patterns (e.g., choosing the same option several times in a row). See Supplemental Methods for greater detail on these algorithmic validity rules. Experienced CNB data validators supervised by neuropsychologists make ultimate decisions to designate a test as invalid by integrating algorithmic rules, assessor comments, and visual data inspection. Though helpful for quality assurance, these validation procedures require significant manual intervention, are infeasible for collecting data at scale, and are not "gold standard" criteria, as they are less likely to detect more subtle task disengagement or insufficient effort than typical PVTs. Nonetheless, associations between our performance validity metrics and these validation rules provide a useful initial test of criterion validity.

Performance validity metrics

We used a combination of three methods to generate performance validity metrics.

Person-fit metric

Person-fit is a measure of how feasible an individual's response pattern is given item characteristics (e.g., difficulty). Many person-fit indicators have been described – see Karabatsos (2003) and Meijer & Sijtsma (2001). We selected two indicators that are least correlated with each other: the point-biserial[1] method (Donlon & Fischer, 1968) and Kane & Brennan (1980) dependability method. Both measures indicate the degree to which a person's response pattern conforms to sample-level response patterns.[2] See Supplemental Methods for more details on these specific methods.

Response time[3] outlier metric

This metric reflects the proportion of items on which the response time (RT) is at least two standard deviations from the expected RT, given that the response is correct or incorrect. Each item's response time is modeled using the following regression equation:

$$Response\ Time = intercept + \beta_1(Response) + error$$

where "*Response*" is a dichotomous variable indicating whether the response is correct or incorrect. In this model, each item has an expected response time (given a correct/incorrect response). If the examinee's response time is > 2.0 SD above or below the expected response time according to the model, then it is considered an outlier. The proportion of responses that are outliers is subtracted from 1.0 to arrive at a score for this metric, with higher values indicating greater validity. For example, if 10% of an individual's responses were outliers, the score for this metric would be 1.00–0.10 = 0.90 (90% valid/non-outliers).

"Easiest Items" metric

Similar to most standalone PVTs, this metric reflects items that are answered correctly by an overwhelming majority of individuals. This metric determines the proportion correct on the three easiest items on each measure, where "easiest" is indicated by the highest proportion correct in the full sample. Typically, these items were answered correctly by >95% of the reference sample. This metric has 4 possible values (0%, 33%, 67%, 100%). Note that this method will inevitably share some variance with person-fit methods but may be particularly sensitive to intentional underperformance, where a common putative strategy is to feign memory loss (Tan et al., 2002), resulting in easy items answered incorrectly.

Statistical analyses

All analyses were performed in R v3.6.1 (R Core Team, 2020).

Simulations

One set of simulations was performed for each test. First, item responses and response times were estimated using parameters generated from the PNC. Next, simulations were conducted using these parameters for $n = 2{,}000$ simulated participants. Item-level response patterns and response time estimates were used to generate varying proportions of invalid responses for a small percentage of simulated participants. Next, the three performance validity metrics were calculated for each measure for each simulated examinee. Finally, a model was built and cross-validated to predict invalid responses in the simulations. More specific steps for performing simulations were as follows:

1. To simulate item responses, we first estimated population parameters (e.g., discriminations, difficulties) of item-level data from the PNC, using the 2-parameter logistic IRT model (2PLM). For reference, eTable 2 shows the item-wise proportions correct for data on which simulations were based, and eTable 3 shows the 2PLM item parameter estimates used in simulations. In rare cases where a negative discrimination parameter was estimated in real data, it was fixed to zero before simulations. Of note, although *responses* from simulated examinees are unrelated to those from real examinees, simulated item responses have the exact same *parameters* (discrimination, difficulty) as items from the PNC. These item parameter estimates were used as population parameters for simulations in #3 below. See Supplementary Methods for details.

2. Response time distributions for correct and incorrect responses were estimated separately for each item using data from the PNC, using the JohnsonFit() function in the *SuppDists* package (Wheeler, 2016). These measures provide item-level response time estimates from a distribution of responses most similar to a target distribution. For each item of each test, the program generated two sets of four estimates (one set each for correct and incorrect responses), used in simulations described in #4 below.

3. Response patterns were then simulated for 2,000 examinees, using population parameters estimated from #1 above and the simdata() function in the *mirt* package (Chalmers, 2012). N = 2,000 was chosen because it is small enough to ensure some variability due to sample size – for example, two simulations of N = 2,000 will vary in their proportions correct (especially of the N = 100 invalid responders) – yet not so large that the phenomena demonstrated here may be limited to extremely large sample sizes, casting doubt on their utility.

4. For each simulated correct/incorrect item response from #3, response *time* was simulated by drawing randomly from a

[1] Note that the correlations used here are *point*-biserial, not biserial. The formula for a point-biserial correlation is identical to the formula for a Pearson correlation. The "point-biserial" term is used simply to indicate that it is a Pearson correlation between continuous and dichotomous variables (and therefore has a maximum absolute value < 1.0).

[2] Note that a weakness of these metrics is that they do not provide values for all-correct or all-incorrect response patterns, with the exception of the Kane-Brennan statistic, which shows perfect fit for perfect scores; therefore, such response patterns (all 0s or all 1s) are coded here as maximum fit.

[3] Note that all response times were log-transformed (natural log) before analysis. If raw RTs were used, all outliers would be on the positive end (too slow) due to the inherent positive skew of response time distributions.

https://doi.org/10.1017/S1355617722000893 Published online by Cambridge University Press

distribution with the same moments as those estimated in #2, using the rJohnson() function in the *SuppDists* package (Wheeler, 2016). Therefore, if real-world examinees tended to have highly skewed response times for Item X on Test Y, response times for simulated examinees for Item X on Test Y were drawn from a similarly skewed distribution.

5. For a subset of the 2,000 simulated examinees, a percentage of their responses (see #8 below) were changed to be "invalid." For these simulated low-validity examinees, the probability of a correct response on a percentage of responses was set at chance level (i.e., unrelated to discrimination and difficulty item parameters), and response times for these chance-level responses were drawn from a random uniform distribution between 200 and 3000 milliseconds. This randomness was meant to simulate careless responding insofar as response times and accuracy were unrelated to the discrimination or difficulty of the item (analogous to random or unmotivated responding). The "invalid" nature of response patterns of these randomly selected examinees was indicated (value=1) in an additional column. Other examinees received a 0 in this column.

Primary analyses simulated invalid performance for 100 (5%) of the 2,000 examinees, which was selected as a reasonable base rate of non-credible performance for non-clinical/community samples (Martin & Schroeder, 2020; Ross et al., 2016). However, given research in healthy undergraduates that found higher rates of invalid performance (Roye et al., 2019), we also modeled invalid performance in 200 (10%) simulated participants as a sensitivity test.

6. With valid and invalid response patterns simulated, performance validity metrics were calculated for each measure for each simulated examinee. Person-fit metrics were calculated using r.pbis() and E.KB() functions in the *PerFit* package (Tendeiro et al., 2016).

7. To assess the performance of the validity metrics in detecting invalid responders in this simulation, we split the sample randomly into a training (75%) and testing (25%) set. In the training set, a logistic regression model was built to predict invalid responders using the new performance validity metrics. Then, the model (built in the training set) was used to predict invalid responses in the testing set, with area under the receiver operating characteristic (ROC) curves generated using each performance validity metric individually (e.g., only the person-fit metric) and using all metrics combined. Of note, to eliminate collinearity among performance validity metrics, they were regressed out of each other before building models that included them all.

8. Steps 3 through 7 were repeated for all possible percentages of invalid responses (#5 above) per test. For example, for the PMAT (24 items), a set of simulations was conducted for when 100 examinees had invalid responses to 24 items (24/24 = 100% invalid), 23 items (23/24 = 95.8% invalid), 22 items (22/24 = 91.7% invalid), etc., to only 1 item invalid. For exploratory purposes, a response pattern simulated to be invalid was considered invalid even if it had only one "random" response. Although examining a large range of invalid responses may not reflect distributions of invalid responses in real-world participants, this method avoids choosing an arbitrary cutoff and examines the methods' utility across a distribution of levels of invalid responding. The ability of performance validity metrics to detect invalid response patterns in the test sample was recorded at each possible percentage of invalid responses. For each test and each percentage of invalid responses, 500 simulations (2000 examinees each) were run. In total, there were: 500 simulations × 179 total items (across 6 tests) = 89,500 simulations total.

Coefficients for each simulation were saved, and final hard-coded equations below represent the average across all simulations.

Application of validity estimates to developmental and adult neurocognitive data

Next, as an initial test of criterion validity, we examined whether the multivariate CNB validity estimates (MCVE) were associated with established CNB validation rules. We generated MCVEs for PNC and MRS-II participants and used these scores to predict previously established valid/invalid classifications for each test using logistic regressions and ROC curves. Specifically, equations created from "Simulations" above were used to calculate a weighted sum of the four individual validity metrics per person, providing out-of-sample validity estimates. These validity estimates were used to "predict" valid/invalid classification rules determined by a human team. As detailed above, although these are not gold standard PVTs, they serve as an important proof-of-concept test.

Finally, to assess whether the MCVE could be confounded by clinical conditions (depression, anxiety, etc.), we examined MCVE data for real PNC participants with and without six psychopathological diagnoses: depression, obsessive-compulsive disorder (OCD), post-traumatic stress disorder (PTSD), psychosis, attention deficit hyperactivity disorder (ADHD), and conduct disorder. See supplementary methods and Calkins et al. (2015) for greater details on psychopathology assessment and Calkins et al. (2014) for clinical composition of the sample. Similar analyses were conducted for MRS-II; because clinical assessments differed, we compared participants with and without: depression, PTSD, anxiety, insomnia, alcohol problems, psychosomatic/physical problems, and marked lack of perspective-taking (low interpersonal reactivity). Specific instruments can be found in Supplemental Methods, and details of score generation for the seven MRS-II psychopathology scales and clinical composition can be found in Moore et al. (2017).

Results

Simulations

We first used simulations and the MCVE prediction model incorporating our three performance validity metrics to classify simulated examinees as valid versus invalid responders. Figure 1 shows the relationship between simulated effort (number of valid responses across items) and the model's ability to classify response patterns as valid or invalid when simulating 5% of examinees as invalid. As expected, the more valid responses a simulated examinee gave (i.e., moving from left to right on the x-axis), the ability of the MCVE to detect invalid responding in these simulations decreased. Further, the individual metrics (e.g., person-fit) differed by test in how well they detected invalid responses. For example, for the ER40, individual metrics except the "easiest items" metric perform well (AUC > 0.80) up to 75% valid responding. By contrast, for the VOLT, the relative utility of the methods was less clear; the easiest items metric outperformed RT outlier and person-fit metrics between ∼30% and ∼90% valid responses, though all fell below AUC = 0.80 when 60% or more of responses were valid. Notably, for every test except PMAT, the combined MCVE performed marginally to substantially better than any individual metric, supporting the utility of our approach. See Supplemental Results for greater details on the shape of curves in Figure 1. As shown in eFigure 1, sensitivity analyses showed similar results when simulating 10% of examinees as invalid, supporting results

https://doi.org/10.1017/S1355617722000893 Published online by Cambridge University Press

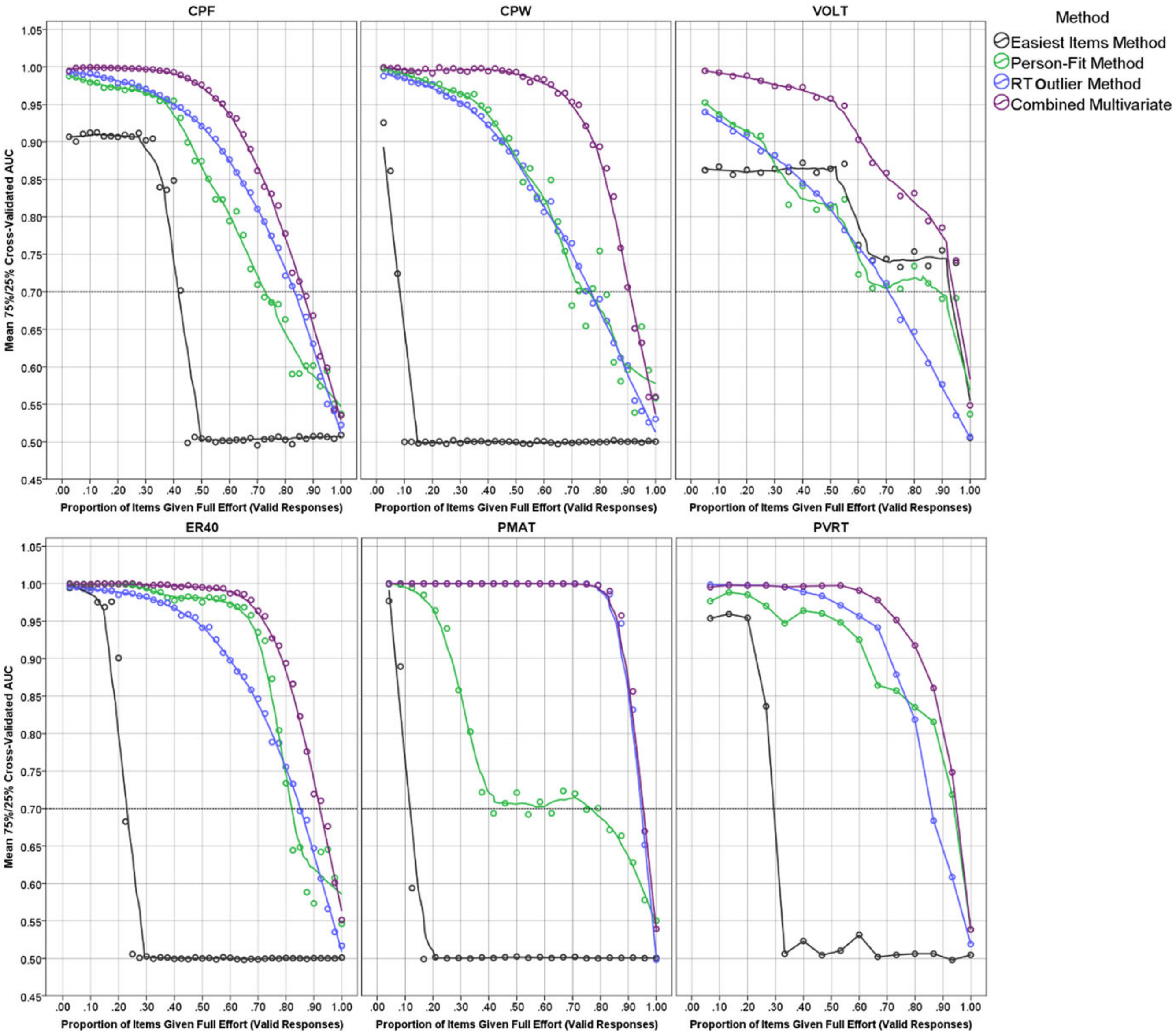

Figure 1. Prediction accuracy (AUC) for the multivariate CNB performance validity estimate in predicting true (Simulated) careless responding in 5% of the simulated examinees, by proportions of valid responses. *Note*. For visual simplification, "Person-Fit Method" is the average of the person-fit indices. CPF = Penn Face Memory Test; CPW = Penn Word Memory Test; VOLT = Visual Object Learning Test; ER40 = Penn Emotion Recognition Test; PMAT = Penn Matrix Analysis Test; PVRT = Penn Verbal Reasoning Test.

above. Additionally, eFigure 2 shows that increases in the proportion of items invalid for each test generally results in decreased internal consistency (Cronbach's α), supporting effects on validity.

The simulations summarized in Figure 1 were used to generate hard-coded equations for use in real data:

$$\mathrm{MCVE}_{\mathrm{mem}} = 0.42 \times \mathrm{Outlier\_Score} + 0.02 \times \mathrm{Acc\_3easy} + 0.05 \times \mathrm{PersonFit1} + 0.50 \times \mathrm{PersonFit2}$$

$$\mathrm{MCVE}_{\mathrm{non\text{-}mem}} = 0.34 \times \mathrm{Outlier\_Score} + 0.00 \times \mathrm{Acc\_3easy} + 0.22 \times \mathrm{PersonFit1} + 0.44 \times \mathrm{PersonFit2}$$

where "$\mathrm{MCVE}_{\mathrm{mem}}$" is for episodic memory tests, "$\mathrm{MCVE}_{\mathrm{non\text{-}mem}}$" is for all non-memory tests, "Outlier_Score" is the RT outlier validity metric, "Acc_3easy" is the easiest items metric, "PersonFit1" is the point-biserial metric, "PersonFit2" is the Kane-Brennan dependability metric, and coefficients reflect averages across each test type (memory and non-memory). Note that non-memory tests are combined because we intend for the MCVE to be generalizable across multiple tests. Furthermore, there is domain specificity in PVTs that might be lost when aggregating across memory and non-memory domains (Erdodi, 2019). Supplementary file "MCVE_estimation.R" provides the R script for calculating validity estimates, where memory and non-memory versions can be specified by commenting out relevant lines indicated.

Application of validity metrics to two independent samples

Next, these equations were used to generate MCVE scores for participants in the PNC and MRS-II, which were used to predict valid/invalid classifications previously determined by the investigators' quality assurance team for each test. eFigure 3 shows results of these ROC curves for the PNC. AUCs ranged from 0.61 for the

https://doi.org/10.1017/S1355617722000893 Published online by Cambridge University Press

Table 1. In-Sample prediction statistics for the six neurocognitive tests predicting human-determined validity, in the Philadelphia Neurodevelopmental cohort (PNC) and marine resiliency Study-II (MRS-II) samples

	PNC				MRS-II			
Test	Thresh	Sens	Spec	AUC	Thresh	Sens	Spec	AUC
CPF	0.86	0.88	0.47	0.73	0.90	0.80	0.73	0.82
CPW	0.90	0.81	0.60	0.75	0.88	0.91	0.80	0.89
VOLT	0.88	0.85	0.54	0.74	0.84	0.93	0.80	0.93
PMAT	0.85	0.78	0.51	0.65	0.82	0.88	0.40	0.65
PVRT	0.83	0.67	0.71	0.72	0.64	0.94	0.80	0.89
ER40	0.87	0.64	0.70	0.71	0.81	0.80	1.00	0.96

Note. PNC = Philadelphia Neurodevelopmental Cohort; MRS-II = Marine Resiliency Study-II; Thresh = threshold; Sens = sensitivity; Spec = specificity; AUC = area under the receiver operating characteristic curve; mean threshold across samples = 0.84.

PMAT to 0.75 for the CPW. Using an AUC > 0.70 cutoff, the MCVE shows acceptable classification accuracy for four of the six tests. eFigure 4 shows ROC results for the MRS-II. AUCs ranged from 0.65 for the PMAT to 0.94 for the ER40. In this sample, the MCVE achieved acceptable (or remarkable) classification accuracy for five of six tests. To facilitate decisions regarding use of MCVE thresholds, eFigures 5 and 6 and Supplementary File MCVE_sensitivity_specificity_coordinates.csv show how sensitivity and specificity vary as this threshold is varied. Overall, the table suggests that a specificity of at least 0.80 can be achieved with a threshold of approximately 0.90 ± 0.05 on the MCVE. Table 1 presents in-sample prediction statistics for the PNC and MRS-II. See Supplementary Results for more information.

Associations with psychopathology

eFigure 7 shows how psychiatric diagnostic groups differed on MCVE scores in the PNC. For OCD, PTSD, and conduct disorder, no significant differences in MCVE were found. Depression was associated with higher MCVE (i.e., more valid performance) on the PVRT ($p < 0.005$), CPW ($p < 0.0005$), and VOLT ($p < 0.005$). Psychosis was associated with lower MCVE on the ER40 ($p < 0.05$) but higher MCVE on the CPW ($p < 0.0005$). Finally, ADHD was associated with lower MCVE on all tests ($p < 0.05$ to $p < 0.0005$).

eFigure 8 shows how psychiatric diagnostic groups differed on MCVEs in the MRS-II. For PTSD, anxiety, insomnia, alcohol problems, and lack of perspective-taking, no significant differences in MCVE were found. Depression was associated with lower MCVE on the ER40 ($p < 0.05$), and psychosomatic/physical problems were associated with lower MCVE on the PVRT ($p < 0.05$) and VOLT ($p < 0.0005$).

Discussion

Here, we used novel, item-level psychometrics to establish data-driven, embedded performance validity metrics for individual tests in a widely used computerized neurocognitive battery. Our performance validity metrics detected patterns of invalid responding in data that simulated careless responding, even at subtle levels. Moreover, a multivariate combination of these metrics predicted previously established validity rules on most measures in independent developmental and adult datasets. Furthermore, there were few differences on this measure by clinical diagnostic groups, even in diagnoses associated with reduced cognitive performance in prior studies (e.g., Barzilay et al., 2019; Gur et al., 2014; Kaczkurkin et al., 2020; Service et al., 2020). These results provide proof-of-concept evidence for the potential utility of data-driven performance validity metrics that leverage existing individual test data. Importantly, the MCVE also allows one to examine performance validity across a spectrum, providing greater flexibility for estimating the likelihood of invalid responses and offering more precise data for investigators interested in understanding their data at a fine-grained level. Though these methods were specifically applied to the PennCNB, they could be applied to any computerized test with accuracy and response time data.

Although robust measures exist to detect insufficient engagement on neurocognitive tests, such measures typically rely on cutoff scores, do not integrate response speed, do not allow one to examine variations in validity across a test battery, and are not scalable for large-scale data acquisition. Our metrics offer a method to examine gradients of performance validity, provide multiple measures across individual tests, and integrate response speed and item-level psychometrics. Similarly, item-level metrics from another computerized battery, the NIH Toolbox-Cognition, have recently shown promise predicting externally validated PVTs in a mild traumatic brain injury sample (Abeare et al., 2021). Such metrics may facilitate administration of neurocognitive tests through telehealth or other remote applications, as well as the integration of cognitive testing into “big data” efforts, including genomics. In addition, compared to standalone PVTs, the nature of these validity metrics may be less amenable to coaching, which can be problematic in forensic cases (Suhr & Gunstad, 2007). Examining variations in performance validity across a test battery may be especially informative in this regard, as sophisticated simulators of cognitive impairment may target specific tests rather than performing poorly across all measures (Lippa, 2018). Critically, we provide equations that PennCNB users can apply to examine performance validity in data already collected. However, these metrics should be integrated with other available data, where possible, including behavioral observations by test administrators and participant clinical history.

Importantly, performance validity metrics should be minimally associated with most clinical conditions, even those with established neurocognitive deficits. There were few associations between our validity metrics and diagnoses of depression, PTSD, anxiety, psychosis spectrum disorders, or externalizing disorders, supporting our approach. This feature is especially relevant considering neurocognitive performance differences in these conditions (e.g., Barzilay et al., 2019; Gur et al., 2014; Kaczkurkin et al., 2020). However, there were diagnoses with reduced validity metrics that should be addressed. First, ADHD showed MCVE reductions across all tests. It is unclear whether these reductions are measuring a behavioral phenotype of ADHD or reflect high base rates of non-credible performance in ADHD (Suhr & Berry, 2017). Caution is warranted in applying the MCVE to individuals with ADHD before additional research is conducted. Second, youth with significant psychotic symptoms and Marines with depression showed slightly lower MCVE on ER40, so caution is warranted in interpreting ER40 MCVE data in these populations. In contrast, lower MCVE from the Marines with psychosomatic/physical problems appears consistent with research showing that individuals with higher psychosomatic complaints have elevated rates of performance validity concerns (Dandachi-FitzGerald et al., 2016; Martin & Schroeder, 2020).

Limitations and future directions

Although initial results are promising, additional studies are needed for validation. As mentioned above, one limitation is the lack of a "gold standard" PVT criterion, which should be examined in future work. Future validation studies could take several forms, including comparison of the MCVE to multiple PVTs or known groups designs in which patients expected to display underperformance because of external incentives are compared to those without external incentives (see Larrabee, 2012). Though these methods could be applied to detect intentional underperformance, they are likely detecting unmotivated responding in these samples given the lack of external incentive for underperforming.

A second limitation is that data were simulated from unidimensional IRT models, though many tests will produce multidimensional data. For example, memory tests often produce data showing two separate dimensions reflecting the ability or tendency to detect a target (e.g., knowing one has seen a face previously) and to reject a foil (e.g., knowing one has *not* seen a face previously). Future studies should examine whether incorporating multidimensional psychometric modeling improves these measures. Relatedly, methods of detecting aberrant response patterns have advanced considerably since the development of metrics used here, especially in clinical and personality research (Falk & Ju, 2020; Lanning, 1991; Tellegen, 1988). While our metrics were selected based on generalizability (e.g., no reliance on model estimation) and appropriateness for neurocognitive tests, future work may benefit from testing more advanced methods.

A third limitation of the present study is that the weighted composites (equations on pg. 13) are based on simulations in which the number of invalid responders is fixed at 5%. While we tested 10% invalid in additional simulations, the weights in our equations could change slightly if the base rates of invalid performance were lower (e.g., high-stakes test) or higher (e.g., with external incentives to malinger).

Ensuring the validity of assessments is critical for accurate interpretation in medicine and psychology. However, PVTs arouse considerable controversy, as there can be substantial consequences to classifying an assessment as invalid, including stigma, loss of benefits, and psychological distress. Thus, most PVTs have emphasized specificity over sensitivity, which should be considered in applying these metrics. Relatedly, researchers must currently use subjective judgment (including examination of the MCVE distribution in their sample) to decide what proportion of the sample to flag or remove. Even in cases where the MCVE distribution shows extreme outliers, the judgment is still subjective, which is a weakness of the metrics proposed here. MCVEs have not been extensively examined in clinical or medicolegal contexts, and further research is needed before they are used in such contexts. Furthermore, neuropsychologists using PVTs should integrate multiple sources of data to determine invalid performance, including behavioral observations and expected cognitive performance given the patient's history (Heilbronner et al., 2009).

Conclusions

These limitations notwithstanding, we provide promising initial data on the development and application of novel, data-driven, dimensional performance validity metrics for individual tests in the PennCNB. We show that these metrics are sensitive to subtle patterns of invalid responding in simulated data, correspond well with established quality assurance metrics in two independent datasets, and are not associated with diagnoses of most psychiatric disorders. These methods may facilitate modeling of unmotivated, random, or disengaged responding in remote or large-scale cognitive data collection, enhancing the validity and precision of such assessments.

Acknowledgements. This work was supported by the NIH (HD095278; MH089983; MH117014, MH119219; T32MH019112) and the Penn-CHOP Lifespan Brain Institute. The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the University of Pennsylvania or the Department of Veterans Affairs. We thank the VISN4 MIRECC Writing Group for helpful comments on a draft of this manuscript. Portions of these data were presented at the 2021 Annual Meeting of the International Neuropsychological Society.

Conflicts of interest. All authors declare no potential conflicts of interest.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/S1355617722000893

References

Abeare, C., Erdodi, L., Messa, I., Terry, D. P., Panenka, W. J., Iverson, G. L., & Silverberg, N. D. (2021). Development of embedded performance validity indicators in the NIH Toolbox Cognitive Battery. *Psychological Assessment*, *33*, 90–96. https://doi.org/10.1037/pas0000958

Acheson, D. T., Geyer, M. A., Baker, D. G., Nievergelt, C. M., Yurgil, K., Risbrough, V. B., & MRS-II Team. (2015). Conditioned fear and extinction learning performance and its association with psychiatric symptoms in active duty Marines. *Psychoneuroendocrinology*, *51*, 495–505. https://doi.org/10.1016/j.psyneuen.2014.09.030

Aliyu, M. H., Calkins, M. E., Swanson, C. L., Lyons, P. D., Savage, R. M., May, R., & PAARTNERS Study Group. (2006). Project among African-Americans to explore risks for schizophrenia (PAARTNERS): Recruitment and assessment methods. *Schizophrenia Research*, *87*, 32–44. https://doi.org/10.1016/j.schres.2006.06.027

Barzilay, R., Calkins, M. E., Moore, T. M., Wolf, D. H., Satterthwaite, T. D., Cobb Scott, J., & Gur, R. E. (2019). Association between traumatic stress load, psychopathology, and cognition in the Philadelphia Neurodevelopmental Cohort. *Psychological Medicine*, *49*, 325–334. https://doi.org/10.1017/S0033291718000880

Basner, M., Savitt, A., Moore, T. M., Port, A. M., McGuire, S., Ecker, A. J., & Gur, R. C. (2015). Development and validation of the cognition test battery for spaceflight. *Aerospace Medicine and Human Performance*, *86*, 942–952. https://doi.org/10.3357/AMHP.4343.2015

Bhatia, T., Mazumdar, S., Wood, J., He, F., Gur, R. E., Gur, R. C., & Deshpande, S. N. (2017). A randomised controlled trial of adjunctive yoga and adjunctive physical exercise training for cognitive dysfunction in schizophrenia. *Acta Neuropsychiatrica*, *29*, 102–114. https://doi.org/10.1017/neu.2016.42

Bilder, R. M., & Reise, S. P. (2019). Neuropsychological tests of the future: How do we get there from here? *The Clinical Neuropsychologist*, *33*, 220–245. https://doi.org/10.1080/13854046.2018.1521993

Calkins, M. E., Merikangas, K. R., Moore, T. M., Burstein, M., Behr, M. A., Satterthwaite, T. D., & Gur, R. E. (2015). The Philadelphia Neurodevelopmental Cohort: Constructing a deep phenotyping collaborative. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *56*, 1356–1369. https://doi.org/10.1111/jcpp.12416

Calkins, M. E., Moore, T. M., Merikangas, K. R., Burstein, M., Satterthwaite, T. D., Bilker, W. B., & Gur, R. E. (2014). The psychosis spectrum in a young U.S. community sample: Findings from the Philadelphia Neurodevelopmental Cohort. *World Psychiatry*, *13*, 296–305. https://doi.org/10.1002/wps.20152

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*, 1–29.

Dandachi-FitzGerald, B., van Twillert, B., van de Sande, P., van Os, Y., & Ponds, R. W. H. M. (2016). Poor symptom and performance validity in regularly referred Hospital outpatients: Link with standard clinical measures, and role of incentives. *Psychiatry Research*, *239*, 47–53. https://doi.org/10.1016/j.psychres.2016.02.061

Donlon, T. F., & Fischer, F. E. (1968). An index of an individual's agreement with group-determined item difficulties. *Educational and Psychological Measurement*, *28*, 105–113.

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*, 155–172. https://doi.org/10.1080/23279095.2017.1384925

Falk, C. F., & Ju, U. (2020). Estimation of response styles using the multidimensional nominal response model: a tutorial and comparison with sum scores. *Frontiers in Psychology*, *11*, 72. https://doi.org/10.3389/fpsyg.2020.00072

Garrett-Bakelman, F. E., Darshi, M., Green, S. J., Gur, R. C., Lin, L., Macias, B. R., & Turek, F. W. (2019). The NASA Twins Study: a multidimensional analysis of a year-long human spaceflight. *Science*, *364*, eaau8650. https://doi.org/10.1126/science.aau8650

Glahn, D. C., Gur, R. C., Ragland, J. D., Censits, D. M., & Gur, R. E. (1997). Reliability, performance characteristics, construct validity, and an initial clinical application of a visual object learning test (VOLT). *Neuropsychology*, *11*, 602–612. https://doi.org/10.1037//0894-4105.11.4.602

Greenwood, T. A., Lazzeroni, L. C., Maihofer, A. X., Swerdlow, N. R., Calkins, M. E., Freedman, R., & Braff, D. L. (2019). Genome-wide association of endophenotypes for schizophrenia from the Consortium on the Genetics of Schizophrenia (COGS) study. *JAMA Psychiatry 76*, 1274–1284. https://doi.org/10.1001/jamapsychiatry.2019.2850

Gulsuner, S., Stein, D. J., Susser, E. S., Sibeko, G., Pretorius, A., Walsh, T., & McClellan, J. M. (2020). Genetics of schizophrenia in the South African Xhosa. *Science*, *367*, 569–573. https://doi.org/10.1126/science.aay8833

Gur, R. C., Calkins, M. E., Satterthwaite, T. D., Ruparel, K., Bilker, W. B., Moore, T. M., & Gur, R. E. (2014). Neurocognitive growth charting in psychosis spectrum youths. *JAMA Psychiatry*, *71*, 366–374. https://doi.org/10.1001/jamapsychiatry.2013.4190

Gur, R. C., Ragland, J. D., Moberg, P. J., Bilker, W. B., Kohler, C., Siegel, S. J., & Gur, R. E. (2001). Computerized neurocognitive scanning: II. The profile of schizophrenia. *Neuropsychopharmacology*, *25*, 777–788. https://doi.org/10.1016/S0893-133X(01)00279-2

Gur, R. C., Richard, J., Calkins, M. E., Chiavacci, R., Hansen, J. A., Bilker, W. B., & Gur, R. E. (2012). Age group and sex differences in performance on a computerized neurocognitive battery in children age 8–21. *Neuropsychology*, *26*, 251–265. https://doi.org/10.1037/a0026712

Gur, R. C., Richard, J., Hughett, P., Calkins, M. E., Macy, L., Bilker, W. B., & Gur, R. E. (2010). A cognitive neuroscience-based computerized battery for efficient measurement of individual differences: Standardization and initial construct validation. *Journal of Neuroscience Methods*, *187*, 254–262. https://doi.org/10.1016/j.jneumeth.2009.11.017

Hartung, E. A., Kim, J. Y., Laney, N., Hooper, S. R., Radcliffe, J., Port, A. M., & Furth, S. L. (2016). Evaluation of Neurocognition in Youth with CKD Using a Novel Computerized Neurocognitive Battery. *Clinical Journal of the American Society of Nephrology*, *11*, 39–46. https://doi.org/10.2215/CJN.02110215

Heilbronner, R. L., Sweet, J. J., Morgan, J. E., Larrabee, G. J., Millis, S. R., & Conference Participants. (2009). American Academy of Clinical Neuropsychology Consensus Conference Statement on the neuropsychological assessment of effort, response bias, and malingering. The Clinical Neuropsychologist, 23, 1093–1129. https://doi.org/10.1080/13854040903155063

Kaczkurkin, A. N., Sotiras, A., Baller, E. B., Barzilay, R., Calkins, M. E., Chand, G. B., & Satterthwaite, T. D. (2020). Neurostructural heterogeneity in youths with internalizing symptoms. *Biological Psychiatry*, *87*, 473–482. https://doi.org/10.1016/j.biopsych.2019.09.005

Kane, M. T., & Brennan, R. L. (1980). Agreement coefficients as indices of dependability for domain-referenced tests. *Applied Psychological Measurement*, *4*, 105–126.

Kanser, R. J., Rapport, L. J., Bashem, J. R., & Hanks, R. A. (2019). Detecting malingering in traumatic brain injury: Combining response time with performance validity test accuracy. *The Clinical Neuropsychologist*, *33*, 90–107. https://doi.org/10.1080/13854046.2018.1440006

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, *16*, 277–298.

Lanning, K. (1991). *Consistency, Scalability, and Personality Measurement*. Springer New York. https://doi.org/10.1007/978-1-4612-3072-4

Larrabee, G. J. (2012). Performance validity and symptom validity in neuropsychological assessment. *Journal of the International Neuropsychological Society*, *18*, 625–630. https://doi.org/10.1017/S1355617712000240

Lippa, S. M. (2018). Performance validity testing in neuropsychology: A clinical guide, critical review, and update on a rapidly evolving literature. *The Clinical Neuropsychologist*, *32*, 391–421. https://doi.org/10.1080/13854046.2017.1406146

Loring, D. W., & Goldstein, F. C. (2019). If invalid PVT scores are obtained, can valid neuropsychological profiles be believed? *Archives of Clinical Neuropsychology*, *34*, 1192–1202. https://doi.org/10.1093/arclin/acz028

Lupu, T., Elbaum, T., Wagner, M., & Braw, Y. (2018). Enhanced detection of feigned cognitive impairment using per item response time measurements in the Word Memory Test. *Applied Neuropsychology. Adult*, *25*, 532–542. https://doi.org/10.1080/23279095.2017.1341410

Martin, P. K., & Schroeder, R. W. (2020). Base rates of invalid test performance across clinical non-forensic contexts and settings. *Archives of Clinical Neuropsychology*. https://doi.org/10.1093/arclin/acaa017

McCormick, C. L., Yoash-Gantz, R. E., McDonald, S. D., Campbell, T. C., & Tupler, L. A. (2013). Performance on the Green Word Memory test following operation enduring freedom/operation Iraqi freedom-era military service: Test failure is related to evaluation context. *Archives of Clinical Neuropsychology*, *28*, 808–823. https://doi.org/10.1093/arclin/act050

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, *25*, 107–135.

Moore, T. M., Reise, S. P., Gur, R. E., Hakonarson, H., & Gur, R. C. (2015). Psychometric properties of the Penn Computerized Neurocognitive Battery. *Neuropsychology*, *29*, 235–246. https://doi.org/10.1037/neu0000093

Moore, T. M., Risbrough, V. B., Baker, D. G., Larson, G. E., Glenn, D. E., Nievergelt, C. M., & Gur, R. C. (2017). Effects of military service and deployment on clinical symptomatology: The role of trauma exposure and social support. *Journal of Psychiatric Research*, *95*, 121–128. https://doi.org/10.1016/j.jpsychires.2017.08.013

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org/

Raven, J. (1989). The Raven Progressive Matrices: A review of national norming studies and ethnic and socioeconomic variation within the United States. *Journal of Educational Measurement*, *26*, 1–16.

Reise, S. P. (1990). A comparison of item-and person-fit methods of assessing model-data fit in IRT. *Applied Psychological Measurement*, *14*, 127–137.

Roalf, D. R., Gur, R. C., Almasy, L., Richard, J., Gallagher, R. S., Prasad, K., & Gur, R. E. (2013). Neurocognitive performance stability in a multiplex multigenerational study of schizophrenia. *Schizophrenia Bulletin*, *39*, 1008–1017. https://doi.org/10.1093/schbul/sbs078

Ross, T. P., Poston, A. M., Rein, P. A., Salvatore, A. N., Wills, N. L., & York, T. M. (2016). Performance invalidity base rates among healthy undergraduate research participants. *Archives of Clinical Neuropsychology*, *31*, 97–104. https://doi.org/10.1093/arclin/acv062

Roye, S., Calamia, M., Bernstein, J. P. K., De Vito, A. N., & Hill, B. D. (2019). A multi-study examination of performance validity in undergraduate research participants. *The Clinical Neuropsychologist*, *33*, 1138–1155. https://doi.org/10.1080/13854046.2018.1520303

Scott, J. C., Lynch, K. G., Cenkner, D. P., Kehle-Forbes, S. M., Polusny, M. A., Gur, R. C., & Oslin, D. W. (2021). Neurocognitive predictors of treatment outcomes in psychotherapy for comorbid PTSD and substance use disorders.

https://doi.org/10.1017/S1355617722000893 Published online by Cambridge University Press

Journal of Consulting and Clinical Psychology, *89*, 937–946. https://doi.org/10.1037/ccp0000693

Scott, J. C., Moore, T. M., Stein, D. J., Pretorius, A., Zingela, Z., Nagdee, M., & Gur, R. C. (2021). Adaptation and validation of a computerized neurocognitive battery in the Xhosa of South Africa. *Neuropsychology*, *35*, 581–594. https://doi.org/10.1037/neu0000742

Service, S. K., Vargas Upegui, C., Castaño Ramírez, M., Port, A. M., Moore, T. M., Munoz Umanes, M., & Freimer, N. B. (2020). Distinct and shared contributions of diagnosis and symptom domains to cognitive performance in severe mental illness in the Paisa population: A case-control study. *The Lancet. Psychiatry*, *7*, 411–419. https://doi.org/10.1016/S2215-0366(20)30098-5

Suhr, J. A., & Berry, D. T. R. (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*, 1427–1428. https://doi.org/10.1037/pas0000535

Suhr, J. A., & Gunstad, J. (2007). Coaching and malingering: A review. Assessment of malingered neuropsychological deficits. Oxford University Press. pp. 287–311.

Tan, J. E., Slick, D. J., Strauss, E., & Hultsch, D. F. (2002). How'd They Do It? Malingering strategies on symptom validity tests. *The Clinical Neuropsychologist*, *16*, 495–505. https://doi.org/10.1076/clin.16.4.495.13909

Tatsuoka, K. K., & Linn, R. L. (1983). Indices for detecting unusual patterns: Links between two general approaches and potential applications. *Applied Psychological Measurement*, *7*, 81–96.

Tellegen, A. (1988). The analysis of consistency in personality assessment. *Journal of Personality*, *56*, 621–663. https://doi.org/10.1111/j.1467-6494.1988.tb00905.x

Tendeiro, J. N., Meijer, R. R., & Niessen, A. S. M. (2016). PerFit: An R package for person-fit analysis in IRT. *Journal of Statistical Software*, *74*, 1–27.

Thomas, M. L., Brown, G. G., Gur, R. C., Hansen, J. A., Nock, M. K., Heeringa, S., & Stein, M. B. (2013). Parallel psychometric and cognitive modeling analyses of the Penn Face Memory Test in the army study to assess risk and resilience in servicemembers. *Journal of Clinical and Experimental Neuropsychology*, *35*, 225–245. https://doi.org/10.1080/13803395.2012.762974

Walters, G. D., Berry, D. T. R., Rogers, R., Payne, J. W., & Granacher, R. P. (2009). Feigned neurocognitive deficit: Taxon or dimension? *Journal of Clinical and Experimental Neuropsychology*, *31*, 584–593. https://doi.org/10.1080/13803390802363728

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale–Fourth Edition (WAIS–IV)*. NCS Pearson.

Wheeler, B. (2016). SuppDists: Supplementary distributions. R package version 1.1.-9.4. https://CRAN.R-project.org/package=SuppDists

https://doi.org/10.1017/S1355617722000893 Published online by Cambridge University Press