Psychological Assessment

Development and Public Release of the Penn Reading Assessment Computerized Adaptive Test (PRA-CAT) for Premorbid IQ

Mikhal A. Yudien, Tyler M. Moore, Allison M. Port, Kosha Ruparel, Raquel E. Gur, and Ruben C. Gur

Online First Publication, June 13, 2019. http://dx.doi.org/10.1037/pas0000738

American Psychological Association

Psychological Assessment

1040-3590/19/$12.00

http://dx.doi.org/10.1037/pas0000738

BRIEF REPORT

Development and Public Release of the Penn Reading Assessment Computerized Adaptive Test (PRA-CAT) for Premorbid IQ

Mikhal A. Yudien
University of Pennsylvania and Swarthmore College

Tyler M. Moore, Allison M. Port, Kosha Ruparel, Raquel E. Gur, and Ruben C. Gur
University of Pennsylvania

An important component of neuropsychological testing is assessment of premorbid intelligence to estimate a patient's ability independent of neurological impairment. A common test of premorbid IQ—namely, the Reading section of the Wide Range Achievement Test (WRAT)—has been shown to have high measurement error in the high ability range, is unnecessarily long (55 items), and is proprietary. We describe the development of an alternative, nonproprietary, computerized adaptive test for premorbid IQ, the Penn Reading Assessment (PRA-CAT). PRA-CAT items were calibrated using a 1-parameter item response theory model in a large community sample (N = 9,498), Ages 8 to 21, and the resulting parameters were used to simulate computerized adaptive testing sessions. Simulations demonstrated that the PRA-CAT achieves low measurement error (0.25; equivalent to Cronbach's alpha = .94) and acceptable measurement error (0.40; Cronbach's alpha = .84) after only 18 and 6 items, respectively (on average). Correlation of WRAT and PRA-CAT scores with numerous clinical, cognitive, demographic, and neuroimaging criteria suggests that validity of PRA-CAT score interpretation is comparable (and sometimes superior) with the WRAT. The fully functioning PRA-CAT for public use (including item parameter estimates reported here) has been built using the open-source program Concerto, and can be installed by anyone on a local computer or on the "cloud." Given the length and proprietary nature of the WRAT, the PRA-CAT shows promise as a potential alternative (and with minimal or no cost). Further validation in the context of neurological injury is needed.

Public Significance Statement
Clinicians often wish to assess the severity of disease- or injury-related cognitive decline, which can be difficult if premorbid cognitive ability is unknown. The PRA-CAT provides a method of assessing this premorbid ability (IQ) that is more efficient than current methods of doing so.

Keywords: premorbid IQ, computer adaptive testing, reading assessment, IQ

Supplemental materials: http://dx.doi.org/10.1037/pas0000738.supp

An important component of neuropsychological evaluation is assessment of premorbid intelligence to estimate a patient's general cognitive capability before neurological impairment. Premorbid IQ tests gauge abilities that remain relatively unaffected by neurological insult and typically rely on one of three approaches: (a) performance on achievement test reading sections ("hold" tests) such as subtests of the Wide Range Achievement Test (WRAT; Jastak & Jastak, 1965; Wilkinson & Robertson, 2006), Wechsler Adult Intelligence Scale (Psychological Corporation, 2002), and National Adult Reading Test (NART; Nelson & Willison, 1991); (b) regression analyses based on demographic information, such as the Barona formula (Barona & Chastain, 1986); and (c) a combination of both reading test performance and demographic information, such as the Oklahoma Premorbid Intelligence Estimate (Krull, Scott, & Sherer, 1995) and BEST-3 (Vanderploeg, Schinka, & Axelrod, 1996).

The method of analyzing performance on reading tests to assess premorbid IQ is used most extensively in neuropsychological testing, as such tests are rapid, easily administered, and relatively

Mikhal A. Yudien, Brain Behavior Laboratory, Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania, and Department of Biology and Department of Psychology, Swarthmore College; Tyler M. Moore, Allison M. Port, Kosha Ruparel, Raquel E. Gur, and Ruben C. Gur, Brain Behavior Laboratory, Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania.

This work was supported by the Lifespan Brain Institute; National Institute of Mental Health grants MH089983, MH019112, MH096891, MH107235, and MH107703; and the Dowshen Program for Neuroscience.

Correspondence concerning this article should be addressed to Tyler M. Moore, Brain Behavior Laboratory, Perelman School of Medicine, University of Pennsylvania, 3400 Spruce Street, Philadelphia, PA 19104. E-mail: tymoore@pennmedicine.upenn.edu

inexpensive. In addition, reading ability has been a reliable predictor of general intelligence across various populations (Contador, Bermejo-Pareja, Del Ser, & Benito-León, 2015) as well as predictive of changes in teenagers' verbal IQ (Ramsden et al., 2013). Substandard reading ability has been related to poor health outcomes (Berkman et al., 2004), and it has been shown to predict poor general cognitive functioning even more than does level of education (Dotson, Kitner-Triolo, Evans, & Zonderman, 2009).

The WRAT, first developed in the 1940s (background covered in Jastak & Jastak, 1965), has undergone several revisions to produce the latest versions: WRAT-4 (Wilkinson & Robertson, 2006) and WRAT-5 (Wilkinson & Robertson, 2017). WRAT has been consistently among the top 10 most commonly used neuropsychological tests (Camara, Nathan, & Puente, 2000; Rabin, Barr, & Burton, 2005). The WRAT-4, which consists of four subtests, serves as a rapid initial assessment of academic abilities such as word reading, sentence comprehension, spelling, and math. The WRAT-4 Word Reading subtest (READ) requires that participants read a list of phonetically regular and irregular words and, in the case of some lower ability participants, that they read a list of letters. Performance on the READ serves as a proxy for premorbid IQ—for example, Griffin, Mindt, Rankin, Ritchie, and Scott (2002)—as well as a predictor of neurocognitive functioning (Sayegh, Arentoft, Thaler, Dean, & Thames, 2014) that is relatively resistant to several forms of neurologic insult (Casaletto et al., 2014; Seidman et al., 2016). READ estimates premorbid neurocognitive functioning by testing vocabulary and recognition of words, relying upon the presumption that word reading is primarily dependent upon previous knowledge rather than current cognitive capabilities. The online supplemental materials include a review of WRAT validation efforts.

In studies of both neurologically impaired and unimpaired populations, it has been reported that a previous version of READ (WRAT-3; Jastak & Wilkinson, 1984) as well as an alternative reading test, NART, despite being accurate measures of average IQ, underestimate the IQ of those in higher intelligence ranges (Johnstone, Callahan, Kapila, & Bouman, 1996; Wiens, Bryan, & Crossen, 1993). Thus, there is a need for a test of premorbid functioning that is more suitable for populations with an average intelligence range as well as for those with a wider range of intelligence. Additionally, as the WRAT is a proprietary test owned by Pearson Education, the availability of a universally accessible, free alternative reading test will serve to benefit future neuropsychological testing. Perhaps most uniquely, no computerized adaptive testing (CAT) measure of reading ability has yet been developed and validated.

The Penn Reading Assessment (PRA) was developed in 2012 during assessment of the Philadelphia Neurodevelopmental Cohort (PNC) by the Brain Behavior Laboratory of The University of Pennsylvania Perelman School of Medicine. Supplementary Table S1 of the online supplemental materials shows the stimuli for each PRA version, and information about how words were selected for the PRA is also provided in the online supplemental materials. The purpose of the present investigation was to psychometrically analyze the PRA as a potential alternative adaptive reading test that can be used to estimate premorbid IQ. Two versions of the PRA were designed as a reading test normed on individuals aged 8 to 21 years that, like READ and similar measures, consists of phonetically regular and irregular words. The psychometric analyses conducted here allowed us to construct a CAT form of the PRA capable of achieving acceptable measurement error in a fraction of the time necessary for the full forms (or the WRAT).

Method

Participants in the PNC PRA standardization subsample consisted of 3,185 youths aged 8 to 21 years (50.2% female), of a total of 9,498 PNC participants assessed by the Brain Behavior Laboratory between 2009 and 2011. All participants received the WRAT, but the PRA was introduced later during the accrual process and was administered to a subsample. Participants were proficient in English and were recruited from the greater Philadelphia, Pennsylvania, community (not clinics) for the Grand Opportunity study funded by the National Institute of Mental Health, as previously described (Calkins et al., 2014, 2015; Gur et al., 2010). All procedures were approved by the University of Pennsylvania and Children's Hospital of Philadelphia Institutional Review Boards. Supplementary Table S2 of the online supplemental materials presents demographic and clinical information on the samples, which are nearly identically matched on age ($M = \sim$13.5 years), race, sex (~50% female), parental education, and psychiatric symptomatology. Note that because the PNC was a community sample, rates of psychiatric illness are comparable with the general population.

As part of the Penn Computerized Neurocognitive Battery (CNB), the WRAT-4 Blue Reading Form was administered first in the battery for all 9,498 PNC participants tested. In addition, the PRA forms were each randomly assigned to 3,185 participants recruited later in the study, and the PRA was administered fifth in the battery for those tested. That is, each of the 3,185 participants was administered both the WRAT and either the PRAa form or the PRAd form. We found no significant difference in gender distribution or age between the groups that received the PRAa versus the PRAd.

Analyses were performed in the *psych* (Revelle, 2018) or *mirt* (Chalmers, 2012) packages in R (R Core Team, 2018). The analysis pipeline described here was designed for construction of a computerized adaptive PRA. We first assessed dimensionality—that is, is the PRA unidimensional enough for item response theory (IRT; Reise, Cook, & Moore, 2015)?—by examining the ratio of first to second eigenvalues of the tetrachoric correlation matrix. One convention is that, if this ratio is greater than 3, a test or scale is considered to be unidimensional (Slocum-Gori & Zumbo, 2011). Relatedly, we assessed internal consistency using Cronbach's alpha.

With dimensionality and internal consistency established, we estimated a one-parameter IRT model described by the following equation:

$$p_i(\theta_j = \frac{1}{1 + e^{-a(\theta - b_i)}}, \quad (1)$$

where $p_i(\theta_j)$ is the probability of a correct response for person j on item i, a is the item discrimination (estimated but constant across all items), b_i is the difficulty of item i, and θ_j is the trait level of person j. The discrimination parameter, a, determines how precisely an item can place an individual on a trait spectrum; higher discrimination is always better. The difficulty parameter, b_i, determines how high on the trait continuum one has to be to have a

50% chance of responding correctly. This model was used for each form of the PRA.

The parameters described in this section and in Equation 1 pave the way for one of the most groundbreaking applications of IRT—CAT (Wainer & Dorans, 2000; Weiss & Kingsbury, 1984)—which we have previously used to develop short and adaptive forms of cognitive and clinical tests (Moore, Calkins, Reise, Gur, & Gur, 2018; Moore, Scott, et al., 2015; Roalf et al., 2016). In CAT, after the first item administration (and response), a scoring algorithm estimates the examinee's trait level (ability), and based on this rough estimate, chooses the most appropriate next item to administer, in which "most appropriate" is determined by how much information[1] it will provide. After this next item administration (and response), the algorithm then uses both item responses to estimate ability. Then, the next most appropriate item is selected, and so on. The test stops when some stopping criterion is met—for example, when the examinee's standard error of measurement reaches some lower bound. Note that because IRT scoring is based on the item parameters (a and b_i), endorsement of an "easy" item will affect a person's score differently than endorsement of a "difficult" item. Because all items were calibrated in the same model (Equation 1), *examinees can be scored on the same scale even if the sets of items administered to each examinee are completely nonoverlapping*. This feature is especially useful for longitudinal studies, in which item repetition can be problematic.

Item parameters estimates were input in Firestar (Choi, 2009), a CAT simulation program, to determine what would have happened if the PRA-CAT had been administered in the original sample. The first item was selected based on maximum information at the mean ($\theta = 0$), and the stopping criterion was to stop when the examinee's standard error of measurement reaches 0.25 or lower. Other Firestar settings were left as default—namely, the item selection method was maximum posterior weighted information (van der Linden, 1998) and the interim theta estimator was expected a posteriori (Bock & Mislevy, 1982).

Scores resulting from the above-described simulated adaptive versions were then used for validation purposes. The adaptive (PRA-CAT) scores were compared with the full-length PRA and WRAT scores in their relationship with numerous validity criteria. These criteria included age, sex, race, mother's education, trauma exposure (count of traumatic experiences, out of nine), global assessment of functioning, other neurocognitive performance, psychopathology, census-based measures of socioeconomic status (SES), and measures of brain volume. Neurocognitive performance was assessed using the Penn CNB (Gur et al., 2010; Moore, Reise, et al., 2015), clinical psychopathology was assessed using a modified version of the KIDDIE-SADS described in Calkins et al. (2014), census-based measures of SES are described in Moore et al. (2016), and neuroimaging in this cohort has been described extensively (Satterthwaite et al., 2014, 2016). Relationships between premorbid IQ (as measured by the WRAT and PRA) and these criteria were compared across form and test. For example, the relationship between PRA-CATa and age was compared with the relationship between the WRAT and age. Because these correlations were dependent—that is, they had a variable in common, the criterion—they were compared using the Steiger method (Steiger, 1980). All p values were adjusted using the false discovery rate method (Benjamini & Hochberg, 1995). The PRA-CAT for public use was built using the free, open-source software Concerto (Scalise & Allen, 2015). Please see the online supplemental materials for further information.

Results and Discussion

Dimensionality and internal consistency of both PRA forms were acceptable for IRT, and fit of the subsequent one-parameter models was acceptable. Further information on internal consistency, dimensionality, and specifics of estimated item parameters is available in the online supplemental materials. To summarize, the above parameter estimates can be used to generate an information function showing how much information each test provides across ability levels. Figure 1 shows these information functions. Although the WRAT is more informative at lower ability levels, both PRA forms reveal more information about average ability—that is, −1 to 1, approximately 68% of the normal distribution—than does the WRAT. Further, the CAT simulations demonstrated that the PRA-CAT achieves low measurement error (0.25; equivalent to Cronbach's alpha = .94) and acceptable measurement error (0.40; Cronbach's alpha = .84) after only 18 and six items, respectively (on average). Below, we test validity of the more precise version (max standard error of measurement [*SEM*] = 0.25; average administration = 18 items).

Correlations between the PRA forms and the WRAT were 0.94 for both the "a" and "d" forms, providing strong evidence for the PRA's convergent validity from the outset. Further, scores on each test (WRAT, PRAa, PRAd, adaptive PRAa, and adaptive PRAd) were correlated with 21 validity criteria, including demographic, neurocognitive, clinical, and neuroimaging criteria. Table 1 shows these correlations, along with indicators of significant difference from the WRAT and from the corresponding PRA full form. The PRAa adaptive form was equally or significantly more correlated with 10 of the predictors than was the WRAT, and equally or more correlated with 20 of the predictors than was the full-form PRAa. The adaptive PRAd was equally or significantly more correlated with eight of the predictors than was the WRAT, and equally or more correlated with 20 of the predictors than was the nonadaptive PRAd. Notably, of the criteria significantly more correlated with the WRAT, the maximum absolute difference in correlations was 0.068 (WRAT vs. PRAd in predicting Black race). The vast majority of correlation differences are less than 0.05, suggesting the statistical significance is due to the large sample. Indeed, when Bonferroni correction is applied (critical $p = .0012$), the PRAa and PRAd no longer appear different than the WRAT for seven of the effects in Table 1: White race, clinical Fear score, and five of the neurocognitive scores.

This study aimed to analyze and establish the validity of the PRA as the first CAT reading measure available for use. Development of the PRA was based on the standard reading test paradigm—consisting of words with irregular grapheme to phoneme translations—that is utilized by many prevalent tests (e.g., WRAT, NART). Two forms of the PRA were administered alongside the WRAT to over 3,000 PNC participants as part of a larger data collection initiative. Factor analysis results indicated that both the WRAT and the PRA measure one factor, presumed to be reading

[1] The concept of psychometric information is beyond the present scope, but see Embretson and Reise (2009). Information is related to the standard error of measurement by the following equation: *SEM* = sqrt(1/info).

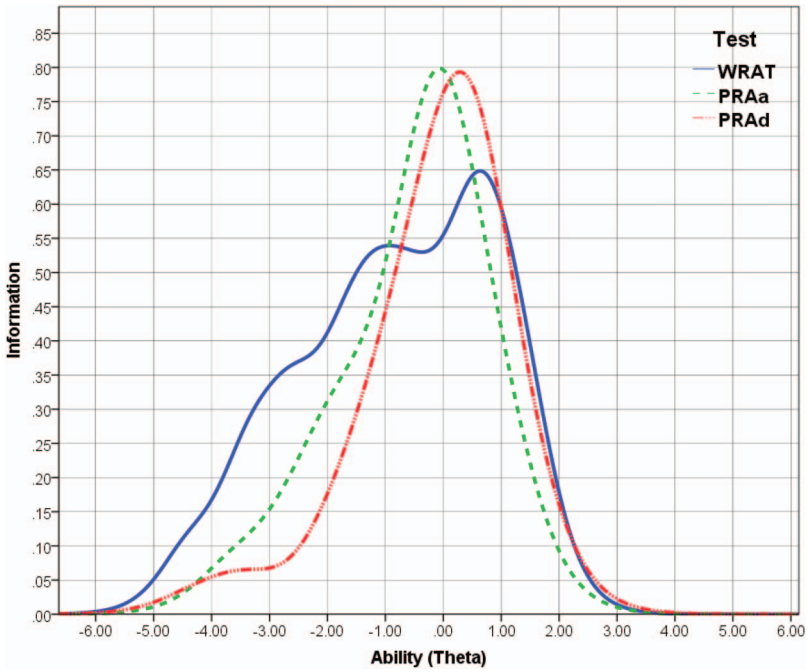

Figure 1. Test information curves for the WRAT, PRAa Full Form, and PRAd Full Form. WRAT = Wide Range Achievement Test; PRA = Penn Reading Assessment. See the online article for the color version of this figure.

ability, though there is some evidence of multidimensionality in the WRAT. Internal consistency data further suggest that the PRA is equally as reliable as the traditional WRAT. By simulating adaptive test sessions for the PRA, we found that the adaptive versions of the PRA are highly correlated with many of the neurocognitive, volumetric, environmental, and clinical factors tested and, for some key predictors, displayed even higher correlations than did the WRAT. Results of the present study demonstrate the validity of the PRA as an alternative reading test to the WRAT. Our results suggest that the PRA can serve as a valid and useful measure of reading ability as a proxy for premorbid IQ.

Table 1
Correlations Between Test Scores and Validity Criteria

Type	Criterion	WRAT4	PRAa	PRAd	PRAa Adaptive	PRAd Adaptive
Demographics	Age	.583	.669	.631	.643[ac]	.608[ac]
	Sex = female[b]	.068	.056	.133	.070	.121[c]
	Neighborhood SES	.320	.276	.289	.269[c]	.283[c]
	Neighborhood crime	−.092	−.148	−.026	−.158	−.029[c]
	Mother's education	.242	.205	.210	.209[c]	.203[c]
	Race = Black[b]	−.374	−.364	−.314	−.365	−.306[c]
	Race = White[b]	.357	.332	.324	.329[c]	.310[c]
Neurocognitive	Overall accuracy	.506	.481	.471	.484[c]	.464[c]
	Overall speed	.136	.115	.114	.113[c]	.120
	Overall efficiency	.427	.411	.388	.412	.391[c]
	Memory (Eff)	.145	.125	.150	.126	.158
	Social cog (Eff)	.343	.363	.309	.367[c]	.309[c]
	Comp reas (Eff)	.438	.412	.413	.409[c]	.412[c]
	Executive (Eff)	.377	.350	.320	.350[c]	.325[c]
Clinical	Global functioning	.131	.072	.089	.083[c]	.098[c]
	Anxious-misery	−.068	−.052	−.098	−.057	−.106[c]
	Psychosis	−.134	−.133	−.181	−.141	−.186[c]
	Externalizing	−.160	−.129	−.191	−.140	−.191[c]
	Fear	−.117	−.100	−.151	−.100	−.140[c]
Brain	Total gray matter	.286	.308	.236	.282	.223[c]
	Total brain volume	.277	.288	.231	.265	.219[c]

[a] significantly ($p < .05$) difference from corresponding PRA full-form correlation. [b] correlations are biserial. [c] significantly ($p < .05$) difference from WRAT4 correlation. SES = socioeconomic status; Eff = efficiency; cog = cognition; comp reas = complex reasoning.

Use of the adaptive version of the PRA will allow for more efficient data collection. The WRAT and NART have been criticized for underestimating the IQ of those with higher intelligence (Griffin et al., 2002; Johnstone et al., 1996; Wiens et al., 1993); our results indicate that the PRA will resolve this issue and can serve to more precisely estimate IQs in the upper range. Additionally, although the original PRAa and PRAd were administered in an average of only approximately 4 min, the CAT form will be completed even more efficiently, as each participant will be tested on fewer items. In continually updating the current estimate of a participant's ability based on previous responses and presenting an item selected on the basis of item information, the CAT version of the PRA will be more streamlined than the currently available measures. The highly precise CAT version presented here (max *SEM* = 0.25) takes (on average) 1.2 min to complete, yielding 70% time saving. The acceptably precise version (max *SEM* = 0.40; validity not tested here) takes (on average) 24 s to complete, yielding 90% time saving.

Strengths of this study include the large, diverse sample and the extensive phenotyping, including clinical and neurocognitive. However, some weaknesses should be noted. First, although standardization of the WRAT4 itself included slightly more than 3,000 participants (Wilkinson & Robertson, 2006)—a sample size comparable with ours—the population studied in standardization of the WRAT4 differed from the population studied here, in that WRAT standardization participants were selected specifically to be proportionate to the U.S. *national* census data. Thus, the WRAT standardization population was older, more geographically diverse, and more closely reflected nationwide socioeconomic/race/ethnicity proportions than did our sample selected from the greater Philadelphia ("tri-state") area. For example, our sample was approximately 33% African American, which is higher than the national average. A second weakness is that although we have shown a relationship between adaptive PRA scores and neurocognitive performance (including WRAT), we have not shown that the PRA measures *premorbid* IQ as such. That is, we have not shown that PRA performance is robust to neurological injury. Finally, it should be noted that the relationships of the adaptive PRA scores to validity criteria were generally lower than those of the WRAT; thus, although PRA scores are comparable with the WRAT, they are not quite as related to validity criteria. Nonetheless, the present psychometric analyses of the PRA provide some evidence for its reliability and validity in a community sample, suggesting that it could be used as an alternative to the currently available tests of reading ability. Future development efforts are focused on constructing a self-proctored version of the PRA in order to further extend its possible uses.

References

Barona, A., & Chastain, R. L. (1986). An improved estimate of premorbid IQ for blacks and whites on the WAIS-R. *International Journal of Clinical Neuropsychology, 8,* 167–173.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B: Methodological, 57,* 289–300. http://dx.doi.org/10.1111/j.2517-6161.1995.tb02031.x

Berkman, N. D., Dewalt, D. A., Pignone, M. P., Sheridan, S. L., Lohr, K. N., Lux, L., . . . Bonito, A. J. (2004). Literacy and health outcomes. *Evidence Report/Technology Assessment (Summary), 1,* 1–8.

Bland, J. M., & Altman, D. G. (1997). Cronbach's alpha. *British Medical Journal, 314,* 572. http://dx.doi.org/10.1136/bmj.314.7080.572

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6,* 431–444. http://dx.doi.org/10.1177/014662168200600405

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, 56,* 1356–1369. http://dx.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. http://dx.doi.org/10.1002/wps.20152

Camara, W. J., Nathan, J. S., & Puente, A. E. (2000). Psychological test usage: Implications in professional psychology. *Professional Psychology: Research and Practice, 31,* 141–154. http://dx.doi.org/10.1037/0735-7028.31.2.141

Casaletto, K. B., Cattie, J., Franklin, D. R., Moore, D. J., Woods, S. P., Grant, I., . . . HNRP Group. (2014). The Wide Range Achievement Test-4 Reading subtest "holds" in HIV-infected individuals. *Journal of Clinical and Experimental Neuropsychology, 36,* 992–1001. http://dx.doi.org/10.1080/13803395.2014.960370

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48,* 1–29. http://dx.doi.org/10.18637/jss.v048.i06

Choi, S. W. (2009). Firestar: Computerized adaptive testing simulation program for polytomous item response theory models. *Applied Psychological Measurement, 33,* 644–645. http://dx.doi.org/10.1177/0146621608329892

Clark, J. M., & Paivio, A. (2004). Extensions of the Paivio, Yuille, and Madigan (1968) norms. *Behavior Research Methods, Instruments & Computers, 36,* 371–383. http://dx.doi.org/10.3758/BF03195584

Contador, I., Bermejo-Pareja, F., Del Ser, T., & Benito-León, J. (2015). Effects of education and word reading on cognitive scores in a community-based sample of Spanish elders with diverse socioeconomic status. *Journal of Clinical and Experimental Neuropsychology, 37,* 92–101. http://dx.doi.org/10.1080/13803395.2014.989819

Dotson, V. M., Kitner-Triolo, M. H., Evans, M. K., & Zonderman, A. B. (2009). Effects of race and socioeconomic status on the relative influence of education and literacy on cognitive functioning. *Journal of the International Neuropsychological Society, 15,* 580–589. http://dx.doi.org/10.1017/S1355617709090821

Embretson, S. E., & Reise, S. P. (2009). *Item response theory*. Mahwah, NJ: Psychology Press.

Griffin, S. L., Mindt, M. R., Rankin, E. J., Ritchie, A. J., & Scott, J. G. (2002). Estimating premorbid intelligence: Comparison of traditional and contemporary methods across the intelligence continuum. *Archives of Clinical Neuropsychology, 17,* 497–507. http://dx.doi.org/10.1093/arclin/17.5.497

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. http://dx.doi.org/10.1016/j.jneumeth.2009.11.017

Jastak, J., & Jastak, S. (1965). *The Wide Range Achievement Test manual*. Wilmington, DE: Guidance Associates.

Jastak, S., & Wilkinson, G. S. (1984). *Wide Range Achievement Test—Revised*. Wilmington, DE: Jastak Associates.

Johnstone, B., Callahan, C. D., Kapila, C. J., & Bouman, D. E. (1996). The comparability of the WRAT-R reading test and NAART as estimates of premorbid intelligence in neurologically impaired patients. *Archives of*

Clinical Neuropsychology, 11, 513–519. http://dx.doi.org/10.1093/arclin/11.6.513

Kareken, D. A., Gur, R. C., & Saykin, A. J. (1995). Reading on the Wide Range Achievement Test-Revised and parental education as predictors of IQ: Comparison with the Barona formula. *Archives of Clinical Neuropsychology, 10,* 147–157. http://dx.doi.org/10.1093/arclin/10.2.147

Kaufman, A. S., & Kaufman, N. L. (2004). *Kaufman Test of Educational Achievement–Comprehensive form*. Circle Pines, MN: American Guidance Service.

Krull, K., Scott, J., & Sherer, M. (1995). Estimation of premorbid intelligence from combined performance and demographic variables. *Clinical Neuropsychologist, 9,* 83–88. http://dx.doi.org/10.1080/13854049508402063

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika, 71,* 713–732. http://dx.doi.org/10.1007/s11336-005-1295-9

Moore, T. M., Calkins, M. E., Reise, S. P., Gur, R. C., & Gur, R. E. (2018). Development and public release of a computerized adaptive (CAT) version of the Schizotypal Personality Questionnaire. *Psychiatry Research, 263,* 250–256. http://dx.doi.org/10.1016/j.psychres.2018.02.022

Moore, T. M., Martin, I. K., Gur, O. M., Jackson, C. T., Scott, J. C., Calkins, M. E., . . . Gur, R. C. (2016). Characterizing social environment's association with neurocognition using census and crime data linked to the Philadelphia Neurodevelopmental Cohort. *Psychological Medicine, 46,* 599–610. http://dx.doi.org/10.1017/S0033291715002111

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. http://dx.doi.org/10.1037/neu0000093

Moore, T. M., Scott, J. C., Reise, S. P., Port, A. M., Jackson, C. T., Ruparel, K., . . . Gur, R. C. (2015). Development of an abbreviated form of the Penn Line Orientation Test using large samples and computerized adaptive test simulation. *Psychological Assessment, 27,* 955–964. http://dx.doi.org/10.1037/pas0000102

Nelson, H. E., & Willison, J. (1991). *National Adult Reading Test (NART)*. Windsor, UK: NFER-Nelson.

Paivio, A., Yuille, J. C., & Madigan, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology, 76,* 1–25.

Psychological Corporation. (2002). *Wechsler Individual Achievement Test* (2nd ed.). San Antonio, TX: Author.

Rabin, L. A., Barr, W. B., & Burton, L. A. (2005). Assessment practices of clinical neuropsychologists in the United States and Canada: A survey of INS, NAN, and APA Division 40 members. *Archives of Clinical Neuropsychology, 20,* 33–65. http://dx.doi.org/10.1016/j.acn.2004.02.005

Ramsden, S., Richardson, F. M., Josse, G., Shakeshaft, C., Seghier, M. L., & Price, C. J. (2013). The influence of reading ability on subsequent changes in verbal IQ in the teenage years. *Developmental Cognitive Neuroscience, 6,* 30–39. http://dx.doi.org/10.1016/j.dcn.2013.06.001

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Reise, S. P., Cook, K. F., & Moore, T. M. (2015). Evaluating the impact of multidimensionality on unidimensional item response theory model parameters. In S. P. Reise & D. Revicki (Eds.), *Handbook of item response theory modeling* (pp. 13–40). New York, NY: Routledge.

Revelle, W. (2018). *Psych: Procedures for personality and psychological research*. Evanston, IL: Northwestern University. Retrieved from https://CRAN.R-project.org/package=psych

Roalf, D. R., Moore, T. M., Wolk, D. A., Arnold, S. E., Mechanic-Hamilton, D., Rick, J., . . . Moberg, P. J. (2016). Defining and validating a short form Montreal Cognitive Assessment (s-MoCA) for use in neurodegenerative disease. *Journal of Neurology, Neurosurgery & Psychiatry, 87,* 1303–1310. http://dx.doi.org/10.1136/jnnp-2015-312723

Satterthwaite, T. D., Connolly, J. J., Ruparel, K., Calkins, M. E., Jackson, C., Elliott, M. A., . . . Hakonarson, H. (2016). The Philadelphia Neurodevelopmental Cohort: A publicly available resource for the study of normal and abnormal brain development in youth. *NeuroImage, 124,* 1115–1119. http://dx.doi.org/10.1016/j.neuroimage.2015.03.056

Satterthwaite, T. D., Elliott, M. A., Ruparel, K., Loughead, J., Prabhakaran, K., Calkins, M. E., . . . Gur, R. E. (2014). Neuroimaging of the Philadelphia Neurodevelopmental Cohort. *NeuroImage, 86,* 544–553. http://dx.doi.org/10.1016/j.neuroimage.2013.07.064

Sayegh, P., Arentoft, A., Thaler, N. S., Dean, A. C., & Thames, A. D. (2014). Quality of education predicts performance on the Wide Range Achievement Test-4th Edition Word Reading subtest. *Archives of Clinical Neuropsychology, 29,* 731–736. http://dx.doi.org/10.1093/arclin/acu059

Scalise, K., & Allen, D. D. (2015). Use of open-source software for adaptive measurement: Concerto as an R-based computer adaptive development and delivery platform. *British Journal of Mathematical and Statistical Psychology, 68,* 478–496. http://dx.doi.org/10.1111/bmsp.12057

Seidman, L. J., Shapiro, D. I., Stone, W. S., Woodberry, K. A., Ronzio, A., Cornblatt, B. A., . . . Woods, S. W. (2016). Association of neurocognition with transition to psychosis: Baseline functioning in the second phase of the North American Prodrome Longitudinal Study. *Journal of the American Medical Association Psychiatry, 73,* 1239–1248. http://dx.doi.org/10.1001/jamapsychiatry.2016.2479

Slocum-Gori, S., & Zumbo, B. D. (2011). Assessing the unidimensionality of psychological scales: Using multiple criteria from factor analysis. *Social Indicators Research, 102,* 443–461. http://dx.doi.org/10.1007/s11205-010-9682-8

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87,* 245–251. http://dx.doi.org/10.1037/0033-2909.87.2.245

Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment, 80,* 99–103. http://dx.doi.org/10.1207/S15327752JPA8001_18

van der Linden, W. J. (1998). Bayesian item-selection criteria for adaptive testing. *Psychometrika, 63,* 201–216. http://dx.doi.org/10.1007/BF02294775

Vanderploeg, R. D., Schinka, J. A., & Axelrod, B. N. (1996). Estimation of WAIS-R premorbid intelligence: Current ability and demographic data used in a best-performance fashion. *Psychological Assessment, 8,* 404–411. http://dx.doi.org/10.1037/1040-3590.8.4.404

Wainer, H., & Dorans, N. J. (2000). *Computerized adaptive testing* (2nd ed.). Mahwah, NJ: Routledge.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21,* 361–375. http://dx.doi.org/10.1111/j.1745-3984.1984.tb01040.x

Wiens, A. N., Bryan, J. E., & Crossen, J. R. (1993). Estimating WAIS-R FSIQ from the National Adult Reading Test-Revised in normal subjects. *Clinical Neuropsychologist, 7,* 70–84. http://dx.doi.org/10.1080/13854049308401889

Wilkinson, G. S., & Robertson, G. J. (2006). *Wide Range Achievement Test–Fourth edition: Professional manual*. Lutz, FL: Psychological Assessment Resources.

Wilkinson, G. S., & Robertson, G. J. (2017). *Wide Range Achievement Test* (5th ed.). Minneapolis, MN: Pearson.

Received September 27, 2018
Revision received April 29, 2019
Accepted May 1, 2019 ■