

















Research Article

Computerized adaptive test strategies for the matrix reasoning subtest of the Wechsler Adult Intelligence Scale, 4th Edition (WAIS-IV)

Steven P. Reise¹ , Emily Wong¹ , Jared Block¹ , Keith F. Widaman² , Joseph M. Gullett³ , Russell M. Bauer³ , Daniel L. Drane⁴ , David W. Loring⁴ , Laura Glass Umfleet⁵ , Dustin Wahlstrom⁶ , Kristen Enriquez⁷ , Fiona Whelan⁷ , Stone Shih⁷  and Robert M. Bilder^{7,1} 

¹Department of Psychology, College of Letters & Science, UCLA, Los Angeles, CA, USA, ²University of California, Riverside, Riverside, CA, USA, ³University of Florida, Gainesville, FL, USA, ⁴Departments of Neurology and Pediatrics, Emory University School of Medicine, Atlanta, GA, USA, ⁵Medical College of Wisconsin, Milwaukee, WI, USA, ⁶Pearson Clinical Assessment, San Antonio, TX, USA and ⁷Department of Psychiatry & Biobehavioral Sciences, UCLA David Geffen School of Medicine, and Jane & Terry Semel Institute for Neuroscience and Human Behavior, Los Angeles, CA, USA

Abstract

Objective: Most neuropsychological tests were developed without the benefit of modern psychometric theory. We used item response theory (IRT) methods to determine whether a widely used test – the 26-item Matrix Reasoning subtest of the WAIS-IV – might be used more efficiently if it were administered using computerized adaptive testing (CAT). **Method:** Data on the Matrix Reasoning subtest from 2197 participants enrolled in the National Neuropsychology Network (NNN) were analyzed using a two-parameter logistic (2PL) IRT model. Simulated CAT results were generated to examine optimal short forms using fixed-length CATs of 3, 6, and 12 items and scores were compared to the original full subtest score. CAT models further explored how many items were needed to achieve a selected precision of measurement (standard error $\leq .40$). **Results:** The fixed-length CATs of 3, 6, and 12 items correlated well with full-length test results (with $r = .90, .97$ and $.99$, respectively). To achieve a standard error of $.40$ (approximate reliability = $.84$) only 3–7 items had to be administered for a large percentage of individuals. **Conclusions:** This proof-of-concept investigation suggests that the widely used Matrix Reasoning subtest of the WAIS-IV might be shortened by more than 70% in most examinees while maintaining acceptable measurement precision. If similar savings could be realized in other tests, the accessibility of neuropsychological assessment might be markedly enhanced, and more efficient time use could lead to broader subdomain assessment.

Keywords: Wechsler Intelligence Scales; Psychometrics; Wechsler Memory Scale; Intelligence tests; Aptitude tests; Psychological tests

(Received 25 October 2022; final revision 16 May 2023; accepted 13 June 2023)

Introduction

Neuropsychological assessment methods would benefit from modernization. Many tests have origins in the 19th century and most elements of modern batteries were in place by the end of World War II (Bilder, 2011). Despite widespread advocacy for the “flexible” approach to neuropsychological assessment, the most widely used methods center on tests that include the Wechsler Adult Intelligence Scale and selected other tests that are considered measures of specific neurocognitive domains (Rabin et al., 2005; Rabin et al., 2016). Most tests used today were developed based on the intuitions and ingenuity of the test developers. In contrast, modern psychometric approaches use *a priori* definitions of constructs and careful psychometric evaluation of how those constructs are measured.

Test refinements over recent decades have led to better understanding of many tests using classical test theory methods. Test publishers have generally improved their standardization

practices through better sampling methods and improved test refinement practices. This work has provided greater insight into the factor structure of widely used tests and better definition of certain test properties (internal consistency, test-retest, and alternate-forms reliability). A few measures have further demonstrated external validity relative to selected demographic variables (e.g., age, education, race, and ethnicity), and by showing differences associated with specific diagnostic groups or treatment outcomes (Holdnack et al., 2011; Wechsler, 2008a, 2008b).

In contrast, modern psychometric theory has been used infrequently in the construction and evaluation of neuropsychological tests, with several noteworthy exceptions (Bilder & Reise, 2019; Crane et al., 2008; Gershon et al., 2014; Moore et al., 2015; Mungas et al., 2003; Mungas et al., 2000; Yudien et al., 2019). Item response theory (IRT) offers the potential to better define and measure latent traits identified in many neuropsychological tests. IRT can further specify the precision of measurement at different levels of the construct, which is often particularly important in the

Corresponding author: Steven P. Reise; Email: reise@psych.ucla.edu

Cite this article: Reise S.P., Wong E., Block J., Widaman K.F., Gullett J.M., Bauer R.M., Drane D.L., Loring D.W., Umfleet L.G., Wahlstrom D., Enriquez K., Whelan F., Shih S., & Bilder R.M. Computerized adaptive test strategies for the matrix reasoning subtest of the Wechsler Adult Intelligence Scale, 4th Edition (WAIS-IV). *Journal of the International Neuropsychological Society*, 1–10, <https://doi.org/10.1017/S1355617723000401>

assessment of individuals with very high or very low levels of ability, and critical to assure that measurement properties are comparable within individuals when change has taken place due to disease or interventions. These methods are further important to assure that our tests measure the same constructs across groups that differ by sex, race, ethnicity, cultural backgrounds, and clinical conditions (Bilder *et al.*, 2022).

IRT can further markedly reduce test length by identifying the information contributed by each item to the measurement of the latent trait(s), and optimizing the combined information generated by subsets of test items administered in an adaptive format. Specifically, computerized adaptive testing (CAT) based on item response theory (IRT) has been extensively researched, and implementations are common around the world. The International Association for Computerized Adaptive Testing lists 39 large scale CAT programs (<https://www.iacat.org/content/operational-cat-programs>). These include the Armed Services Vocational Aptitude Test Battery and the Graduate Record Examination, which are used in high-stakes placement and admissions decisions. A clinical example is the Patient Reported Outcomes Measurement Information System (PROMIS) that enables efficient, fixed precision assessment of depression, anxiety, self-reported cognitive ability/dysfunction, and more (<https://www.healthmeasures.net/resource-center/measurement-science/computer-adaptive-tests-cats>). Unfortunately, CAT research has seldom been applied to neuropsychological assessment, with a few exceptions (Moore *et al.*, 2015, 2023; Yudien *et al.*, 2019).

Adaptive testing starts by administering an item of intermediate difficulty. If the examinee passes that item, they next get an item that is more difficult, but if they fail the item, they will next get an easier item. This process continues, adjusting the difficulty of the next-selected item until a “good” estimate (with predefined precision) of the examinee’s ability level is obtained. Computerized adaptive testing (CAT) (Choi, 2009, 2020) often reduces test length by 50% or more, resulting in markedly increased efficiency of assessment, which in turn can reduce assessment time and cost, and thereby increase access to additional care (Bilder & Reise, 2019; Gibbons *et al.*, 2008; Reeve *et al.*, 2007; Reise & Waller, 2009). Most current practice models relying on traditional testing have led neuropsychological assessment to be among the most time-consuming of medical diagnostic procedures, with associated high costs that differentially impact individuals with fewer financial resources, and often lead to long waiting lists so that care, even when provided, is often delayed. CAT has an additional advantage: because item selection and scoring are typically done automatically, additional time and cost savings accrue relative to traditional assessment that depends on manual scoring by highly trained individuals.

The National Neuropsychology Network (NNN) was created to enable assessment of widely used neuropsychological tests using IRT and other modern psychometric methods, and to facilitate development of more efficient methods to measure latent traits (Loring *et al.*, 2021). The NNN enables these kinds of analyses because all data are being acquired at the item level, in contrast to databases that comprise only summary scores.

As a proof-of-concept demonstration, we examined the Matrix Reasoning (MR) subtest of the Wechsler Adult Intelligence Scale, 4th Edition (WAIS-IV; Wechsler, 2008). The MR subtest was introduced in the WAIS-III (Wechsler, 1997), but historical roots of this test may be traced to the much earlier development of the Raven Progressive Matrices (Penrose & Raven, 1936). Factor analytic work has shown that the Matrix Reasoning subtest loads

together with other measures (Visual Puzzles, Block Design) to form the WAIS-IV Perceptual Reasoning Index (Wechsler *et al.*, 2008; Wechsler, 2008a). We recently used confirmatory factor analysis to demonstrate that the WAIS-IV factor structure shows strong measurement invariance across a heterogeneous patient group from the NNN and the original healthy standardization sample (Bilder *et al.*, 2022).

The primary goal of our analyses was to demonstrate possible savings in administration time of the Matrix Reasoning subtest that could be gained using IRT-based CAT item-selection strategies, under the assumption of unidimensionality. Preliminary analyses were also necessary to determine if the assumption of unidimensionality was justified in our sample.

Method

Participants

Inclusion/exclusion

Because this project involved care-as-usual, no *a priori* restrictions on inclusion of participants were used, except that the study included only adults (ages 18 years or older) and only those whose primary language was English. Data were collected from 2197 patients who were administered the Matrix Reasoning subtest of the WAIS-IV as part of a routine clinical neuropsychological evaluation. Patients were between the ages of 18 and 90 years ($M = 51.68$, $SD = 18.15$), about half of whom indicated that their biological sex assigned at birth was Female ($N = 1163$). A small minority identified as Hispanic ($N = 45$), while the vast majority identified as not Hispanic ($N = 2059$); ethnicity was unknown for the remaining 93 individuals.

Demographic and clinical variables

We recorded age, educational attainment, sex, race and ethnicity following protocols developed by the National Human Genome Research Institute’s “PhenX” (phenotypes and genotypes) project (McCarty *et al.*, 2014) that were endorsed by the NIMH as Common Data Elements for demographic variables (Barch *et al.*, 2016). Complete data dictionaries for the NNN database are available online at www.nnn.ucla.edu. This study was not preregistered.

Human subjects

All procedures were conducted with approval from the Institutional Review Boards at each site, using reliance agreements implemented by SmartIRB. Initially we obtained informed consent (for the first 2138 cases), and excluded participants if there were concerns about capacity to provide informed consent. Subsequently we received a waiver of informed consent so all clinic patients could be included. For participants older than 89, we coded age as “90+”. The UCLA IRB was the IRB of record. Participants were identified by Global Unique Identifiers (GUIDs) or pseudo-GUIDs, as defined by the NIMH. Some participants had multiple neuropsychological evaluations during their clinical care; in these cases, results of the first examination only were included for each examinee. An “examination” was operationally defined as a set of tests administered within a period of 30 days, intended to represent a single episode of care.

Data sources and measures

All clinics administered the Matrix Reasoning subscale following standard administration and scoring methods set forth in the

Table 1. Matrix reasoning classical test theory statistics and item response theory parameter estimates

	Mean	s	raw.r	r.drop	IRT Slope	IRT Location
					α	β
V1	.99	.08				
V2	.99	.09				
V3	.99	.10				
V4	.99	.11				
V5	.95	.21				
V6	.90	.30	.43	.39	1.59	-2.00
V7	.87	.33	.41	.36	1.23	-1.99
V8	.78	.42	.57	.52	1.68	-1.08
V9	.77	.42	.60	.55	1.86	-1.01
V10	.68	.47	.70	.66	2.41	-.49
V11	.71	.45	.62	.56	1.79	-.71
V12	.64	.48	.68	.63	2.13	-.38
V13	.64	.48	.74	.70	2.88	-.31
V14	.61	.49	.69	.64	2.23	-.25
V15	.52	.50	.68	.63	2.23	.04
V16	.52	.50	.74	.70	3.01	.09
V17	.41	.49	.67	.62	2.56	.40
V18	.46	.50	.74	.70	3.54	.25
V19	.33	.47	.62	.56	2.48	.62
V20	.36	.48	.67	.62	3.12	.52
V21	.32	.47	.66	.60	3.41	.61
V22	.35	.48	.65	.60	2.94	.54
V23	.21	.41	.55	.49	3.07	.95
V24	.18	.39	.52	.47	3.13	1.03
V25	.11	.31	.43	.38	3.58	1.27
V26	.08	.26	.36	.32	3.14	1.49

Note. IRT slope (α in the 2PL equation) is a discrimination parameter; IRT Location (β in the 2PL equation) is a difficulty parameter; s is standard deviation; raw.r is item to test score correlation; r.drop is item to test score correlation if item dropped.

manual (Wechsler, 2008). The Matrix Reasoning subtest includes two sample items (A and B), and 26 test items. Item 4 is administered first unless intellectual disability is suspected, in which case administration begins with item 1. If examinees do not obtain perfect scores on either items 4 or 5, then preceding items are administered in reverse order until the examinee obtains perfect scores on two consecutive items. Following the administration manual, items were otherwise administered in order of difficulty (easiest to hardest). For each item, individuals received a score of either 1 (correct) or 0 (incorrect). After three consecutive scores of 0, the test was discontinued. Missing values after the test was discontinued following three failures were later coded as 0's (i.e., incorrect) for purposes of CAT simulation.

Dimensionality assessment

The Matrix Reasoning subtest is universally scored as a single construct in practice. Nevertheless, the IRT model to be applied here assumes "essential" unidimensionality, that is, the responses primarily reflect a single common dimension. Thus, it is important, prior to estimating an IRT model, to first empirically establish the viability of a unidimensional model in our sample. To evaluate unidimensionality in the present data, we first estimated tetrachoric correlations among the items (deleting the first five items due to very high proportion corrects, see Table 1) and then factor analyzed the tetrachoric correlations using *minres* extraction. To establish essential unidimensionality, we examined the ratio of the first to the second eigenvalue, the magnitude of factor loadings, and several indices of statistical fit. Specifically, we used the *lavaan* library (Rosseel, 2012) to fit a confirmatory unidimensional model specifying the items as ordinal and using diagonally weighted least squares estimation. We examined model chi-square,

the scaled comparative fit index, scaled root mean squared error of approximation, and standardized root mean square residual. By traditional conventions, values of these indices $>.90$, $<.05$, and $<.08$ would be considered "good".

Item response theory (IRT)

Compared to classical test theory (CTT) which focuses on test-level functioning, IRT focuses on item-level functioning. The chief goal of IRT is to fit a statistical model, called an item response function (IRF), which describes how the probability of responding correctly to an item changes as a function of ability or "trait" level (generally denoted as θ) and properties of an item (e.g., its difficulty and discrimination). It is assumed that the probability of responding correctly monotonically increases as a function of θ . This interpretation assumes that any variation in item response is driven by one dominant dimension (i.e., factor), which embodies the unidimensionality assumption.

As noted, the IRF describes the probability of responding correctly given ability or "trait level" (θ), and it is defined by several parameters (e.g., item difficulty, guessing rate, discrimination). For this application we selected a model that is analogous to the well-known and extensively used item-level factor analytic model. Specifically, we selected the so-called two-parameter logistic model (2PL)¹ as shown in Equation 1.

$$P(x = 1|\theta) = \frac{1}{1 + \exp(-\alpha(\theta - \beta))} \quad (1)$$

In the above, examinee individual differences in trait level are symbolized by θ , which in this study (and most IRT studies) is assumed to be like a Z-score with a mean θ of zero and standard deviation of 1.0. Test items vary in location (β), which is the point along the trait continuum at which the probability of responding correctly is 50%. Thus, the location parameter has a scale such that positive values indicate a more difficult item, that is, an item that would require higher than average trait level to get correct. A negative difficulty item indicates an "easier" item, meaning that even individuals below the population mean may have a high chance to get it correct. The so-called "discrimination" parameter (α) controls the slope of the IRF at its reflection point – higher values indicate more discriminating items; more "discriminating" means that the item is better able to distinguish between individuals in the trait range around the item location.

In IRT, items can vary not only in discrimination, but also in how much statistical information they provide in discriminating among individuals. Specifically, once an IRF is estimated for each item, it can be easily transformed into an item information function (IIF) as shown in Equation 2.

$$\text{Info}|\theta = \alpha^2 P(x = 1|\theta)(1 - P(x = 1|\theta)) \quad (2)$$

An IIF describes how well an item can discriminate between individuals at different levels of ability. Items with higher discrimination (α) provide more information, but where that information is concentrated is determined by the location parameter (i.e., where the probability of a correct response is 50%).

¹Many additional psychometric analyses, such as evaluation of statistical and graphical item fit for alternative IRT models (e.g., a 3-parameter model) were conducted. However, discussion of these are beyond the scope and are not detailed here. They are available upon request.

Finally, it is critical to note that IIFs are additive across items and thus can be summed to form an overall scale information function (SIF). The SIF is key because it allows us to study how the standard error of measurement changes as a function of trait level for a given set of items administered, whether the full, complete battery or only a few CAT items are administered. Specifically, Equation 3 shows how scale information is converted to a conditional standard error of measurement. This is critical because information, and thus the standard error (SE), is leveraged in CAT to more efficiently select items for administration.

$$SE|\theta = \frac{1}{\sqrt{\text{INFO}|\theta}} \quad (3)$$

Our first analyses centered on performing traditional classical test theory psychometric analyses on the data. This included item-test correlations, item means and standard deviations. These values were obtained using the R library *psych* (Revelle, 2023), *alpha* command. The IRT 2PL model parameters were estimated using the *mirt* library in R (Chalmers, 2012). Although there are no definitive approaches to evaluating model fit in IRT, we examined root mean square error of approximation, standardized root mean square residual, and comparative fit index which are output from *mirt* (Maydeu-Olivares et al., 2011). These are simply IRT versions of the fit indices shown previously and similar “benchmarks” would apply.

Computerized adaptive testing

As noted, CAT (Wainer et al., 2000; van der Linden & Glas, 2000) has been investigated carefully in many assessment domains, including psychopathology assessment (see Gibbons et al., 2016) but has received scant attention in neuropsychological assessment. In CAT, an item is administered, usually of medium location parameter (e.g., around zero) and a response is collected and scored. Based on that response, a trait level estimate is made, and a new item is selected that is typically easier (if the examinee got the item wrong) or harder (if the examinee got the item right). More technically, the next item selected to administer is the one that maximizes the psychometric information (i.e., provided the most discrimination) at the current trait level estimate. This process of administering items, updating the latent trait estimate, and selecting new items to administer, continues until a termination criterion is met.

There are two major termination criteria used. First, items are adaptively administered until a fixed number, for example, 10, are administered. That is called fixed length adaptive testing. Second, items are administered until a standard error criterion is met. For example, items are administered until the standard error is at or below .30 (which would correspond roughly to an alpha reliability of .90). Of course, the procedure used is limited by the item bank one has on hand. If the test does not have enough items to provide sufficient psychometric information to reduce the standard error below a threshold, then a more liberal threshold is required.

In this research, for demonstration purposes, we conducted a real data simulation. We began again by filling in all cells of the data matrix after each person’s termination criterion was met with zeros to eliminate missing data. Therefore, the number correct for an individual is the number correctly answered until the stopping criterion was reached, and the number wrong is the sum of items missed prior to the stopping criterion plus all items after the stopping criterion. This demonstration is hypothetical and a “proof

of concept,” but this is true of most real data simulations of CAT (Thompson & Weiss, 2011).

The specific CAT algorithms evaluated here were as follows. We began by selecting the most informative item (most discriminating item) at trait level = 0. This item was always Item #13 which had discrimination of 2.88 and location of -0.31. Each examinee began with a trait level estimate of 0 and based on the response to the first item, trait level estimate and standard error were updated using the expected a posteriori (EAP) method of scoring (Bock & Mislevy, 1982). The next item selected was the one that provided the most psychometric information at the current trait level estimate.

The first set of algorithms we examined were fixed test length. Specifically, we limited the CAT administration to 3, 6, and 12 items respectively. We then examined a standard error-based CAT. Specifically, we continued administering items in CAT format until an examinee’s trait level estimates had a standard error below .40 (16% error variance or reliability of roughly .84). The three key outcome statistics are (1) the average standard error of measurement for each CAT condition, (2) the root mean square deviation between CAT and full scale trait level estimates (RMSD), and (3) the correlations, both Pearson and Spearman, between CAT trait level estimates and full scale estimates (i.e., based on all the items). The first and second indices provide degree of uncertainty around the true score and full scale score, respectively. The CAT versus full scale correlations are part-whole correlations and must be positive. They should be interpreted as descriptive statistics reflecting the specific simulation and not as estimates of a “population” parameter. High correlations (e.g., >.90) imply that CAT scores would have very similar external correlates as the full scale scores.

Results

In the current analysis, patients were originally administered between 1 and 26 test items. On average, participants received around 17 items (Mean = 17.30, Median = 19, Mode = 24). A frequency graph of the number of items administered is shown in Supplemental Figure 1. In Supplemental Figure 2 is displayed a graph of the time, in minutes, to complete a given number of items. The line is a linear regression, and the curves are locally estimated scatterplot smoothing (LOESS) plots. In addition, Supplementary Table 1 shows the mean and median times to complete each number of items administered using no trimming and 5% trimming to reduce the effect of outliers. These graphs and tables indicate that people taking between 20 and 23 items tend to spend about 8–10 min on the MR substest.

Dimensionality assessment

The first five eigenvalues of the tetrachoric matrix were 13.09, 1.30, 0.71, 0.64, and 0.56, and thus $13.09/21 = 62\%$ of the item variance was explained by the first factor. The ratio of the 1st to 2nd eigenvalue was 10.07, much larger than the frequently noted benchmark for “unidimensionality” of 3. This pattern suggests a very strong common dimension as expected. The results of fitting a unidimensional confirmatory model using *lavaan* (Rosseel, 2012) produced a standard chi-square of 583.26 (robust 769.73) on 189 *df*, and scaled comparative fit index = .996, scaled root mean squared error of approximation = .037, and standardized root mean square residual = .052, all indicating an acceptable, if not excellent, fit.

Classical and IRT model fitting results

Basic classical test theory psychometric values are shown in the first four columns of Table 1. Using these results, we decided to eliminate the first five items from subsequent analyses; these items are nearly universally passed (too high correct response rate) to provide any discrimination among individuals. The first set of columns contain the item-test total score correlations, and the item-test correlation if the item dropped from the total score. These values tend to be higher for items around #13 to #22 – items that are completed correctly by 35–65% of the participants and the item variance is largest. The item means in the third column indicate that the items are well ordered by proportion correct (easy to difficult) as would be expected given how the test is administered and our method of treating missing data. Thus, our easiest item is Item #6 with a .90 correct rate, and the hardest is Item #26 with a .08 correct rate. Coefficient alpha was .92 for both the 26-item and 21-item versions.

Estimated item parameters from the 2PL model are shown in the last two columns of Table 1 and the corresponding item information curves are shown in Figure 1. The fit statistics output by the *mirt* program for this model were root mean square error of approximation = 0.04, standardized root mean square residual = .07, and comparative fit index = .99, agreeing almost exactly with the CFI results shown previously and revealing excellent model fit (Maydeu-Olivares et al., 2011).

Finally, one of the major features of IRT modeling is the graphical display of item and test functioning. First, in Figure 2 we display trait level (on a scale with $M = 0$, $SD = 1.0$) versus estimated reliability conditional on trait level (Note: the concept of “test score” reliability does not exactly apply to IRT, and thus this is a rough approximation based on error variance of trait level estimates). This graph shows that for 1.5 standard deviation below the mean to 2 standard deviations above the mean, the reliability of trait scores is above .80, a commonly used “benchmark”. Most importantly, in Figure 3 scale information and conditional standard errors are shown for this 21-item version of Matrix Reasoning under the 2PL model. This figure tells the same story, but now in purely IRT terms; from around -1.5 to $+2$ standard deviations from the mean, measurement precision is relatively high, but is much lower at the extremes.

Computerized adaptive testing

For the fixed-length simulations for 3, 6, and 12 items, the Pearson correlation between CAT trait level estimates and full-length trait level estimates were .90, .97, and .99, respectively, and Spearman correlations were .91, .96, and .99. The average standard errors of measurement were .47, .35, and .29, respectively (.22, .12, and .08 error variance)². These can be compared against the average standard error of .27 for the full-length test (.07 error variance). In these conditions, all individuals received the same number of items, but the specific items administered depended on their response pattern. Graphs of full length trait level estimates versus CAT trait level estimates for these three conditions are shown in Supplementary Figures 3 through 5, respectively. Clearly, as the number of items increases, there are: (1) an increasing range of CAT scores, (2) an increase in the number of different CAT scores (possible response patterns and corresponding trait level estimates equals 2 raised to the power of the number of items), and (3) lower

²These standard errors correspond to marginal “reliability” values of .78, .88, and .92, respectively, compared with .93 for the full length test.

spread of full scale scores conditional on CAT scores. Finally, RMSD were .42, .22, and .08, indicating that, as the number of CAT items increases, the typical difference between CAT and full-scale trait level estimates gets smaller.

In the “standard error less than or equal to .40” CAT condition (16% error variance or reliability of roughly .84), the mean number of items administered was 6.72 ($S = 5.70$), the average standard error of measurement was .38, the RMSD was .25 and the Pearson correlation between CAT trait level estimates and full-length trait level estimates was $r = .97$ (Figure 4); the Spearman was .94. However, that CAT algorithm did not work for all individuals as described below.

In Figure 4, the trait level scores are color coded by the number of items received; these are grouped into 3–4, 5–13, and 21 (no one received 14–20 items). The above reported correlations are inflated, and RMSD deflated, due to the CAT algorithm not converging for individuals with very high or very low estimated trait levels, for whom all items needed to be administered. Supplemental Table 2 shows the number of people (and proportion of people) receiving specific numbers of items. The 283 individuals who required all 21 items (about 13%) had very high or low trait level estimates – locations on the trait where it was impossible to achieve a standard error of .40 or below. This is shown in Supplemental Figure 6. Because the same items were administered, the CAT trait level estimate and full test trait level estimate are the same for these individuals. This is a limitation of the item pool in the original test; specifically at very low and very high levels of performance, the original test has a standard error greater than .40. This can be seen in the Scale Information Function plot in Figure 3, which as in most tests, shows substantial increases in error at the lower and upper ends of the trait. We thus estimated indices again, but with the 283 individuals who received all items eliminated. The new values are: average standard error = .37, RMSD = .26, Pearson = .93 and Spearman = .90.

Discussion

This proof-of-concept investigation demonstrated that one of the most widely administered neuropsychological tests – the Matrix Reasoning subtest of the WAIS-IV – might be administered with markedly greater efficiency using a computerized adaptive test (CAT) method, based on item response theory (IRT). The CAT models demonstrated that tests with 3 to 12 items correlated well ($r = .90$ to $.99$) with full-length subtest scores based on 21 items. When using a fixed-precision strategy, an average of 6.7 items was needed to yield a standard error of less than .4 (which indicates reliability of approximately .84), yielding strong correlations ($r = .929$ and $.902$ when excluding people who received all items) between the CAT estimate of ability and the estimate based on the full-length test. Many participants (1431/2197 or about 65%) required only 3 or 4 items to reach this level of precision.

For the Matrix Reasoning subtest alone, the results suggest the total administration time could be reduced from 8–10 min³ to 2–3 min on average. Given estimates for total WAIS-IV administration time of 60–90 min, these findings suggest that the entire WAIS-IV might be completed in 20–45 min following an adaptive testing format. This is based on the fact that 8 of 10 core

³Some estimates of time to complete the MR subtest are slightly shorter, but available estimates are for different versions in different samples. For example, Ryan, Glass and Brown, 2007 reported for WISC-IV that the Matrix Reasoning subtest took on average 6:06 (SD 2:32) with a range from 2:18 to 16:15. Axelrod 2001 found the WAIS-III Matrix Reasoning subtest took on average 5.3 (SD 3.8) minutes to administer.

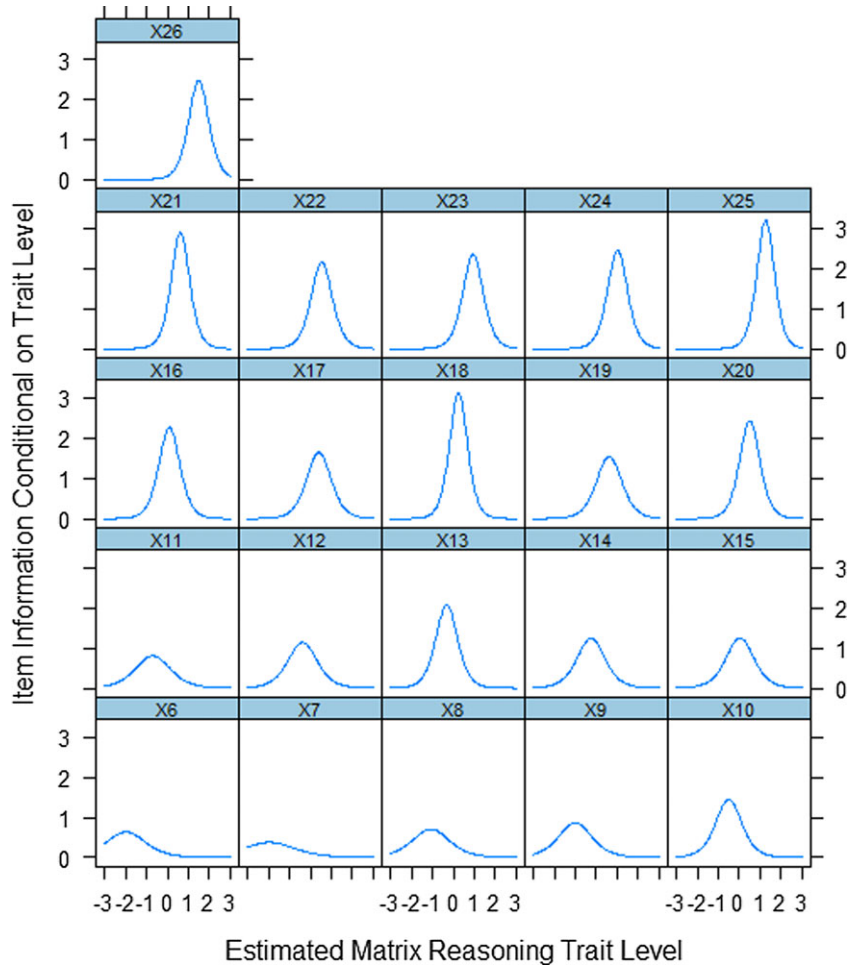


Figure 1. Item information functions for items 6–26. *Note.* The item information functions for items 1–5 could not be estimated as nearly all individuals got these questions correct, and therefore provide no information.

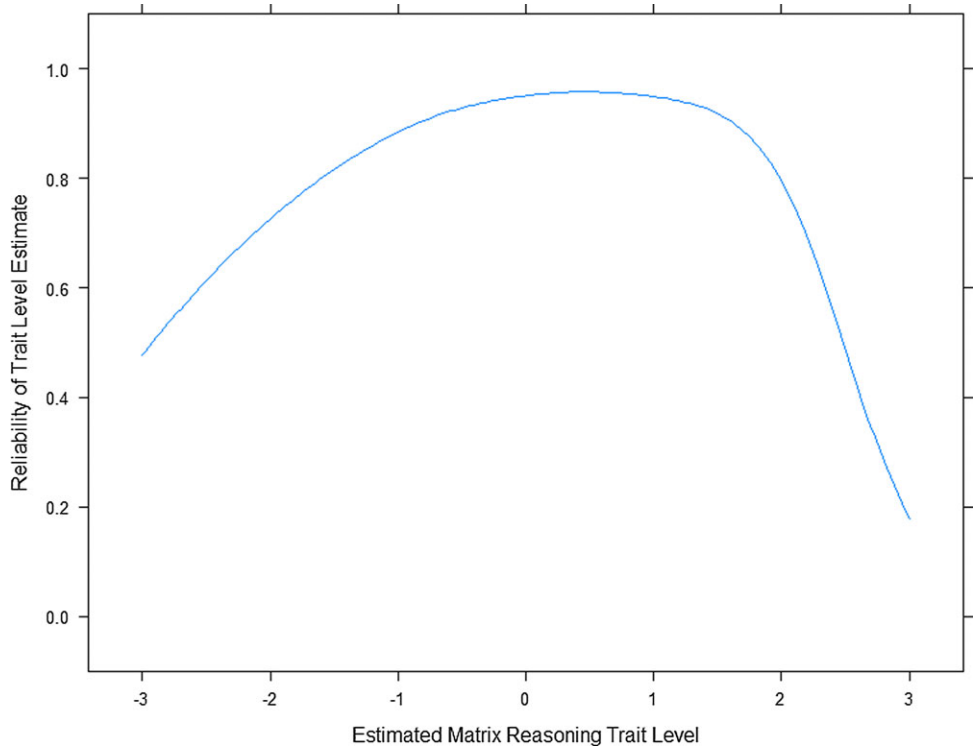


Figure 2. Reliability estimate conditional on trait level.

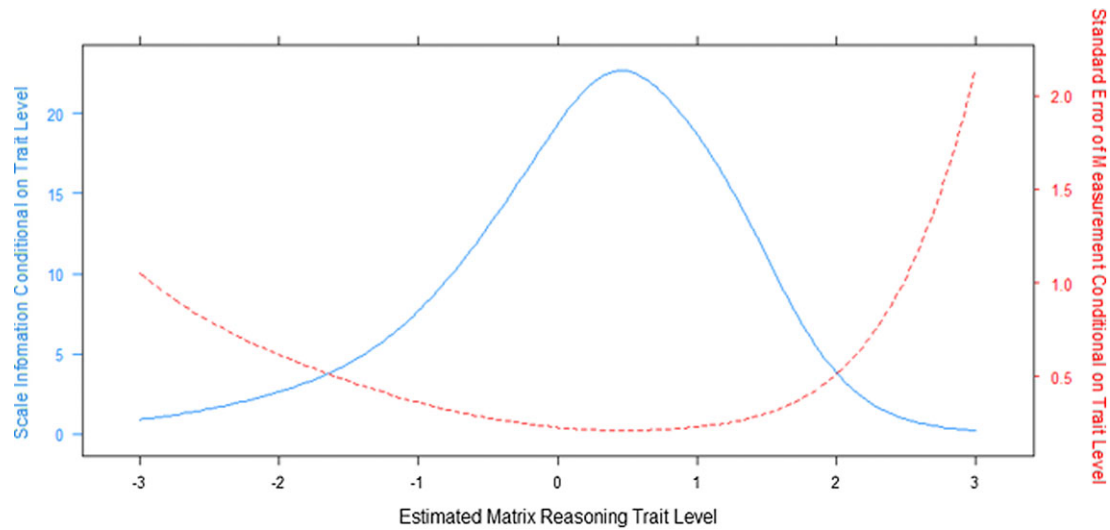


Figure 3. Test information and standard error conditional on trait level.

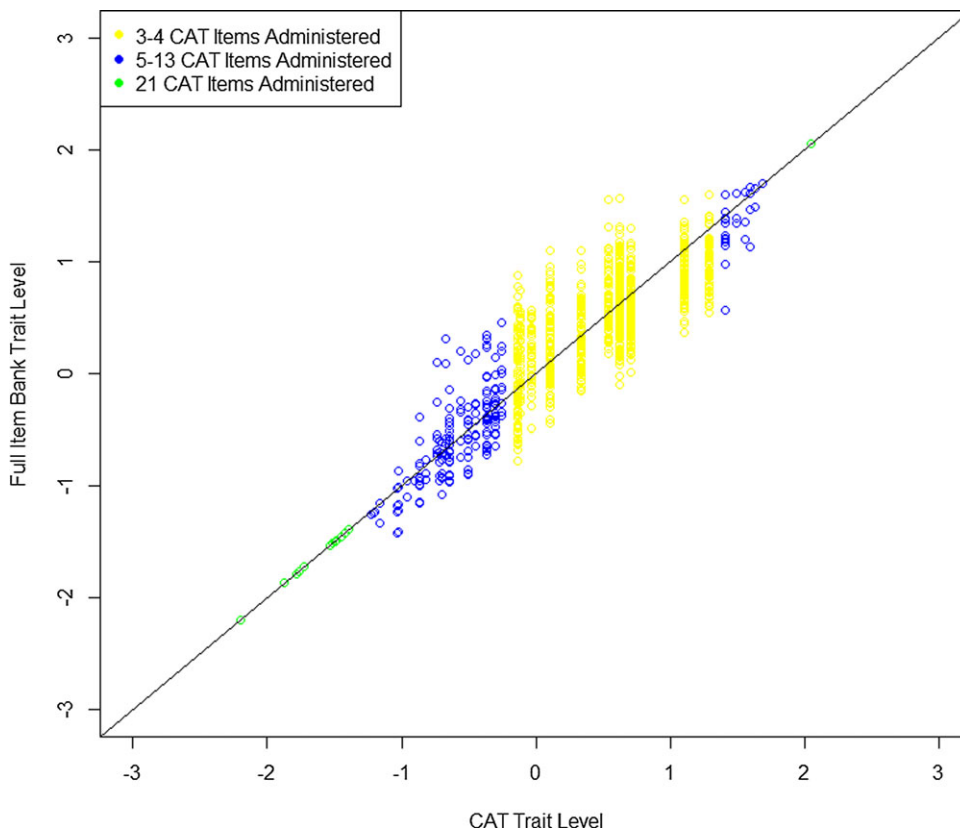


Figure 4. Computerized adaptive test estimated matrix reasoning trait levels versus full scale estimated trait level.

subtests follow an item format. The two core subtests that do not are Coding and Symbol Search. It remains unclear how much time is necessary to arrive at accurate estimates of true scores for these subtests; this could be a valuable topic for future research. Our prior confirmatory factor analyses further showed that WAIS-IV index scores, including the Perceptual Reasoning Index to which MR contributes, might be well estimated with only 2 subtests per index (Bilder et al., 2022).

Many other factors might alter these estimates of time needed for examination, and in cases where performance is

inconsistent from trial to trial an adaptive algorithm might take longer to execute successfully. If our results hold for other tests, however, we could anticipate at least doubling of efficiency, which would have a major impact on the potential throughput of cases in existing clinics nationwide. Increasing throughput has two major implications: (1) increasing access to neuropsychological assessment by the many patients who may benefit from it; and (2) decreasing the costs in both time and money for the patients who do receive neuropsychological services.

If clinical neuropsychology increases use of adaptive testing methods, it will be important to determine how much precision we want in these assessments. The current practice, using fixed-length tests, probably “over-tests” many patients and provides a higher level of precision in measurement of the tested construct than is clinically useful. A classic example may be drawn from studies of the Wisconsin Card Sorting Test (WCST) in schizophrenia. There was once heated debate about whether it might be reasonable to use the “short” (64 card) version of the WCST rather than the standard version (128 cards). A study (Prentice *et al.*, 2008) examining individual item responses found, however, that only four cards were needed to provide most of the information relevant to differentiate the schizophrenia group from healthy volunteers.

The model fitting results from the current study revealed that item difficulty is well specified by the order of Matrix Reasoning items following standard administration. In our sample, we also found very little information was provided by the first five items. It may be that these items are helpful in some populations of very low ability, but it is noteworthy that these items were of little value in our sample, which included patients with severe neurological and psychiatric disorders. The IRT modeling further showed that precision of measurement is very good from about -1.5 to $+2$ standard deviations around the mean, but worse as values move to extremes. For applications of this test that particularly require assessment of patients with very low levels of ability, additional easy items would be needed. For research on individuals of exceptionally high abilities, more difficult items would be needed.

A key assumption in this work is that the Matrix Reasoning subtest measures a unitary construct. Our dimensionality assessment revealed a ten-fold difference between the first eigenvalue (13.09) and the next highest eigenvalue (1.30) and the next three eigenvalues were less than 1, suggesting a strong first factor. Finally, a confirmatory factor analysis showed that the unidimensional solution had acceptable fit. Thus, the assumption of unidimensionality is generally sound. On the other hand, it is plausible that more than one dimension might be measured by the Matrix Reasoning subtest. Prior research on similar tests, such as the Raven Advanced Progressive Matrices (Raven, 1998), has suggested that this task involves *Gf* (general fluid intelligence, following Carroll (1993)), and more specifically inductive reasoning. Carpenter *et al.* (1990) examined multiple logical rules involved in different RAPM problems and suggested a dual-process model centering on inductive reasoning (finding abstract relations and rules) and working memory (goal management) processes. It is possible that distinct, separable processes like these might be more prominent in samples with specific deficits or strengths in the relevant abilities.

Understanding item-level responses on the Matrix Reasoning subtest may further help development of novel procedures that will enable open access assessment using adaptive testing methods. One interesting example of this kind of innovation already has been undertaken with the Matrix Reasoning Item Bank (MaRs-IB; Chierchia *et al.*, 2019). In theory, future development could lead to creation of large, shared, open-access banks of items along with their IRT parameters, that would enable free and flexible administration of adaptive testing of the processes involved in solving these widely used tests.

Limitations

A strength and potential weakness of the NNN sample is that it is heterogeneous, with inclusion of all patients examined in

participating clinics. A potential weakness is that patients represent a wide range of conditions so the findings may not generalize to any one specific condition or to people who have no neuropsychological complaints. However, this diversity is also a strength since the findings are more likely to generalize to other clinics nationwide.

Another limitation is that our analyses considered each item dichotomously (as correct or incorrect) and did not investigate the possible selection of different response alternatives or qualitative response features. Further, we did not use information about how much time it took examinees to provide either correct or incorrect responses. Theoretically, valuable information may be provided by knowing which incorrect response option an examinee took, and further in how long it took them to arrive at those answers. Particularly relevant to potential future development of performance validity indicators, it may be possible in the future to identify selection of unusual response options, with atypically long or short response latencies, that do not “fit” with other estimates of the individual’s true ability as determined by other metrics. Very large numbers of examinees must be used to analyze data in this way, given that – by definition – such rare response selections are infrequent. We hope as our sample grows we will be able to address these issues.

Finally, several possible limitations to generalizability should be mentioned. First, this is a simulation rather than implementation study, so it is impossible to say how the CAT might change examinee response patterns. Standard administration involves progression from easy to more difficult items, which might enable examinees to learn as the test progresses. It is also possible that MR, by sequencing items in order of difficulty and including a stop rule, rather than giving items randomly, as is done in some achievement tests, might yield different scores because examinees may not get all items. To perform an IRT calibration study, we would suggest administering all items, not in a fixed order, so that item characteristics can be estimated free from possible item-order effects. Finally, we are sensitive to the fact that application of adaptive tests requires the use of computers and/or internet access. This raises a concern about unequal access that may disproportionately impact patients who may not be able to access the necessary technology. Internet access is now prevalent in some regions (e.g., 93% in North America) but lower in others (43% in Africa). Even in regions with high overall access, those without access are disproportionately poor, from minority groups, and experience other adverse social and structural determinants of health outcomes⁴ Federal rules to eliminate “digital discrimination”⁵ raise hope that action will soon reduce current disparities at least in the United States. Despite these possible limitations, we believe the results of this demonstration suggest major practical advantages of using adaptive methods to increase efficiency and improve access to neuropsychological assessment.

Supplementary material. The supplementary material for this article can be found at <https://doi.org/10.1017/S1355617723000401>

Funding statement. This work was supported by a grant from the National Institute of Mental Health (R01MH118514).

Competing interests. Robert M. Bilder has received consulting fees and/or honoraria over the last 12 months from Atai Life Sciences, Karuna Therapeutics, the Institute of Digital Media and Child Development, and VeraSci. Dustin

⁴Miniwatts Marketing Group, © 2022, <https://www.internetworldstats.com/stats.htm>, accessed 1/21/2023.

⁵<https://www.fcc.gov/task-force-prevent-digital-discrimination>, accessed 1/21/2023.

Wahlstrom is a salaried employee of Pearson Clinical Assessment, which publishes the WAIS-IV, WMS-IV, CVLT-3, and D-KEFS.

References

- Barch, D. M., Gotlib, I. H., Bilder, R. M., Pine, D. S., Smoller, J. W., Brown, C. H., Huggins, W., Hamilton, C., Haim, A., & Farber, G. K. (2016). Common measures for National Institute of Mental Health funded research. *Biological Psychiatry*, 79(12), E91–E96. <https://doi.org/10.1016/j.biopsych.2015.07.006>
- Bilder, R. M. (2011). Neuropsychology 3.0: Evidence-based science and practice. *Journal of the International Neuropsychological Society*, 17(1), 7–13. <https://doi.org/10.1017/s1355617710001396>
- Bilder, R. M., & Reise, S. P. (2019). Neuropsychological tests of the future: How do we get there from here? *The Clinical Neuropsychologist*, 33(2), 220–245. <https://doi.org/10.1080/13854046.2018.1521993>
- Bilder, R. M., Widaman, K. F., Bauer, R. M., Drane, D., Loring, D. W., Umfleet, L. G., Reise, S. P., Vannier, L. C., Wahlstrom, D., Fossum, J. L., Wong, E., Enriquez, K., Whelan, F., & Shih, S. (2022). Construct identification in the neuropsychological battery: What are we measuring? *Neuropsychology*, 37(4), 351–372. <https://doi.org/10.1037/neu0000832>
- Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, 6(4), 431–444. <https://doi.org/10.1177/014662168200600405>
- Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: a theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review*, 97(3), 404. <https://doi.org/10.1037/0033-295X.97.3.404>
- Carroll, J. B. (1993). *Human Cognitive Abilities: A Survey of Factor-Analytic Studies (No. 1)*. Cambridge University Press.
- Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48(6), 1–29. <https://doi.org/10.18637/jss.v048.i06>
- Chierchia, G., Fuhrmann, D., Knoll, L. J., Pi-Sunyer, B. P., Sakhardande, A. L., & Blakemore, S. J. (2019). The matrix reasoning item bank (MaRs-IB): Novel, open-access abstract reasoning items for adolescents and adults. *Royal Society Open Science*, 6(10), 190232. <https://doi.org/10.1098/rsos.190232>
- Choi, S. W. (2009). Firestar: Computerized adaptive testing simulation program for polytomous item response theory models. *Applied Psychological Measurement*, 33(8), 644. <https://doi.org/10.1177/0146621608329892>
- Choi, S. W. (2020). Firestar: Computerized Adaptive Testing (CAT) Simulation Program. R package version 1.9.2. <https://github.com/choi-phd/Firestar>.
- Crane, P. K., Narasimhalu, K., L., E., Mungas, D. M., Haneuse, S., Larson, E. B., Kuller, L., Hall, K., & van Belle, G. (2008). Item response theory facilitated cocalibrating cognitive tests and reduced bias in estimated rates of decline. *Journal of Clinical Epidemiology*, 61(10), 1018–1027.e9.
- Gershon, R. C., Cook, K. F., Mungas, D., Manly, J. J., Slotkin, J., Beaumont, J. L., & Weintraub, S. (2014). Language measures of the NIH Toolbox Cognition Battery. *Journal of the International Neuropsychological Society*, 20(6), 642–651. <https://doi.org/10.1017/S1355617714000411>
- Gibbons, R. D., Weiss, D. J., Frank, E., & Kupfer, D. (2016). Computerized adaptive diagnosis and testing of mental health disorders. *Annual Review of Clinical Psychology*, 12(1), 83–104. <https://doi.org/10.1146/annurev-clinpsy-021815-093634>
- Gibbons, R. D., Weiss, D. J., Kupfer, D. J., Frank, E., Fagiolini, A., Grochocinski, V. J., Bhaumik, D. K., Stover, A., Bock, R. D., & Immekus, J. C. (2008). Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatric Services*, 59(4), 361–368. <https://doi.org/https://doi.org/10.1176/ps.2008.59.4.361>
- Holdnack, J. A., Zhou, X., Larrabee, G. J., Millis, S. R., & Salthouse, T. A. (2011). Confirmatory factor analysis of the WAIS-IV/WMS-IV. *Assessment*, 18(2), 178–191.
- Loring, D. W., Bauer, R. M., Cavanagh, L., Drane, D. L., Enriquez, K. D., Reise, S. P., Shih, K., Umfleet, L. G., Wahlstrom, D., Whelan, F., Widaman, K. F., Bilder, R. M., & NNN Study Group (2021). Rationale and design of the National Neuropsychology Network. *Journal of the International Neuropsychological Society*, 28(1), 1–11. <https://doi.org/10.1017/S1355617721000199>
- Maydeu-Olivares, A., Cai, L., & Hernández, A. (2011). Comparing the fit of item response theory and factor analysis models. *Structural Equation Modeling: A Multidisciplinary Journal*, 18(3), 333–356. <https://doi.org/10.1080/10705511.2011.581993>
- McCarty, C. A., Huggins, W., Aiello, A. E., Bilder, R. M., Hariri, A., Jernigan, T. L., Newman, E., Sanghera, D. K., Strauman, T. J., & Zeng, Y. (2014). PhenX RISING: Real world implementation and sharing of PhenX measures. *BMC Medical Genomics*, 7(1), 16.
- Moore, T. M., Di Sandro, A., Scott, J. C., Lopez, K. C., Ruparel, K., Njokweni, L. J., Santra, S., Conway, D. S., Port, A. M., D'Errico, L., Rush, S., Wolf, D. H., Calkins, M. E., Gur, R. E., & Gur, R. C. (2023). Construction of a computerized adaptive test (CAT-CCNB) for efficient neurocognitive and clinical psychopathology assessment. *Journal of Neuroscience Methods*, 386, 109795. <https://doi.org/10.1016/j.jneumeth.2023.109795>
- Moore, T. M., Scott, J. C., Reise, S. P., Port, A. M., Jackson, C. T., Ruparel, K., Savitt, A. P., Gur, R. E., & 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(3), 955–964. <https://doi.org/10.1037/pas0000102>
- Mungas, D., Reed, B. R., & Kramer, J. H. (2003). Psychometrically matched measures of global cognition, memory, and executive function for assessment of cognitive decline in older persons. *Neuropsychology*, 17(3), 380–392. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=12959504
- Mungas, D., Reed, B. R., Marshall, S. C., & Gonzalez, H. M. (2000). Development of psychometrically matched English and Spanish language neuropsychological tests for older persons. *Neuropsychology*, 14(2), 209–223.
- Penrose, L. S., & Raven, J. C. (1936). A new series of perceptual tests: Preliminary communication. *British Journal of Medical Psychology*, 16, 97–104. <https://doi.org/10.1111/j.2044-8341.1936.tb00690.x>
- Prentice, K. J., Gold, J. M., & Buchanan, R. W. (2008). The Wisconsin card sorting impairment in schizophrenia is evident in the first four trials. *Schizophrenia Research*, 106(1), 81–87. <https://doi.org/10.1016/j.schres.2007.07.015>
- 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(1), 33–65. <https://doi.org/10.1016/j.acn.2004.02.005>
- Rabin, L. A., Paolillo, E., & Barr, W. B. (2016). Stability in test-usage practices of clinical neuropsychologists in the United States and Canada over a 10-year period: A follow-up survey of INS and NAN members. *Archives of Clinical Neuropsychology*, 31(3), 206–230. <https://doi.org/10.1093/arclin/acw007>
- Raven, J. C. (1998). APM Raven's Advanced Progressive Matrices.
- Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., Thissen, D., Revicki, D. A., Weiss, D. J., Hambleton, R. K., Liu, H., Gershon, R., Reise, S. P., Lai, J. S., & Cella, D. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical Care*, 45(5 Suppl 1), S22–S31. <https://doi.org/10.1097/01.mlr.0000250483.85507.04>
- Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology*, 5(1), 27–48. <https://doi.org/10.1146/annurev.clinpsy.032408.153553>
- Revelle, W. (2023). *psych: Procedures for Psychological, Psychometric, and Personality Research*. R package version 2.3.3. Evanston, IL: Northwestern University.
- Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48, 1–36. <https://doi.org/10.18637/jss.v048.i02>
- Thompson, N. A., & Weiss, D. A. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research, and Evaluation*, 16(1), 1. <https://doi.org/10.7275/wqzt-9427>
- van der Linden, W. J., & Glas, C. A. W. (2000). *Computer adaptive testing: Theory and practice*. Kluwer Academic Publishers.
- Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., & Mislevy, R. J. (2000). *Computerized adaptive testing: A primer*. Routledge. <https://doi.org/10.4324/9781410605931>

- Wechsler, D. (1997). *WAIS-3, WMS-3: Wechsler Adult Intelligence Scale, Wechsler Memory Scale: Technical Manual*. Psychological Corporation.
- Wechsler, D. (2008). *Wechsler Adult Intelligence Scale-Fourth Edition (WAIS-IV)* (Vol. 22, p. 498). NCS Pearson.
- Wechsler, D. (2008a). *Wechsler Adult Intelligence Scale, Fourth Edition (WAIS-IV) Manual*. Pearson.
- Wechsler, D. (2008b). *Wechsler Memory Scale – Fourth Edition (WMS-IV)*. Psychological Corporation.
- Wechsler, D., Coalson, D. L., & Raiford, S. E. (2008). *Wechsler Adult Intelligence Scale: Fourth Edition. Technical and interpretative manual*. NCS Pearson, Inc.
- Yudien, M. A., Moore, T. M., Port, A. M., Ruparel, K., Gur, R. E., & Gur, R. C. (2019). Development and public release of the Penn Reading Assessment Computerized Adaptive Test (PRA-CAT) for premorbid IQ. *Psychological Assessment, 31*(9), 1168–1173. <https://doi.org/10.1037/pas0000738>