


## RESEARCH ARTICLE

# Adaptive measurement of cognitive function based on multidimensional item response theory

Robert D. Gibbons<sup>1</sup>  | Diane S. Lauderdale<sup>2</sup> | Robert S. Wilson<sup>3</sup> | David A. Bennett<sup>3</sup> |  
 Tesnim Arar<sup>4</sup> | David A. Gallo<sup>4</sup>

<sup>1</sup>Departments of Medicine and Public Health Sciences and Center for Health Statistics, University of Chicago, Chicago, Illinois, USA

<sup>2</sup>Department of Public Health Sciences, University of Chicago, Chicago, Illinois, USA

<sup>3</sup>Department of Neurological Sciences and Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago, Illinois, USA

<sup>4</sup>Department of Psychology, University of Chicago, Chicago, Illinois, USA

## Correspondence

Robert D. Gibbons, Blum-Riese Professor of Statistics, University of Chicago, Chicago, IL 60637, USA.

Email: [rdg@uchicago.edu](mailto:rdg@uchicago.edu)

## Funding information

NIA, Grant/Award Numbers: R56 AG066127, R56 AG0840701

## Abstract

**INTRODUCTION:** Up to 20% of older adults in the United States have mild cognitive impairment (MCI), and about one-third of people with MCI are predicted to transition to Alzheimer's disease (AD) within 5 years. Standard cognitive assessments are long and require a trained technician to administer. We developed the first computerized adaptive test (CAT) based on multidimensional item response theory (MIRT) to more precisely, rapidly, and repeatedly assesses cognitive abilities across the adult lifespan. We present results for a prototype CAT (pCAT-COG) for assessment of global cognitive function.

**METHODS:** We sampled items across five cognitive domains central to neuropsychological testing (episodic memory [EM], semantic memory/language [SM], working memory [WM], executive function/flexible thinking, and processing speed [PS]). The item bank consists of 54 items, with 9 items of varying difficulty drawn from six different cognitive tasks. Each of the 54 items has 3 response trials, yielding an ordinal score (0–3 trials correct). We also include three long-term memory items not designed for adaptive administration, for a total bank of 57 items. Calibration data were collected in-person and online, calibrated using a bifactor MIRT model, and pCAT-COG scores validated against a technician-administered neuropsychological battery.

**RESULTS:** The bifactor MIRT model improved fit over a unidimensional IRT model ( $p < 0.0001$ ). The global pCAT-COG scores were inversely correlated with age ( $r = -0.44$ ,  $p < 0.0001$ ). Simulated adaptive administration of 11 items maintained a correlation of  $r = 0.94$  with the total item bank scores. Significant differences between mild and no cognitive impairment (NCI) were found (effect size of 1.08 SD units). The pCAT-COG correlated with clinician-based global measure ( $r = 0.64$ ).

**DISCUSSION:** MIRT-based CAT is feasible and valid for the assessment of global cognitive impairment, laying the foundation for the development of a full CAT-COG that will draw from a much larger item bank with both global and domain specific measures of cognitive impairment.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2024 The Author(s). Alzheimer's & Dementia: Translational Research & Clinical Interventions published by Wiley Periodicals LLC on behalf of Alzheimer's Association.

**KEYWORDS**

Alzheimer's disease, bifactor model, cognitive impairment, computerized adaptive testing, multidimensional item response theory, neuropsychological assessment

**Highlights**

- As Americans age, numbers at risk for developing cognitive impairment are increasing.
- Aging-related declines in cognition begins decades prior to the onset of obvious cognitive impairment.
- Traditional assessment is burdensome and requires trained clinicians.
- We developed an adaptive testing framework using multidimensional item response theory.
- It is comparable to lengthier in-person assessments that require trained psychometrists.

## 1 | BACKGROUND

As the country's population ages, the number of older adults at risk for developing cognitive impairment is poised to increase substantially. It is estimated that up to 20% of people over 65 have mild cognitive impairment (MCI), and up to one-third with MCI transition to Alzheimer's dementia (AD) within 5 years, the most common form of dementia.<sup>1</sup> While AD is currently irreversible, there are promising interventions to slow its progression. This would be greatly facilitated by measuring early stages of cognitive decline in young adults which is a key risk factor for dementia and other health problems later in life.<sup>2,3</sup> Precise and repeatable measures of cognitive ability are essential for diagnosing MCI and dementia, as well as basic research applications, such as validating new biomarkers and characterizing the cognitive variability in healthy adult samples for behavioral research.<sup>4</sup>

Standardized cognitive task batteries administered via neuropsychological evaluation have long been the gold standard; data suggest that some domains of age-related change in cognitive performance can begin in early adulthood.<sup>5</sup> Although longitudinal designs are needed to assess cognitive change in different cognitive domains from early to late adulthood,<sup>6,7</sup> the repeated administration of traditional measures is costly (i.e., lengthy in-person test sessions), introduces practice effects on repeated items that can obscure cognitive changes,<sup>8,9</sup> and floor and ceiling effects when testing younger and older adults both with and without cognitive impairment. These obstacles limit cognitive testing in younger populations, limit access to cognitive assessment in underserved populations, and limit the interpretation of repeat assessments over time.

By leveraging recent technology in cloud-based computing as well as in CAT, we are positioned to overcome these limits and greatly enhance the speed, precision, and frequency of cognitive assessment over time. Researchers have demonstrated that administering cognitive tests online is feasible, equally reliable,<sup>10,11</sup> and well tolerated by younger and older participants. Sliwinski<sup>12</sup> demonstrated very

high (>0.97) reliability using brief, ambulatory measurement of standard (non-adaptive) cognitive tasks in a diverse sample (25–65 years, see also Brouillette<sup>13</sup>). Importantly, Sliwinski<sup>12</sup> also reported significant yet lower within-person reliability (0.41–0.53), highlighting the need to develop more precise assessment tools. Existing computerized approaches that repeatedly use the same items cannot overcome this limitation, as they rely on a direct and simplified translation of cognitive tasks for computers (or smart devices). However, recent advances in MIRT<sup>14</sup> have made it possible to develop adaptive cognitive tests based on very large item banks, to measure an individual's global cognition as well as several (sub)domains (e.g., memory or language). These techniques can increase precision while using far fewer items, dramatically reducing assessment time, and providing constant measurement precision across the entire range of cognitive ability, enhancing detection of cognitive impairment at its onset.<sup>14</sup> Current uses of IRT for cognitive tests are unidimensional and limited to specific domains (e.g., NIH Toolbox uses CAT only for a single language task), and so do not harness this full potential.

The most popular relatively brief cognitive screeners are the Mini Mental Status Exam (MMSE) and the Montreal Cognitive Assessment (MoCA). While their brevity is attractive, they are imprecise compared to longer test batteries, do not assess cognitive subdomains, subject to learning confounds (e.g., repeatedly being asked to identify a picture of a rhino),<sup>15</sup> and suffer from significant ceiling and floor effects.

MIRT-based CAT has already enjoyed considerable success in the measurement of mental health and substance use disorders and suicide risk stratification.<sup>16</sup> For example, Gibbons<sup>17</sup> demonstrated that diagnosis of major depressive disorder could be obtained in less than a minute and maintain sensitivity of 0.95 and specificity of 0.87 relative to an hour-long DSM-5 structured clinical interview. Gibbons<sup>18</sup> showed that a 389-item bank for depression could be adaptively administered in 2 min (average of 10 items) while maintaining correlation of  $r = 0.95$  with the 389-item bank score. Similar results have been obtained for other psychiatric conditions,<sup>19–22</sup> in English

and Spanish<sup>23</sup> and diverse environments (primary care, emergency medicine, and clinics<sup>24–27</sup>). Current applications of the CAT-MH (the collection of these adult CATs) include the largest national survey of mental health and substance use disorders<sup>28</sup>; availability in all United States Department of Veterans Affairs (US VA) clinics<sup>29</sup>; and screening community college students for mental health disorders and suicide risk in California.<sup>30</sup>

In addition to measuring global cognition, more precise measurement of different cognitive subdomains is critically important in research and practice. For example, processing speed (PS) and episodic memory (EM) exhibit more rapid age-related decline compared to semantic memory.<sup>8</sup> PS is sensitive to white matter degradation with aging and cerebrovascular disease, whereas EM is sensitive to pathological decline in medial temporal regions that characterize MCI and AD,<sup>31,32</sup> highlighting the need to assess specific cognitive domains.

To demonstrate feasibility, we created a pilot bank of cognitive items designed specifically for adaptive administration to assess global cognitive ability. The items were drawn from five domains,<sup>33</sup> with one task representing the domains of EM, PS, semantic memory/language (SM), and working memory (WM), and two tasks tapping the executive function domain, which we refer to here as flexible cognition/reasoning (FC). We call this pilot CAT the prototype CAT (pCAT-COG), which is foundational to the development of the full CAT-COG. Here we describe the item bank, model calibration, validation of the items in the pCAT-COG, and validation of the resulting estimated primary cognitive dimension score. The CAT-COG will be based on a nine-fold expansion of the item bank and add scoring of the five domains in addition to the global cognition dimension. The CAT-COG will be implemented in a cloud-computing environment (see [Supplementary Material A](#)).

## 2 | METHODS

The development of the pCAT-COG is a multistep process that involves (i) the creation of an item bank suitable for adaptive testing, (ii) calibration of items on participant samples, (iii) item analysis, (iv) selection of CAT tuning parameters, and (v) validation against extant cognitive measures and clinical diagnostic categories (MCI).

### 2.1 | Participants

We collected calibration data from both in-person and online samples. In-person data were collected by the Rush Alzheimer's Disease Center (RADC)<sup>34</sup> and yielded 84 participants (age 18–101 years [mean = 77.42, SD = 7.55], 66% female, 0% Hispanic individuals, mean years of education 16.47 [SD = 2.59], 99% White individuals, and 1% Black individuals), who were administered (i) the complete RADC neuropsychological battery<sup>35</sup> and (ii) in a separate session, the pCAT-COG task. Twenty-four RADC participants had MCI, and one had MCI/early-AD. We also recruited 646 participants online (Prolific), (18–89 years of age [mean = 48.41, SD = 20.42], 53% female, 2% Hispanic individuals, mean years of education 14.99 [SD = 3.80], 81% White individuals, 7% Black individuals, and 8% Asian individuals), from which we created

### RESEARCH IN CONTEXT

- Systematic review:** With the aging of the American population, the number of older adults at risk for developing cognitive impairment has increased substantially. Recent research points to aging-related change in cognitive performance beginning decades prior to the onset of cognitive impairment, highlighting the need for cognitive assessment that is valid across the entire adult lifespan.
- Interpretation:** Cognitive function has long been assessed using standardized cognitive tasks administered via neuropsychological evaluation, requiring lengthy in-person assessments with trained personnel. We developed a new approach based on a computerized adaptive test (CAT) developed through multidimensional item response theory (MIRT) to assess cognitive function, either in clinic or remotely (online).
- Future directions:** Our approach will revolutionize computer-based cognitive testing (ultimately in a platform independent way), providing precise estimation of an individual's cognitive ability overall and on specific domains with minimal respondent burden, using a sufficiently large bank of items so that the same individual's cognitive ability can be assessed repeatedly and efficiently without reusing items or stimuli.

the pCAT-COG. The final CAT-COG will be based on a sample that is more nationally representative in terms of racial/ethnic diversity.

### 2.2 | RADC core measures and assessments

Cognitive function is tested in the RADC cohort studies via a battery of tests administered annually as home visits.<sup>35</sup> Nineteen tests across a range of cognitive abilities are used to construct a global composite measure of cognitive function and separate summary measures of five domains similar to those in the pCAT-COG.

### 2.3 | pCAT-COG item bank

The pCAT-COG included six cognitive tasks: episodic recognition memory, object naming, PS, digit span forward and backward, Stroop, and rule identification. For each of these tasks, we included 9 items (three trials each) that varied in difficulty, yielding 54 items. We also included nine EM trials (word stimuli – three ordinal items), to compare longer delay to our EM items with a shorter delay that are designed for adaptive testing. The pCAT-COG starts with an instruction phase with a practice item for each task type (approximately 5 min, repeatable), and then to simulate adaptive testing conditions, we administered the items in (i) three blocks of increasing difficulty, and (ii) within each block, randomly administered to mimic participants' ability to switch

between item-types similar to how a CAT would be administered. More details of the six pCAT-COG tasks are presented in the [Supplementary Material B](#).

## 2.4 | Bifactor model

To estimate a person's cognitive ability on the global dimension when the items are sampled from five different cognitive domains, we used a bifactor MIRT model for ordinal response data.<sup>14</sup> Traditional IRT assumes that the item-intercorrelations are completely explained by a single latent variable. For complex constructs like cognition, made up of multiple domains, this assumption is invalid and can yield biased estimates of item parameters, ability estimates and their uncertainty.<sup>14</sup> To accommodate multidimensionality, the bifactor model allows each item to load on the primary dimension and one subdimension that describes the domain from which the item was selected (e.g., WM). The bifactor model has the advantage of estimating a cross-cutting measure of ability while at the same time incorporating multidimensionality, also simplifying adaptive testing. Briefly, the bifactor model for a binary response item (used here for simplicity) is described by the factor pattern matrix

$$\alpha = \begin{bmatrix} \alpha_{11} & \alpha_{12} & 0 \\ \alpha_{21} & \alpha_{22} & 0 \\ \alpha_{31} & 0 & \alpha_{33} \\ \alpha_{41} & 0 & \alpha_{43} \end{bmatrix}$$

where for this four-item example the first column describes the primary dimension upon which all items load and the second and third columns describe domain-specific factors (e.g., EM and PS), which absorb the residual correlation between items drawn from a specific domain. The marginal probability of the response pattern for subject  $i$  is given by

$$P = \int_{-\infty}^{\infty} \left\{ \prod_{v=2}^d \int_{-\infty}^{\infty} \left[ \prod_{j=1}^n \left( \Phi \left[ \frac{\gamma_j - \alpha_{j1}\theta_1 - \alpha_{jv}\theta_v}{\sqrt{1 - \alpha_{j1}^2 - \alpha_{jv}^2}} \right] \right)^{v_{jv}} \right] g(\theta_v) d\theta_v \right\} g(\theta_1) d\theta_1,$$

where  $\gamma_j$  is a threshold that represents the difficulty of the item,  $\alpha_{j1}$  is the loading of item  $j$  on the primary dimension,  $\alpha_{jv}$  is the loading of item  $j$  on the domain-specific factor  $v$ ,  $\theta_1$  and  $\theta_v$  are the abilities on the primary and domain-specific factors,  $v_{jv}$  is an indicator for the domain from which item  $j$  was drawn,  $\Phi$  is the cumulative normal distribution function, and  $g$  is the distribution of the latent variable  $\theta$ . Further details are presented in the [Supplementary Material C](#). The bifactor model for the pCAT-COG included a primary dimension and the five cognitive domains that form the basis of the item bank.

## 2.5 | Model calibration

First, items were inspected for zero frequency score categories and eliminated as too easy or too difficult. Second, overall model fit was

determined by comparing observed versus estimated response proportions over all items and response categories and computing their correlation. Third, we also compared models using the ordinal scoring (number correct out of three similar tasks) and analysis of the individual binary tasks. We studied our Stroop test, since preliminary analysis revealed that it is essentially unidimensional. This allowed us to compare two different modeling approaches; a graded unidimensional IRT model for the nine ordinal items and a binary bifactor model (the equivalent of a testlet IRT model<sup>14</sup>) that accommodated the nesting of the three binary response trials within each of the nine items. Fourth, we compared on-line and in-person percentage correct item responses in age-matched cognitively normal older adults to explore potential bias in mode of administration. Fifth, using our in-person sample, we compared performance of two kinds of EM items.

## 2.6 | Model selection

We used Bayesian information criterion (BIC), where a smaller number indicates better fit of the model to the data, to select the most parsimonious model among three alternatives: a unidimensional IRT model, the best fitting (in terms of number of dimensions) unrestricted MIRT model, and a bifactor MIRT model.

## 2.7 | Validation

Pearson product-moment correlation coefficients were used to assess the association between pCAT-COG scores and age, and the RADC composite measure. Student's  $t$ -tests were used to compare older RADC participants with and without MCI, and effect sizes were compared between the pCAT-COG, RADC composite measure, and the MoCA.

# 3 | RESULTS

## 3.1 | Feasibility of the pCAT-COG Items

All online participants across the aging spectrum were able to complete the pCAT-COG on their own, and only 17/84 elderly RADC participants required help of an RA. Only 3 (MCI) of the 84 participants had difficulty understanding the instructions for some tasks, and 83/84 completed the pCAT-COG.

## 3.2 | Item difficulty

For RADC participants, we found significant differences in performance (the proportion of trials correctly answered) across items we designed to be easy versus hard on each task: RuleID 0.94 (SE = 0.026) > 0.67 (SE = 0.051), Stroop 0.69 (SE = 0.050) > 0.64 (SE = 0.052), EM 0.90 (SE = 0.033) > 0.68 (SE = 0.051), PS 0.67 (SE = 0.051) > 0.32

( $SE = 0.051$ ),  $SM\ 0.99$  ( $SE = 0.011$ )  $> 0.94$  ( $SE = 0.026$ ), span  $0.70$  ( $SE = 0.050$ )  $> 0.51$  ( $SE = 0.055$ ) (all  $p$ 's  $< 0.05$ ). As expected SM was relatively easy for all participants, as well as the Stroop, which did not achieve a wide range of difficulty. We removed the hardest span items, as they were too hard for MCI, and we excluded these items from preliminary analyses. These results demonstrate that we can create items with sufficient variability in difficulty to measure participant ability across the aging spectrum.

### 3.3 | Estimated adaptive test timing

The cognitively unimpaired online participants usually required only one instructional/practice phase (~5 min), and on average, they completed the 57 items in 34.2 min  $SD = 9.3$  min (36 s per item) across the wide age range (upper limit = 39.7 m ( $SD = 7.3$ ) for those in 76–86 age range). The simulated pCAT-COG required an average of 11 items in order to reproduce the total item bank score with correlation of  $r = 0.94$ . This translates to an average of 6.5 min plus 5 min of practice, or a total of 11.5 min to assess global cognition. By comparison, the RADC battery takes 60–90 min by a trained administrator. MCI participants with in-person help took the longest on average, completing the 57 item pCAT-COG item bank in 47.3 min,  $SD = 9.4$  min (49 s per item), 9 min of task, plus 1–2 practice/instruction rounds (5–10 min), or 14–19 min to estimate global cognition in MCI.

### 3.4 | Online and in-person sample comparison

We compared an age-matched subset of cognitively normal older adults tested online (Prolific,  $n = 45$ , mean age = 79 years, range 76–86) and tested in-person (RADC,  $n = 35$ , mean age = 79 years, range = 69–87). Using classical scoring of the pCAT-COG task (percent correct), we found nearly identical performance between these two samples on five pCAT-COG measures (EM, short delay =  $0.82$  [ $SE = 0.057$ ] vs.  $0.81$  [ $SE = 0.066$ ]; EM, long delay =  $0.70$  [ $SE = 0.068$ ] vs.  $0.74$  [ $SE = 0.074$ ];  $SM = 0.95$  [ $SE = 0.032$ ] vs.  $0.97$  [ $SE = 0.029$ ]; Stroop =  $0.72$  [ $SE = 0.067$ ] vs.  $0.71$  [ $SE = 0.077$ ]; Rule ID =  $0.84$  [ $SE = 0.055$ ] vs.  $0.81$  [ $SE = 0.066$ ]). The only difference was in the speed task ( $0.64$  [ $SE = 0.072$ ] vs.  $0.53$  [ $SE = 0.084$ ]), as older participants tested online were somewhat faster than those tested in-person. These data demonstrate that online and in-person testing can yield meaningful and comparable older adult data.

### 3.5 | Episodic item delay

In the RADC sample, we found performance on the adaptive EM items (which use a 15 s math filler task between study and test to clear WM rehearsal) correlated highly with the longer-delayed EM items (given at the end of the entire pCAT-COG task,  $r = 0.64$ ,  $p < 0.001$ ). Although the immediate items were easier than the delayed items (proportion correct  $0.76$  [ $SE = 0.047$ ] vs.  $0.64$  [ $SE = 0.052$ ],  $p < 0.01$ ), these two kinds of EM items had comparably high correlations with the RADC EM composite score, which pools immediate and delayed item recall

( $r = 0.56$ ) and recognition and story memory ( $r = 0.60$ ). Item discrimination parameters were also equivalent between shorter and longer delayed EM items.

### 3.6 | Model calibration

After removing the 5 hardest span items, 46 ordinal items remained for a total of  $46 * 3 = 138$  binary trial items (excludes three long-term EM items). The fit of the bifactor model to the observed data was excellent (see Figure 1,  $r = 0.99$  between the observed and estimated proportions). For example, for item 10, 65.8% of the sample received a perfect score of 3 and the model estimate was 65.7%.

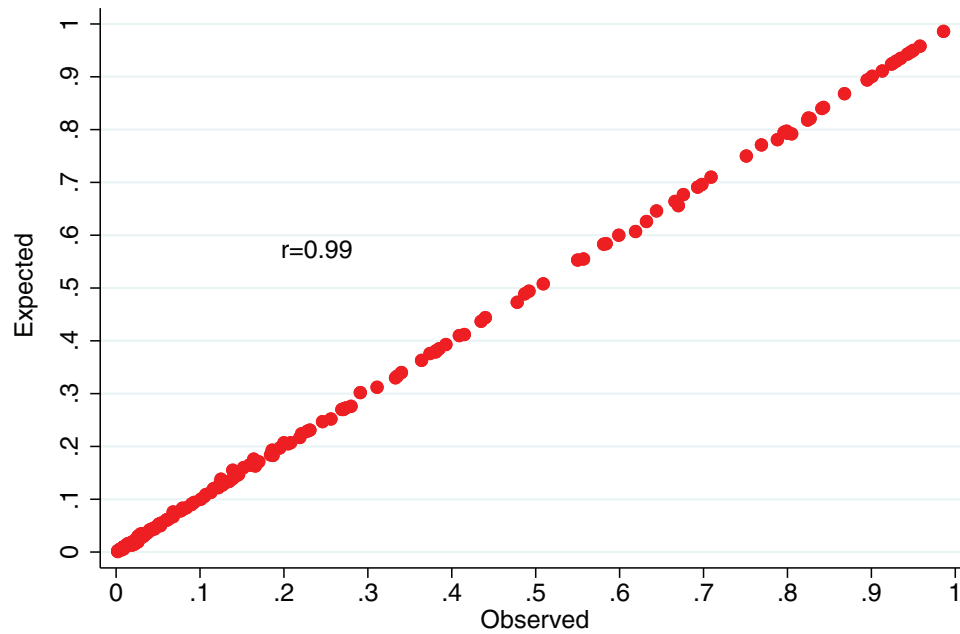
### 3.7 | Validation

The primary (global) dimension was significantly (inversely) associated with age ( $r = -0.44$ ,  $p < 0.0001$ ), see Figure 2. As shown in Figure 2, there is a linear decrease in global cognitive ability with age (plotted at decile midpoints), from ages 35 to 75 years, with a more rapid decrease of 0.8 SD units from ages 75 to 85 years.

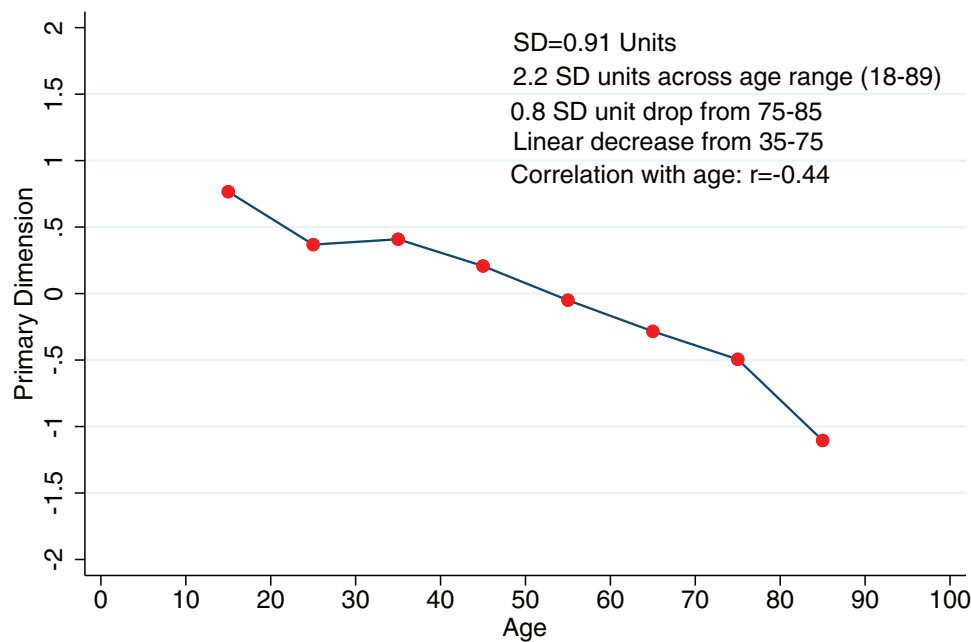
The primary pCAT-COG measure for global cognition was strongly correlated with the RADC Global dimension ( $r = 0.64$ ), as well as the other RADC composite scores with large effect sizes (Figure 3). The pCAT-COG scores significantly differentiated older RADC participants with MCI ( $n = 25$ ) and no cognitive impairment (NCI,  $n = 46$ ,  $t = 4.4$ ,  $df = 69$ ,  $p < 0.0001$ ) with effect size of 1.08 SD units, similar to the effect size for the RADC global composite (effect size = 1.47 SD units) and the MoCA (effect size = 1.14 SD units). Note that this is a lower bound on the final CAT-COG effect size, since the pCAT-COG is based on only 10% of the final item bank being developed. Although we refer to RADC as a “gold standard,” there is no perfect way of measuring cognitive function.

### 3.8 | Model selection

For the unrestricted MIRT models, the one-factor (traditional unidimensional IRT model) had  $BIC = 54,209$ , and the best fitting MIRT model was the four-factor model  $BIC = 53,593$ , with both three-factor and five-factor models exhibiting poorer fit ( $BICs\ 53,649$  and  $53,823$  respectively). The best fitting 4-factor MIRT model produced factors with (1) a mix of SM and executive function items, (2) a mix of SM and PS items, (3) a mix of executive function and PS items, and (4) a mix of WM and executive function items. We are unaware of any neuropsychological theory that would naturally group these tasks in this way, and in fact, the theoretically-guided bifactor model exhibited the best fit of all models tested with  $BIC = 53,106$  for the bifactor model with a primary and five pre-specified cognitive domains. Another advantage of the bifactor model is that it provides a global cognitive ability estimate, whereas the unrestricted MIRT model did not. For many applications the global score may be all that is required, further reducing participant burden.



**FIGURE 1** Bifactor model observed versus expected marginal item category proportions.

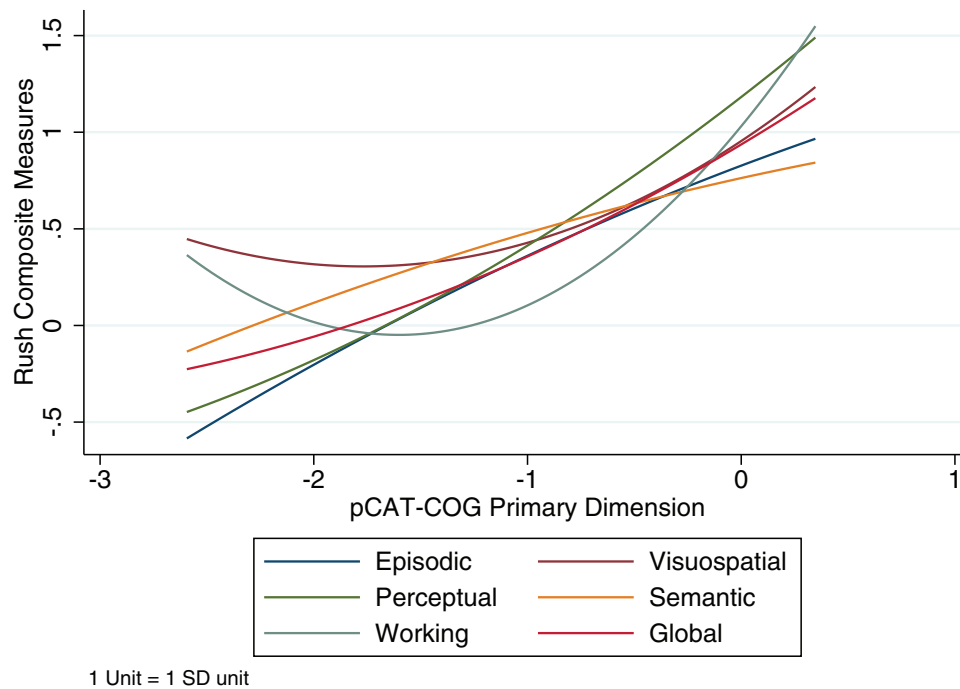


**FIGURE 2** Relationship between global pCAT-COG score and age. pCAT-COG, prototype computerized adaptive test.

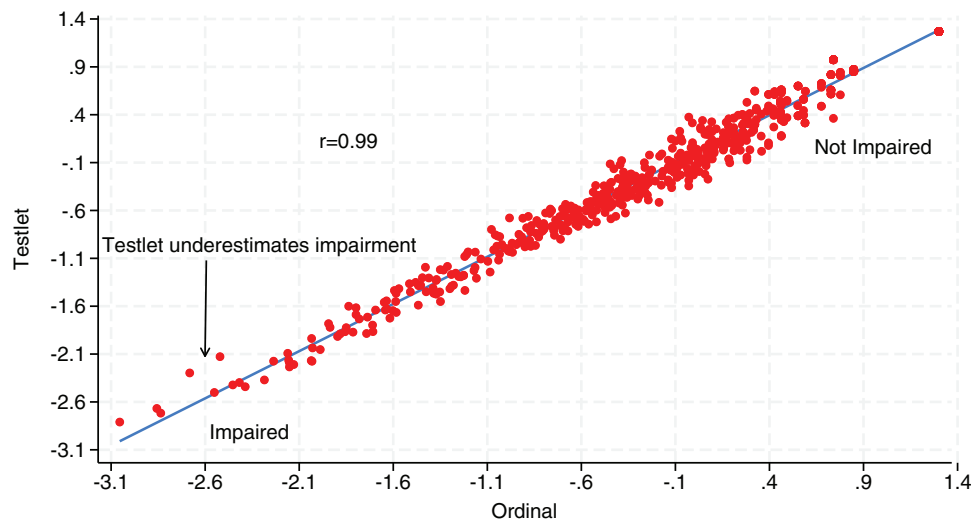
### 3.9 | Results for binary trial versus ordinal item scoring

Comparison of item parameter estimates based on ordinal and binary task scoring revealed that the ordinal item scoring produced a wider range of thresholds (-2.44 to 0.31) than the binary trial scoring (-1.61 to -0.15), indicating that the ordinal scoring yields more information

across the ability range. In terms of the estimated factor loadings, the ordinal scoring yielded large and homogeneous loadings across all nine ordinal items (0.65 to 0.79), whereas the binary trial model showed more heterogeneity with some trials less strongly related to the latent primary dimension (0.31 to 0.80). The two model results are highly correlated ( $r = 0.99$ ), however, the ordinal model had greater precision for the most cognitively impaired participants (Figure 4).



**FIGURE 3** pCAT-COG global cognition versus RADC composite measures. pCAT-COG, prototype computerized adaptive test; RADC, Rush Alzheimer's Disease Center.



**FIGURE 4** Binary testlet versus ordinal score (0–3 correct) – Stroop.

## 4 | DISCUSSION

We developed and tested a new approach to testing cognitive function. Applying adaptive testing to cognitive assessment and representing cognition through a multidimensional model are the two key innovations, allowing more rapid, precise, accessible, and repeatable measurements. Traditional measurement fixes the number of items administered and allows measurement uncertainty to vary. By contrast, a MIRT-based CAT fixes the acceptable level of measurement

uncertainty and allows the number of items to vary to achieve that level. This innovation dramatically reduces the number of items needed to measure cognitive ability with higher precision than traditional fixed-length tests. To date, most cognitive assessments have been based on classical test theory (380 out of 384 based on a recent review<sup>36</sup>) and the few IRT applications were non-adaptive and based on unidimensional IRT models which provide poor fit to these data as we showed. By contrast, MIRT allows us to measure global and ultimately domain-specific cognition, and CAT will select an optimal

set of test items for each individual targeted to their level of cognitive ability.

Our pilot study had limitations as well. First, future work will need to incorporate a more diverse sample with more Hispanic and Black participants and fewer years of education and examine measurement invariance. Second, the pCAT-COG was studied by simulating adaptive administration from complete item responses and therefore does not directly test adaptive administration. Historically, results are quite similar.<sup>18,21–23</sup> Third, our results are for the pilot item bank, and may not fully generalize to the much larger item bank currently being collected.

The advantages of MIRT-based CAT have the potential to transform cognitive assessment, improving cognitive screening and associated health outcomes (clinical impact) and allowing researchers to ask new kinds of questions (scientific impact). In addition to in-person administration in clinical, healthcare, or research settings, we demonstrated that a CAT-COG can be self-administered online, at home, in cognitively unimpaired younger and older adults. Increasing accessibility in this way will allow earlier detection of cognitive decline, including in populations that face barriers to assessment in formal healthcare settings. It also would facilitate large-scale descriptive and etiologic studies of individual or population-level cognition and permit high frequency cognitive assessments over short timespans within individuals (i.e., burst designs to capture intra-individual variability). It would allow a more precise and multidimensional cognitive assessment in omnibus population surveys. This would yield new kinds of data that can catalyze discoveries across the adult lifespan. For example, the ability to make frequent assessments will revolutionize research on the possible acute effects of major life events or risky behaviors on cognitive ability (e.g., new health problems or treatments, retirement, and losing a spouse). The possibility of online self-administration will allow for the concurrent measurement of transient psychological or physiological states (via self-report and biosensors) as well as environmental factors, such as amount of sleep, time-of-day, weather events, or diet, leading to new discoveries of the impact of such factors on cognitive ability across adulthood and among those with MCI.

## ACKNOWLEDGMENTS

We acknowledge Melissa Trevino and Jonathan King for their guidance in the development of this program of research. We also acknowledge Taylor A. Chamberlain, Coen D. Needell, and Martynas Snarskis for their contributions coding the tasks. This work was supported by NIA grants R56 AG066127 and R56 AG0840701.

## CONFLICT OF INTEREST STATEMENT

Robert Gibbons is a founder of Adaptive Testing Technologies which licenses computerized adaptive tests for the assessment of mental health and substance use disorders and suicidality. Robert Gibbons has been an expert witness for Amgen, AstraZeneca, Gerber, and GSK. Diane Lauderdale, Robert Wilson, David Bennett, Tesnim Arar, and David Gallo have no conflicts of interest to disclose. Author disclosures are available in the [Supporting Information](#).

## CONSENT STATEMENT

All human subjects provided informed consent.

## ORCID

Robert D. Gibbons  <https://orcid.org/0000-0002-6463-2280>

## REFERENCES

1. Alzheimer's Association. 2023 Alzheimer's disease facts and figures. *Alzheimers Dement*. 2023;19 (4):1598-1695.
2. Osler M, Christensen GT, Garde E, Mortensen EL, Christensen K. Cognitive ability in young adulthood and risk of dementia in a cohort of Danish men, brothers and twins. *Alzheimers Dement*. 2017; 13:1355-1363.
3. Huang AR, Strombotne KL, Horner EM, Lapham SJ. Adolescent cognitive aptitudes and later-in-life Alzheimer disease and related disorders. *JAMA Netw Open*. 2018;1:1-12.
4. Sperling RA, Aisen PS, Beckett LA, et al. Toward defining the pre-clinical stages of Alzheimer's disease: recommendations from the National Institute on Aging and the Alzheimer's Association Workgroup. *Alzheimers Dement*. 2011;7:1-13.
5. Salthouse TA. When does age-related cognition decline begin?. *Neurobiol Aging*. 2009;30:507-514.
6. Rönnlund M, Nyberg L, Bäckman L, Nilsson LG. Stability, growth, and decline in adult life span development of declarative memory: cross-sectional and longitudinal data from a population-based study. *Psychol Aging*. 2005;20:3-18.
7. Rönnlund M, Nilsson LG. Adult life-span patterns in WAIS-R block design performance: cross-sectional versus longitudinal age gradients and relations to demographic factors. *Intelligence*. 2006;34:63-78.
8. Salthouse TA. Continuity of cognitive change across adulthood. *Psychon Bull Rev*. 2016;23:932-939.
9. Salthouse TA. Shared and unique influences on age-related cognitive change. *Neuropsychology*. 2017;31:11-19.
10. Timmers C, Maeghs A, Vestjens M, Bonnemayer C, Hamers H, Blokland A. Ambulant cognitive assessment using a smartphone. *Appl Neuropsychol Adult*. 2014;21: 136-42.
11. Moore RC, Swendsen J, Depp CA. Applications for self-administered mobile cognitive assessments in clinical research: a systematic review. *Int J Methods Psychiatr Res*. 2017;26:1,12.
12. Sliwinski MJ, Mogle JA, Hyun J, Munoz E, Smyth JM, Lipton RB. Reliability and validity of ambulatory cognitive assessments. *Assessment*. 2016;25:1-17.
13. Brouillette, RM, Foil H, Fontenot S, et al. Feasibility, reliability, and validity of a smartphone based application for the assessment of cognitive function in the elderly. *PLoS One*, 2013;8:e65925.
14. Bock RD, Gibbons RD. *Item Response Theory*. Wiley; 2021.
15. Trzepacz PT, Hochstetler H, Want S, Walker B, Saykin AJ; for the Alzheimer's Disease Neuroimaging Initiative. Relationship between the Montreal Cognitive Assessment and Mini-mental State Examination for assessment of mild cognitive impairment in older adults. *BMC Geriatr*. 2015;15:107.
16. Gibbons RD. Computerized adaptive diagnosis and testing of mental health disorders. *Annu Rev Clin Psychol*. 2016;12:83-104.
17. Gibbons RD, Hooker G, Finkelman MD, et al. The CAD-MDD: a computerized adaptive diagnostic screening tool for depression. *J Clin Psychiatry*. 2013;74:669-674.
18. Gibbons RD, Weiss DJ, Pilkonis PA, et al. The CAT-DI: a computerized adaptive test for depression. *JAMA Psychiatry*. 2012;69:1104-1112.
19. Gibbons RD, Weiss DJ, Pilkonis PA, et al. Development of the CAT-ANX: a computerized adaptive test for anxiety. *Am J Psychiatry*. 2014;171:187-194.



20. Achtyes ED, Halstead S, Smart L, et al. Validation of computerized adaptive testing in an outpatient non-academic setting. *Psychiatr Serv*. 2015;66:1091-1096.
21. Gibbons RD, Kupfer D, Frank E, Moore T, Boudreaux ED. Development of a computerized adaptive suicide scale. *J Clin Psychiatry*. 2017;78:1376-1382.
22. Gibbons RD, Alegria M, Markle S, et al. Development of a Computerized Adaptive Substance Abuse Scale – The CAT-SUD. *Addiction*. 2020;115:1382-1394.
23. Gibbons RD, Alegria M, Cai L, et al. Successful validation of the CAT-MH scales in a sample of Latin American migrants in the U.S. and Spain. *Psychol Assess*. 2018;30:1267-1276.
24. Graham AK, Minc A, Staab E, Beiser DG, Gibbons RD, Laiterapong N. Validation of a computerized adaptive test for mental health in primary care. *Ann Fam Med*. 2019;17(1):23-30.
25. Beiser D, Vu M, Gibbons RD. Test-retest reliability of a computerized adaptive depression test. *Psychiatr Serv*. 2016;67:1039-1041.
26. Kim JJ, Silver RK, Elue R, et al. The experience of depression, anxiety and mania among perinatal women. *Arch Womens Ment Health*. 2017;19:94-100.
27. Aschebrook-Kilfoy B, Ferguson BA, Angelos P, et al. Development of the ThyCAT: A clinically useful computerized adaptive test to assess quality of life in thyroid cancer survivors. *Surgery*. 2017;163:137-142.
28. Ringeisen H, Edlund M, Guyer H, et al. *Mental and Substance Use Disorders Prevalence Study (MDPS): Findings Report*. RTI International; 2023.
29. Researchers from the Veteran's Health Administration, Academia, and Adaptive Testing Technologies validate the Computerized Adaptive Test Suicide Scale (CAT-SS) for U.S. Military Veteran. Adaptive Testing Technologies; 2022. Accessed 3/23/2024. <https://www.prnewswire.com/news-releases/researchers-from-the-veterans-health-administration-academia-and-adaptive-testing-technologies-validate-the-computerized-adaptive-test-suicide-scale-cat-ss-for-us-military-veterans-301467885.html>
30. Wen A, Wolitzky-Taylor, K, Gibbons RD, Craske MG. A randomized controlled trial on using predictive algorithms to adapt level of psychological care for community college students: STAND triaging and adapting to level of care study protocol. *Trials*. 2023;24(1):1-19.
31. Braak H, Braak E. Neuropathological staging of Alzheimer's related changes. *Acta Neuropathol*, 1991;82:239-259.
32. Guillozet AL, Weintraub S, Mash DC, Mesulam MM. Neurofibrillary tangles, amyloid, and memory in aging and mild cognitive impairment. *Arch Neurol*. 2003;60:729-736.
33. Weintraub S, Dikmen S, Heaton R, Tulsy D, Zelazo P, Slotkin J, Gershon R. The Cognition Battery of the NIH Toolbox for assessment of neurological and behavioral function: validation in an adult sample. *J Int Neuropsychol Soc*. 2014;20:567-578.
34. Bennett DA, Buchman AS, Boyle PA, Barnes LL, Wilson RS, Schneider JA. Religious Orders Study and Memory And Aging Project. *J Alzheimers Dis*. 2018;64:5161-5189.
35. Marquez DX, Glover CM, Lamar M, et al. Representation of older latinxs in cohort studies at the Rush Alzheimer's Disease Center. *Neuroepidemiology*. 2020;54(5):404-418.
36. McGory S, Doherty JM, Austin EJ, Starr JM, Shenkin SD. Item response theory analysis of cognitive tests in people with dementia: a systematic review. *BMC Psychiatry*. 2014;14:47.

### SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Gibbons RD, Lauderdale DS, Wilson RS, Bennett DA, Arar T, Gallo DA. Adaptive measurement of cognitive function based on multidimensional item response theory. *Alzheimer's Dement*. 2024;10:e70018. <https://doi.org/10.1002/trc2.70018>