Title

Measurement Precision Across Cognitive Domains in the Alzheimer's Disease Neuroimaging Initiative (ADNI) Data Set

Journal

Neuropsychology, 37(4)

ISSN

0894-4105

Authors

Crane, Paul K

Choi, Seo-Eun

Lee, Michael

et al.

Publication Date

2023-05-01

DOI

10.1037/neu0000901

Peer reviewed

Published in final edited form as:

Neuropsychology. 2023 May ; 37(4): 373–382. doi:10.1037/neu0000901.

Measurement precision across cognitive domains in the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset

Paul K. Crane, MD MPH[1,*], **Seo-Eun Choi, PhD**[1], **Michael Lee, PhD**[1], **Phoebe Scollard, MA**[1], **R. Elizabeth Sanders, BA**[1], **Brandon Klinedinst, PhD**[1], **Connie Nakano, MPH**[1], **Emily H. Trittschuh, PhD**[2], **Jesse Mez, MD MS**[3], **Andrew J. Saykin, PsyD**[4], **Laura E. Gibbons, PhD**[1], **Chun Wang, PhD**[5], **Dan Mungas, PhD**[6], **Ruoyi Zhu, MS**[5], **Nancy S. Foldi, PhD**[7], **Melissa Lamar, PhD**[8], **Roos Jutten, PhD**[9], **Sietske A.M. Sikkes, PhD**[10], **Evan Grandoit, PhD**[11], **Laura A. Rabin, PhD**[12], **Richard N. Jones, ScD**[13], **Doug Tommet, MS**[13] **Alzheimer's Disease Neuroimaging Initiative**

Shubhabrata Mukherjee, PhD[1]

[1]Department of Medicine, University of Washington, Seattle, WA, USA

[2]Department of Psychiatry and Behavioral Sciences, University of Washington, and VA Puget Sound Health Care System, Geriatrics Research, Education, and Clinical Core (GRECC), both in Seattle, WA, USA

[3]Department of Neurology, Boston University, Boston, MA, USA

[*]Correspondence concerning this article should be addressed to: Paul K. Crane, MD MPH, Box 359780, 325 Ninth Avenue, Seattle, WA 98104 pcrane@uw.edu.

Author Contributions

Paul Crane – Conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, supervision, validation, visualization, writing – original draft, writing – review and editing.
Seo-Eun Choi - Conceptualization, data curation, formal analysis, investigation, methodology, software, visualization, writing – review and editing.
Michael Lee - Conceptualization, data curation, formal analysis, investigation, methodology, software, visualization, writing – review and editing.
Phoebe Scollard -- Conceptualization, data curation, formal analysis, investigation, methodology, software, visualization, writing – review and editing.
Elizabeth Sanders -- Conceptualization, data curation, formal analysis, investigation, methodology, software, visualization, writing – review and editing.
Brandon Klinedinst – Writing – review and editing.
Connie Nakano – Writing – review and editing.
Emily Trittschuh - Conceptualization, data curation, investigation, methodology, supervision, writing – review and editing.
Jesse Mez -- Conceptualization, data curation, investigation, methodology, supervision, writing – review and editing.
Andrew Saykin -- Conceptualization, data curation, investigation, methodology, supervision, writing – review and editing.
Laura Gibbons -- Conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, software, supervision, validation, visualization, writing – review and editing.
Chun Wang – Conceptualization, methodology, resources, supervision, visualization, writing – review and editing
Dan Mungas – Conceptualization, methodology, resources, supervision, visualization, writing – review and editing
Ruoyi Zhu– Conceptualization, methodology, visualization, writing – review and editing
Nancy Foldi – Conceptualization, writing – review and editing
Melissa Lamar -- Conceptualization, writing – review and editing
Roos Jutten -- Conceptualization, writing – review and editing
Sietske Sikkes -- Conceptualization, writing – review and editing
Evan Grandoit -- Conceptualization, writing – review and editing
Laura Rabin -- Conceptualization, writing – review and editing
Richard Jones -- Conceptualization, methodology, writing – review and editing
Doug Tommet -- Conceptualization, methodology, writing – review and editing
Shubhabrata Mukherjee – Conceptualization, data curation, formal analysis, investigation, methodology, supervision, visualization, writing – review and editing

We have no conflict of interest to disclose.

[4]Indiana University Alzheimer's Center, Indianapolis, IN, USA

[5]College of Education, University of Washington, Seattle, WA, USA

[6]Department of Neurology, University of California at Davis, Sacramento, CA, USA

[7]Department of Psychology, Queens College and The Graduate Center, City University of New York & Department of Radiology, Brain Health Imaging Institute, Weill Cornell Medicine, New York, NY

[8]Rush University Alzheimer's Center and the Department of Psychiatry and Behavioral Sciences, Rush University Medical Center, Chicago, IL, USA

[9]Department of Neurology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA

[10]Alzheimer Center Amsterdam, Neurology, Vrije Universiteit Amsterdam, Amsterdam UMC Location, VUmc & Amsterdam Neuroscience, Neurodegeneration & Department of Clinical, Neuro, and Developmental Psychology, Vrije Universitiet Amsterdam, Amsterdam, NL

[11]Department of Psychology, Northwestern University, Chicago, IL, USA

[12]Department of Psychology, Brooklyn College and the Graduate Center, City University of New York, Brooklyn, NY, USA

[13]Department of Psychiatry, Brown University, Providence, RI, USA

Abstract

Objective: To demonstrate measurement precision of cognitive domains in the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset.

Method: Participants with normal cognition (NC), mild cognitive impairment (MCI), and Alzheimer's disease (AD) were included from all ADNI waves. We used data from each person's last study visit to calibrate scores for memory, executive function, language, and visuospatial functioning. We extracted item information functions for each domain and used these to calculate standard errors of measurement. We derived scores for each domain for each diagnostic group and plotted standard errors of measurement for the observed range of scores.

Results: Across all waves, there were 961 people with NC, 825 people with MCI, and 694 people with AD at their most recent study visit (data pulled February 25, 2019). Across ADNI's battery there were 34 memory items, 18 executive function items, 20 language items, and 7 visuospatial items. Scores for each domain were highest on average for people with NC, intermediate for people with MCI, and lowest for people with AD, with most scores across all groups in the range of −1 to +1. Standard error of measurement in the range from −1 to +1 was highest for memory, intermediate for language and executive functioning, and lowest for visuospatial.

Conclusions: Modern psychometric approaches provide tools to help understand measurement precision of the scales used in studies. In ADNI, there are important differences in measurement precision across cognitive domains.

Keywords

cognition; measurement precision; psychometrics

Introduction

Measurement precision refers to how close measurements of the same individual would be to each other on repeated measurement with the same instrument. Measurement precision is related to measurement error; the more precise a measurement, the lower the measurement error. In the case of cognitive tests, concerns about participant burden and retest effects result in measurement with each instrument once per participant per encounter. This is in contrast to practice in other branches of science, where replicate measurements are incorporated in routine workflows to directly determine whether scientific conclusions are impacted by the precision or imprecision of the measurements. In cognitive testing, measurement precision typically represents a thought experiment regarding how similar repeated measures would be if the same individual was tested with the same conditions, including the same experience (or lack of experience) with the instrument (i.e., without retest effects).

In physical sciences, many instruments are characterized by having equal levels of measurement precision across the entire dynamic range of the instrument. With a meter stick with equally spaced 1 mm gradations, an object that is about 20 cm will be measured with essentially the same precision as an object that is about 80 cm (Figure 1, meter stick A). With cognitive tests, in contrast, the instruments we use to measure each cognitive domain may be characterized by uneven measurement precision, where the measurement precision for people at different ability levels may differ because of the items included in the instrument.

Imagine a second meter stick with 1 mm gradations from 0 to 50 cm, and then 1 cm gradations from 50 cm to 1 m (Figure 1, meter stick B). With meter stick B, an object that is about 20 cm would be measured very precisely while an object that is about 80 cm would be measured much less precisely. The density of the markings on the meter stick near the size of the object determines the measurement precision for that object, whether the object is measured a single time or repeatedly.

Imagine a third meter stick with 2 mm gradations from 0 to 30 cm, 5 mm gradations from 30 to 70 cm, and 1 cm gradations from 70 cm to 1 m (Figure 1, meter stick C). When measuring the object that is about 20 cm, meter stick C has gradations of 2 mm near that size, while meter stick B had gradations of 1 mm near that size. Meter stick C will have lower measurement precision for an object about 20 cm long compared to meter stick A or meter stick B. When measuring an object about 80 cm long, both meter sticks B and C have gradations of 1 cm near that size. The measurement precision for objects that size is identical (and not great!) for meter sticks B and C. When we communicate the size of the item measured with a meter stick, it is useful to have an understanding of the measurement precision of our meter stick specifically for items around that size.

The situation with cognitive testing is somewhat analogous to physical measurement with meter sticks, but using latent factors estimated from groups of indicators. Modern psychometrics provides a toolkit to help understand measurement precision in this more abstract area of science.

Elsewhere in this special issue we have demonstrated how we use modern psychometric methods to co-calibrate measures of domains such as memory, executive functioning, language, and visuospatial abilities across studies that administered different cognitive testing batteries. Those methods result in scores on the same metric ("co-calibrated") across studies even when the studies use different specific instruments to measure a domain.

In this paper we focus on measurement precision across different domains. We chose the Alzheimer's Disease Neuroimaging Initiative (ADNI) to illustrate these points. ADNI administers a cognitive battery at each study visit. The battery includes several tests, and granular data are available from the LONI website (http://adni.loni.usc.edu/).

We have previously used modern psychometric approaches to develop composite scores from the ADNI battery for memory (Crane, Carle, Gibbons, Insel, Mackin, Gross, Jones, Mukherjee, Curtis, & Harvey, 2012), executive functioning (Gibbons et al., 2012), and language and visuospatial functioning (Choi et al., 2020).

Our previous papers on these scales from ADNI focused on each domain in turn. The present paper specifically emphasizes measurement precision of four different domains included in the battery, providing further insights into the relative performance of the different domains as assessed in ADNI. Since the ADNI battery is widely used and its specific tests are widely familiar to neuropsychologists, this will provide an introduction to considerations of measurement precision that can then be useful in considering co-calibrated data from a single domain across multiple studies.

Several decades ago, Chapman and Chapman noted the importance of considering measurement properties across different domains in clinical neuropsychological practice (L.J. Chapman & Chapman, 1978). Patterns of findings across multiple domains can be influenced by the psychometric properties of the scales used to measure each domain, and furthermore by the relative measurement properties across domains. Studies that make use of all four of our composite scores for the ADNI battery should be informed by an understanding of the relative measurement precision of the composite scores.

Modern Psychometric Approach

Modern psychometric theory has been the dominant paradigm in educational testing settings since at least the 1960s with the publication of Lord and Novick's highly influential book (Lord & Novick, 1968). We provide an extended discussion of modern psychometric approaches and measurement precision in Supplemental Materials 1.

If we consider a test taker of a particular ability level, a scale that has many items of difficulty levels close to that ability level will provide a lot of measurement precision for that test taker, while a second scale that has few such items will provide much less measurement

precision. This is directly analogous to the discussion of the three meter sticks above. In modern psychometrics, measurement precision is quantified as *test information*.

While practitioners and students of modern psychometric theory are taught to have intuition regarding information content and information curves, others not trained in this way may find it difficult to consider quantifications of measurement precision with completely unfamiliar units. Fortunately, there is a mathematical relationship between information and the standard error of measurement (SEM), which is on the same scale as the test score and may be more intuitive

Again considering tests administered to a test taker of a particular ability level, the test with many items of an appropriate difficulty level will have a lot of information at that ability level, and a correspondingly low SEM, while a test with few items of an appropriate difficulty level will have a small amount of information at that ability level, and a high SEM.

If SEMs are large this would lead to concern that lack of measurement precision could be influencing results – analogous to a meter stick with gradations that are spread far apart near the relevant ability level. At the same time, if SEMs are small, analogous to a meter stick with gradations that are close together near the relevant ability level, this can provide reassurance to investigators finding differences in associations across domains. Additional information about modern psychometrics and SEMs can be found in Supplemental Materials.

Goals of this paper

The goals of this paper are to briefly review the items included in each of the four domain scores for the ADNI study, and to compare measurement precision quantified as the standard error of measurement across four cognitive domains.

Method

Participants

Data used in the preparation of this article were obtained from the ADNI database; additional details can also be found at this site (https://adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, see www.adni-info.org. We used data from all waves of the ADNI study. As outlined in our companion paper (Mukherjee et al. this issue), we used cognitive data from the most recent visit for each participant. In the context of cognitive decline over time that is associated with AD this strategy maximizes the distribution of cognitive abilities while limiting the analytic sample to a single observation per person

ADNI has had several funding cycles, with somewhat differing enrollment goals. ADNI 1 in particular enrolled people with normal cognition, MCI, and AD in a 1:2:1 ratio. ADNI

enrolled participants from AD-specialty clinic research settings across the United States and Canada. Participants consented to research including an extensive array of neuroimaging studies. Higher proportions of ADNI participants have advanced levels of education and a higher proportion of the cohort self-reports non-Hispanic white race compared to the general population of older adults (Petersen et al., 2010). Subsequent enrollment waves had different goals. We analyzed data from ADNI1, ADNIGO, and ADNI2 for this paper.

Procedure

Study staff administered cognitive tests to all participants at every ADNI study visit. We downloaded item-level data from the LONI website. As outlined in our companion paper (Mukherjee et al. this issue) our expert panel considered each item and assigned it to a specific domain (memory, executive function, language, or visuospatial, or none of these) and theory-driven sub-domain (e.g. working memory, phonemic fluency). ADNI was one of the three studies we used in our initial calibration step, referred to as one of the “legacy” studies in our companion paper (Mukherjee et al. this issue). Across ADNI’s battery there were 34 memory items, 18 executive function items, 20 language items, and 7 visuospatial items. We detail each of the items in each of the domains in the Supplement for this paper, which was duplicated from the Supplement of our companion paper (Mukherjee et al. this issue).

Measures

We have previously published detailed methods and validation results for composite scores for memory (Crane, Carle, Gibbons, Insel, Mackin, Gross, Jones, Mukherjee, Curtis, & Harvey, 2012), executive function (Gibbons et al., 2012), and more recently language and visuospatial (Choi et al., 2020). The psychometric methods here were nearly identical. The one exception is the treatment of the fluency items. Phonemic and semantic fluency items were treated as indicators of executive function in our earlier work (Gibbons et al., 2012), but now are assigned to the language domain (Choi et al., 2020). The four cognitive scores we generated here are thus mutually independent of each other, with no overlapping item content.

Determination of item information content and domain standard error of measurement

As outlined in our companion paper (Mukherjee et al. this issue), for three of the four domains, we used bifactor models to co-calibrate data from the legacy studies including ADNI. We used Mplus for our calibrations (Muthen & Muthen, 1998-2012). Mplus exports graphs of item information functions for each item that loads on the general factor. If we consider the multiple learning trials for the Rey Auditory Verbal Learning Test (RAVLT) as independent indicators of memory without accounting for their commonality, the strength of association with memory will be inflated, and this would be reflected in information content that was too large. But we used bifactor models and included a secondary factor to address the methods-specific correlation for those memory indicators. The information data output by Mplus for the memory domain (and for all of the bifactor models we employed) accounts for the secondary domain structure.

We extracted the data for item information from Mplus. Due to local independence, we are able to add item information together to obtain total test information. This additive property on the information scale makes it easier to deal directly with information until the very end, when we can take the inverse square root of the information to arrive at the standard error of measurement.

Distribution of domain scores in the ADNI cohort

Each participant in the ADNI study is categorized into mutually distinct categories of cognitive and functional status at every study visit—normal cognition (NC), mild cognitive impairment (MCI), or Alzheimer's disease (AD). Based on ADNI's operational definitions of each of these categories, for ADNI waves with additional diagnoses, we mapped early MCI and subjective impairment categories to normal cognition, and late MCI to MCI (https://adni.loni.usc.edu). We plotted scores for each of these groups for each domain to determine the region of the ability scale where scores were most commonly observed.

Standard error of measurement and Illustration of implications of modifying a scale

We then plotted the standard error of measurement for each domain, focusing on the region of the ability scale where scores were most commonly observed in ADNI.

To illustrate the usefulness of information and standard errors of measurement in thinking about potential revisions to a scale, we focused on the visuospatial domain. There are several ways one could improve the measurement precision of a domain. One way would be to add additional items. This would add to staff time needed to administer the battery, and thus costs of data collection per visit. Furthermore, this would also add respondent burden. Most study participants are fully aligned with the scientific goals of the study and happy to participate. Some people, however, might be barely tolerating study procedures, and would be potentially very unhappy with additional cognitive assessment, even possibly dropping out of the study if burden is perceived to be too high. Another way to improve measurement precision is to consider scoring the same assessment tool in a way that provides additional gradations.

To make this concrete, we consider scoring for the interlocking pentagons item. Folstein used this item in the MMSE, where correct copying of the image results in 1 point and any other response results in 0 points (Folstein, Folstein, & McHugh, 1975). This 0/1 scoring is what is available in the ADNI dataset. Teng et al. used this same item in the Modified MMSE (Teng & Chui, 1987) and in the Cognitive Abilities Screening Instrument (Teng et al., 1994). Their scoring, however, was out of 10 points – 0 to 4 points for the left pentagon, 0 to 4 points for the right pentagon, and 0 to 2 points for the overlapping quadrilateral. Manuals for the 3MS and the CASI describe specific criteria that are used for scoring on this 0-10 point scale.

It should be noted that changing the scoring for the interlocking pentagons item would also have costs associated with it. Someone (or some computer algorithm) would need to rescore all of those items. The drawings of pentagons would need to be located and delivered to the scoring center for scoring. Care would need to be taken to ensure that the correct study ID and visit were applied to the correct drawing. Quality control procedures would need to be

developed for the new scoring, and care taken in data entry. On the other hand, since this is rescoring as opposed to new data collection, this approach would not be associated with increased respondent burden or staff time for data collection.

To illustrate the ways one could use modern psychometric tools to understand the implications of modified scoring of an item, we considered the visuospatial domain. We used co-calibrated item parameters for the 0/1 version of the pentagons item (as available from ADNI). We also used co-calibrated item parameters for the 0-10 version of the item used in the CASI and the 3MS from another study (see Mukherjee et al., this issue). We then consider two potential versions of ADNI's visuospatial domain: one with binary scoring for the pentagons item (ADNI's scoring) and one with the ordinal scoring for the pentagons item. For this comparison, we plotted the overall standard error of measurement for the binary version and for the 0-10 version of the pentagons item.

We used all data from ADNI1, ADNI2, amd ADNIGO for our analyses. The overall sample size was around 3,000, well over those found to be necessary for similarly complex models(Jiang, Wang, & Weiss, 2016). We report all data exclusions (if any), all manipulations, and all measures in the study, and we follow JARS (Kazak, 2018). Data are available from http://adni.loni.usc.edu/. Analysis code and research materials are available on request. Data were managed with Stata (StataCorp, 2018) and analyses were performed in Stata (StataCorp, 2018) and in Mplus (Muthen & Muthen, 1998-2012). This study's design and its analysis were not pre-registered.

Results

There were 961 people who had a diagnosis of NC at their most recent ADNI study visit, there were 825 people with MCI, and there were 773 people with AD. Clinical diagnosis was missing on 457 individuals. Demographic and clinical characteristics for the ADNI sample evaluated here are shown in Table 1.

Scores for the most recent study visit across all waves are shown in Figure 2. The scoring metric is arbitrary. It is scaled such that the distribution of the latent ability level in the population has a mean of 0 and SD of 1. Clear differences are found in scores across each diagnostic group, as expected, with the highest scores for people with normal cognition, intermediate scores for people with MCI, and lowest for people with AD. Also as expected these contrasts were clearest for memory, and the contrasts were somewhat attenuated for visuospatial functioning. As shown with the blue shaded box in Figure 2, for each domain the bulk of the scores were found in the region bounded by +1 and −1.

We plotted SEM curves for each domain as shown in Figure 3. Figures 2 and 3 capitalize on the fact that ability and item difficulty are on the same scale in modern psychometrics; we can go back and forth between Figures 2and 3 to further our understanding of measurement properties (the SEM data plotted in Figure 3) specifically in those regions of the ability metric where ADNI participants' ability levels lie (the box plots in Figure 2). The blue boxes in both Figure 2 and Figure 3 in the regions from −1 to +1 help to cement this relationship between person ability levels and test or item difficulty levels. Figure 3 demonstrates that

the extensive battery of measurement items led to excellent measurement precision for the memory domain, as shown in the blue curve at the bottom of the graph. In the −1 to +1 region highlighted by the blue shaded box, the SEM for the memory domain approaches 0.10, meaning that the information content in this region is approaching 100, which is a tremendously high level of information.

The SEMs for the language (green curve in Figure 3) and executive function (red) domains are intermediate between memory and visuospatial. Considering the threshold of 0.30 units discussed earlier, the executive function domain has better measurement precision than that for the entire region between +1 and −1, but the language domain has greater SEM to the right side of that region; its SEM climbs above 0.30 at an ability level of about +0.2.

The SEM for the visuospatial domain has an SEM that is below 0.3 only for a small region, from about −1.6 to −1.1. Considering the orange box plots in Figure 2, we can see that no one with normal cognition had a visuospatial score this low, very few people with MCI had a visuospatial score this low, and a small minority of people with AD had a visuospatial score this low (the left side of the box in the box-and-whiskers plot denotes the 25th percentile). In the entire region highlighted by the blue shaded box, the SEM for the visuospatial domain is quite high. Indeed, somewhere around −0.2, values for the SEM for the visuospatial domain climb higher than 0.60.

At higher levels of visuospatial ability, the two information curves overlap each other. This makes sense as a score of 1 on binary scoring corresponds exactly to a score of 10 on ordinal scaling for this item. At the top end, there are no inputs that could improve the precision provided by this item. However, at lower levels of visuospatial functioning, ordinal scoring provides better measurement precision, as shown by the higher levels of information.

The Y axis for this graph is the information scale, which has the inverse square root relationship with standard error of measurement discussed earlier. While ordinal scoring for the pentagons item is more informative than binary scoring for the pentagons item, it still is not a very informative item, as the total information peaks at less than 0.6 on the information metric.

We can compare measurement precision for the entire visuospatial ability domain with the binary version and the ordinal version of the pentagons item. The test information functions for the scale with the binary version and the ordinal version are shown in the left panel of Figure 5, and the corresponding standard error of measurement curves are shown in the right panel of Figure 5.

In essence, ordinal scoring for the pentagons item provides more categories to differentiate among people who would receive a 0 for their pentagon effort under binary scoring. We would thus expect that the incremental measurement precision from ordinal scoring compared with binary scoring would be for people with lower level of visuospatial ability, and that is what we see in Figure 4. Because of a property called local independence, we are able to simply add item information levels together to obtain a test information curve, as shown in the left panel of Figure 5. The small magnitude of the incremental effect of ordinal as compared to binary scoring for the pentagons item is illustrated there. The

information content scale is harder to wrap one's head around than the standard error of measurement scale; the standard error of measurement has the distinct advantage of being on the same metric as the measurement itself. The right panel of Figure 5 shows the same small magnitude of the incremental effect of ordinal as compared to binary scoring for the pentagons item, but on the standard error of measurement scale. Note that the binary version of the standard error of measurement graph is the same as that shown in Figure 3. Taken together, these results suggest that if there were minimal costs to switch to ordinal scoring for the pentagons item for ADNI, there's every reason to do so, but that this would not be expected to have a major impact on improving measurement precision for the visuospatial domain.

It may be useful to consider the analogy of the three different meter sticks discussed in the Introduction (Figure 1). Differences between the density of markings in one part of the scale (around 20 cm, for example) have little to do with the measurement precision of a different part of the scale (around 80 cm, for example). In this instance, at the lower end of the scale, moving from dichotomous to polytomous scoring makes a very modest impact on measurement precision. But at the higher end of the scale, this move makes no impact at all.

Discussion

We found quantitatively and qualitatively very different measurement precision across the four cognitive domain composite scores in ADNI. The memory domain was characterized by excellent measurement precision in the entire ability region. The language and executive function domains had intermediate measurement precision, with the SEM for executive function lying below the 0.3 level across the entire +1 to −1 region and that for language lying below that level below about 0.2 and somewhat higher than that level above 0.2. The SEM for visuospatial was quite different than the SEMs for the other three domains, with values above 0.30 in the entire −1 to +1 region, and even above 0.6 (the highest value shown in the graph) for the region above about −0.2.

Measurement properties for these domains in ADNI

There is considerable research interest in characterizing multiple cognitive domains in studies such as ADNI that focus on Alzheimer's dementia, MCI, and imaging and fluid biomarkers. Heterogeneity across domains among people with AD is an important research priority, and has been recognized clinically (Lam, Masellis, Freedman, Stuss, & Black, 2013). The findings reported here provide reassurance for use of composite scores for memory, language, and executive function from the ADNI study. The low levels of the SEM in the region where the bulk of the scores were found across all waves of the ADNI study provide reassurance to investigators finding differences in associations across different domains.

The story is quite different for the visuospatial domain. Visuospatial has poor measurement precision in the region where the bulk of the ADNI cohort had their scores. Investigators finding an association of some factor of interest with memory but failing to find an association with visuospatial should consider the possibility that poor measurement precision for the visuospatial domain may be an explanation for this pattern of findings.

Our experience with data from other studies evaluating cognition in older adults suggests that poor measurement precision for the visuospatial domain is a general issue, not at all specific to ADNI (see Mukherjee et al. paper in this volume). Investigators designing measurement strategies for future waves of the ADNI study, or those designing cognitive batteries for studies based on ADNI's cognitive battery or for clinical trials, could consider augmenting the assessment of visuospatial by adding additional items from the visuospatial item bank (see Mukherjee et al. paper). The tools of modern psychometrics provide a rational basis to identify items that could be added to an assessment that would have the greatest increment in measurement precision over a particular range of ability levels.

These results of different levels of measurement precision across domains is a variant of what Chapman and Chapman warned about several decades ago (L.J. Chapman & Chapman, 1978). Differential measurement properties across domains can influence conclusions one could draw at the individual level. Here we also have a similar concern at the group level, that is, that conclusions drawn about the lack of relationship with visuospatial may be driven by the measurement properties for visuospatial in ADNI.

There is another important finding for visuospatial ability to notice in Figure 2, which is that the right side of the box plots all max out at a value of just over 1.0. This is the highest score observed in the ADNI study, corresponding to correct responses to every indicator of visuospatial ability used in ADNI, also known as a ceiling effect. Indeed, the box and whisker plot for people with normal cognition (the yellow box at the bottom of the top four boxes) shows that the median value for visuospatial scores for that group was at the ceiling. Our companion paper on the Framingham study discusses ceiling effects at some length (see Scollard et al. this issue). Ceiling effects are very challenging for modeling changes over time. Ceiling effects are not the same thing as poor measurement precision. The visuospatial domain in ADNI appears to have both poor measurement precision in the ability region where the vast majority of participants had scores, AND to have ceiling effects such that the median score for those with normal cognition was at the ceiling.

Modern psychometric tools to promote rational test (re-)design

Item response theory, the "modern" psychometric method underlying this study, was first articulated nearly 70 years ago and has had a major impact on educational testing and high stakes aptitude and achievement testing. These methods have not been widely used in neuropsychology, although there are notable exceptions. The Spanish and English Neuropsychological Assessment Scales (Mungas, Reed, Crane, Haan, & Gonzalez, 2004; Mungas, Reed, Marshall, & Gonzalez, 2000) used IRT methodology for rational test design applied to newly created item pools to create scales that met goals of psychometric matching described by Chapman and Chapman (Loren J. Chapman & Chapman, 1973). Similar IRT methods were applied to pre-existing tests to create matched episodic memory and executive function scales (Mungas, Reed, & Kramer, 2003). More recently, IRT methods figured prominently in development of the NIH Toolbox Cognitive Health Battery (Dikmen et al., 2014; Gershon et al., 2014; Weintraub et al., 2013).

This study builds on the available literature applying IRT to neuropsychological tests. The modern psychometric methods we used to calibrate data from ADNI provide helpful tools

to aid in rational test design. For example, investigators responsible for the design of the ADNI cognitive assessments could decide that they would like better measurement precision for the visuospatial domain. One option would be to switch from binary to ordinal scoring for the pentagons item. Our illustration of this scenario in Figures 4 and 5 suggests that there would be improved measurement precision if one integrated ordinal scoring instead of binary scoring for the pentagons item, but that the overall effect of such a switch on measurement precision for the visuospatial domain would be fairly small.

Another option would be to simply add additional visuospatial measures. Item information is known for all items in our growing item bank. The same tools used to develop the data for Figures 4 and 5 could show the impact of incorporating additional items with known properties into the battery.

A related option would be to trade off measurement precision in a different domain to avoid increasing respondent burden, which can be approximated by the time of test administration. The data summarized here suggests that the memory domain in particular could be a useful target for trimming. The same simple calculations that would indicate how much additional precision would be obtained by adding additional visuospatial items can be used to ensure that a proposed revised memory battery still would have sufficient measurement precision even after removing specific memory items.

We should point out that there is no such thing as “too much measurement precision.” Other things being equal, we would always advocate for longer scales and more testing. However, participants have a limited willingness to endure psychometric testing, and staff time to administer cognitive testing items is carefully budgeted in research studies. Indeed, cognitive testing is among the most burdensome aspects cited by patients about their involvement in trials (Ottenhof et al., 2021). Given these tradeoffs, the information and SEM tools of modern psychometric methods provide a useful and valuable way to rationally propose to revise measurement strategies to meet the study’s goals.

Implications for research

Beyond considerations for potential modifications to how a study measures each cognitive domain, these psychometric tools enhance understanding of measurement across studies. Each score from the harmonization and co-calibration efforts we have made is accompanied by a standard error that quantifies our certainty of each measurement. This information may be useful in identifying individual observations characterized by marginal measurement precision that could be excluded to ensure that conclusions were not driven by those observations in particular. Furthermore, understanding measurement precision of any particular domain across studies will enable further understanding of cross-study findings.

Implications for clinical practice

Earlier we highlighted Chapman and Chapman’s insights several decades ago on the importance of considering measurement properties in understanding each individual’s relative performance across cognitive domains, which in turn is critical in formulating differential diagnosis. In this paper we have focused on the relevance of measurement

precision particularly for research. But these concepts may indeed be relevant and useful to clinicians seeking to evaluate a single patient.

Limitations

Our study has limitations that should be considered. We took great pains with our modeling as outlined in our companion paper (Mukherjee et al. this issue). The information content and the SEM curves shown here are derived from those models, so mis-specification of these models would almost certainly result in incorrect values for the information and the directly related SEM. The ADNI cohort has limited racial and ethnic diversity. Proposed enrollment for the next wave of ADNI emphasizes increased diversity, and it will be important to ensure that these scales work equivalently across groups. While it is beyond the scope of the present investigation, there are extensive tools to address test bias or differential item functioning incorporated in the modern psychometric toolkit.

Conclusion

In conclusion, in this paper we have provided some discussion of measurement precision, information, and SEMs in understanding measurement of cognitive domains, using the widely used ADNI study as an example. Investigators using cognitive composite scores we developed should be confident using scores for memory, language, and executive function, but caution is warranted for the visuospatial domain. The tools of modern psychometric theory are very useful in visualizing important measurement properties across cognitive domains. These tools also can be used for rational test design or to determine the measurement properties of revised cognitive batteries that could be proposed.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

Data collection and sharing for this project were funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.;Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.;Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

R01 AG029672 (P Crane, PI)

K25 AG055620 (S Mukherjee, PI)

U19 AG066567 (E Larson, P Crane, A Lacroix, MPI)

P30 AG066509 (T Grabowski, PI)

P30 AG010133 (A Saykin, PI)

R01 AG019771 (A Saykin, PI)

U01AG006781 (E Larson, P Crane, MPI)

SC3GM122662 (N Foldi, PI)

R15AG066039 (L Rabin, PI)

References

Chapman LJ, & Chapman JP (1973). Disordered thought in schizophrenia. NY: Appleton-Century-Crofts.

Chapman LJ, & Chapman JP (1978). Measurement of differential deficits. J Psychiatr Res, 14, 303–311. [PubMed: 722633]

Choi SE, Mukherjee S, Gibbons LE, Sanders RE, Jones RN, Tommet D, … Crane PK (2020). Development and validation of language and visuospatial composite scores in ADNI. Alzheimers Dement (N Y), 6(1), e12072. doi:10.1002/trc2.12072 [PubMed: 33313380]

Crane PK, Carle A, Gibbons LE, Insel P, Mackin RS, Gross A, … Mungas D (2012). Development and assessment of a composite score for memory in the Alzheimer's Disease Neuroimaging Initiative (ADNI). Brain Imaging Behav, 6(4), 502–516. doi:10.1007/s11682-012-9186-z [PubMed: 22782295]

Crane PK, Carle A, Gibbons LE, Insel P, Mackin RS, Gross A, … Harvey D (2012). Development and assessment of a composite score for memory in the Alzheimer's Disease Neuroimaging Initiative (ADNI). Brain Imaging and Behavior, 6(4), 502–516. [PubMed: 22782295]

Crane PK, Narasimhalu K, Gibbons LE, Mungas DM, Haneuse S, Larson EB, … van Belle G (2008). Item response theory facilitated cocalibrating cognitive tests and reduced bias in estimated rates of decline. J Clin Epidemiol, 61(10), 1018–1027 e1019. [PubMed: 18455909]

Dikmen SS, Bauer PJ, Weintraub S, Mungas D, Slotkin J, Beaumont JL, … Heaton RK (2014). Measuring episodic memory across the lifespan: NIH Toolbox Picture Sequence Memory Test. J Int Neuropsychol Soc, 20(6), 611–619. doi:10.1017/S1355617714000460 [PubMed: 24960230]

Folstein MF, Folstein SE, & McHugh PR (1975). "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res, 12(3), 189–198. [PubMed: 1202204]

Gershon RC, Cook KF, Mungas D, Manly JJ, Slotkin J, Beaumont JL, & Weintraub S (2014). Language measures of the NIH Toolbox Cognition Battery. J Int Neuropsychol Soc, 20(6), 642–651. doi:10.1017/S1355617714000411 [PubMed: 24960128]

Gibbons LE, Carle AC, Mackin RS, Harvey D, Mukherjee S, Insel P, … Crane PK (2012). A composite score for executive functioning, validated in Alzheimer's Disease Neuroimaging Initiative (ADNI) participants with baseline mild cognitive impairment. Brain Imaging Behav, 6(4), 517–527. doi:10.1007/s11682-012-9176-1 [PubMed: 22644789]

Jiang S, Wang C, & Weiss DJ (2016). Sample size requirements for estimation of item parameters in the multidimensional graded response model. Front Psychol, 7, 109. doi:10.3389/fpsyg.2016.00109 [PubMed: 26903916]

Kazak AE (2018). Editorial: Journal article reporting standards. American Psychologist, 73(1), 1–2. doi:10.1037/amp0000263 [PubMed: 29345483]

Lam B, Masellis M, Freedman M, Stuss DT, & Black SE (2013). Clinical, imaging, and pathological heterogeneity of the Alzheimer's disease syndrome. Alzheimers Res Ther, 5(1), 1. doi:alzrt155 [pii] 10.1186/alzrt155 [PubMed: 23302773]

Lord FM, & Novick MR (1968). Statistical theories of mental test scores, with contributions by Allan Birnbaum. Reading, MA: Addison-Wesley.

McNeish D, & Wolf MG (2020). Thinking twice about sum scores. Behav Res Methods, 52(6), 2287–2305. doi:10.3758/s13428-020-01398-0 [PubMed: 32323277]

Mungas D, Reed BR, Crane PK, Haan MN, & Gonzalez H (2004). Spanish and English Neuropsychological Assessment Scales (SENAS): further development and psychometric characteristics. Psychol Assess, 16(4), 347–359. [PubMed: 15584794]

Mungas D, Reed BR, & Kramer JH (2003). Psychometrically matched measures of global cognition, memory, and executive function for assessment of cognitive decline in older persons. Neuropsychology, 17(3), 380–392. [PubMed: 12959504]

Mungas D, Reed BR, Marshall SC, & Gonzalez HM (2000). Development of psychometrically matched English and Spanish language neuropsychological tests for older persons. Neuropsychology, 14(2), 209–223. [PubMed: 10791861]

Muthen LK, & Muthen BO (1998-2012). Mplus user's guide (7 ed.). LA: Muthen & Muthen.

Ottenhof L, Vijverberg EGB, Visser LNC, Verrijp M, Prins ND, van der Flier WM, & Sikkes SAM (2021). Can we improve clinical trial design in Alzheimer's disease? The participants point of view. Retrieved from https://alz.confex.com/alz/2021/meetingapp.cgi/Person/53650

Petersen RC, Aisen PS, Beckett LA, Donohue MC, Gamst AC, Harvey DJ, … Weiner MW (2010). Alzheimer's Disease Neuroimaging Initiative (ADNI). Clinical characterization. Neurology, 74(3), 201–209. [PubMed: 20042704]

Proust-Lima C, Amieva H, Dartigues JF, & Jacqmin-Gadda H (2007). Sensitivity of four psychometric tests to measure cognitive changes in brain aging-population-based studies. Am J Epidemiol, 165(3), 344–350. [PubMed: 17105962]

StataCorp. (2018). Stata: Release 15.1 Statistical Software. College Station, TX: StataCorp LP.

Teng EL, & Chui HC (1987). The Modified Mini-Mental State (3MS) examination. J Clin Psychiatry, 48(8), 314–318. [PubMed: 3611032]

Teng EL, Hasegawa K, Homma A, Imai Y, Larson E, Graves A, … et al. (1994). The Cognitive Abilities Screening Instrument (CASI): a practical test for cross-cultural epidemiological studies of dementia. Int Psychogeriatr, 6(1), 45–58; discussion 62. [PubMed: 8054493]

Weintraub S, Dikmen SS, Heaton RK, Tulsky DS, Zelazo PD, Bauer PJ, … Gershon RC (2013). Cognition assessment using the NIH Toolbox. Neurology, 80(11 Suppl 3), S54–64. doi:10.1212/WNL.0b013e3182872ded [PubMed: 23479546]

Key Points

Question:

How do ADNI's cognitive domains compare in terms of measurement precision?

Findings:

Memory is characterized by better measurement precision, and visuospatial by worse measurement precision, with intermediate values for language and executive function, in the range where scores were observed in ADNI.

Importance:

Measurement properties such as measurement precision may be useful in interpreting findings from ADNI, and may be useful in management of burden / precision trade-offs for researchers designing cognitive assessment approaches

Next Steps:

Familiarity with measurement precision issues and metrics may be useful in understanding data from existing studies and in designing cognitive evaluation strategies for future studies.

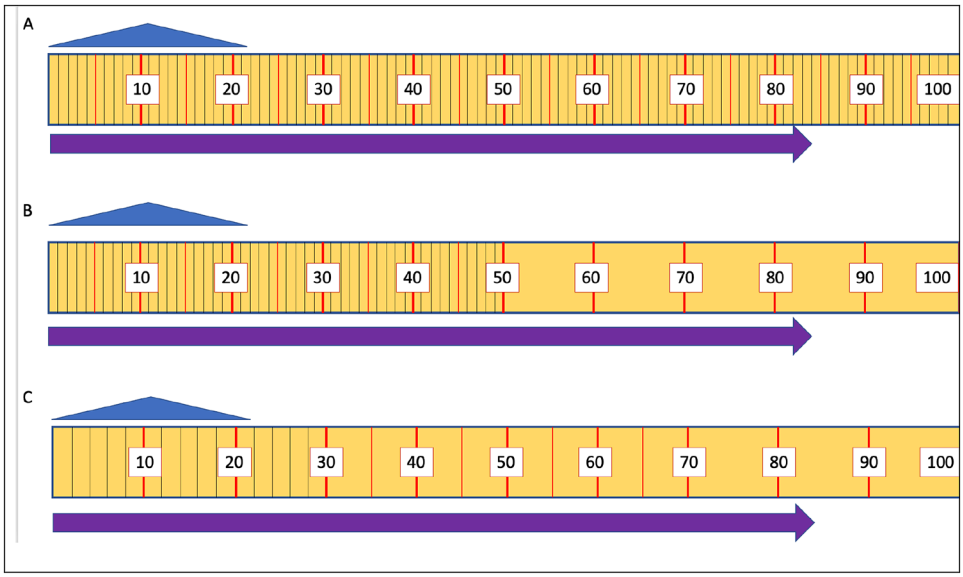

Figure 1.
Three meter sticks with varying measurement precision

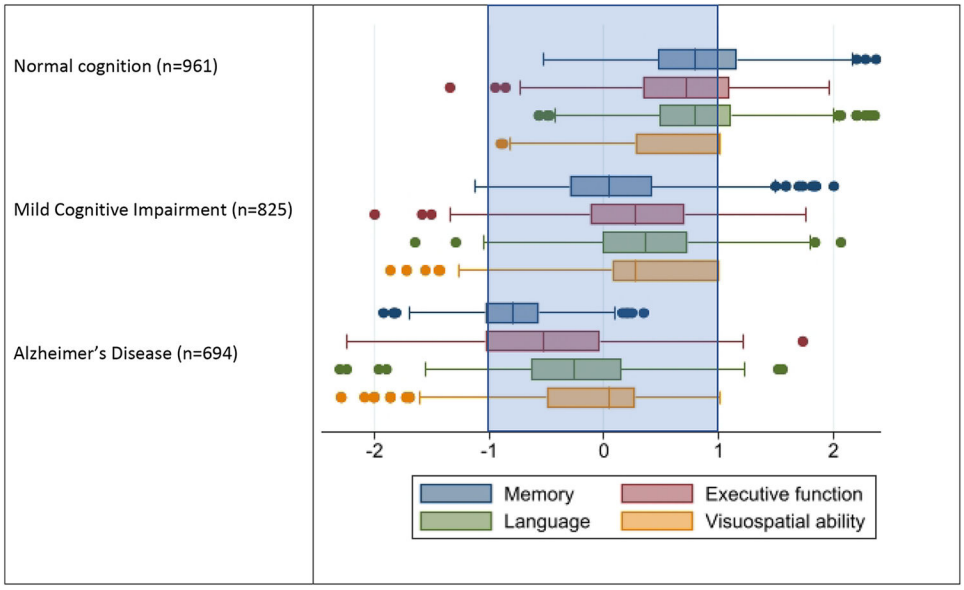

Figure 2.

Distribution of scores for memory, executive function, language, and visuospatial for people with normal cognition, mild cognitive impairment, and Alzheimer's disease at their most recent ADNI study visit across all waves*

* The blue shaded box from −1 to +1 shows the region where the bulk of the scores were clustered across the different diagnostic categories. The x axis is the ability level for each domain.

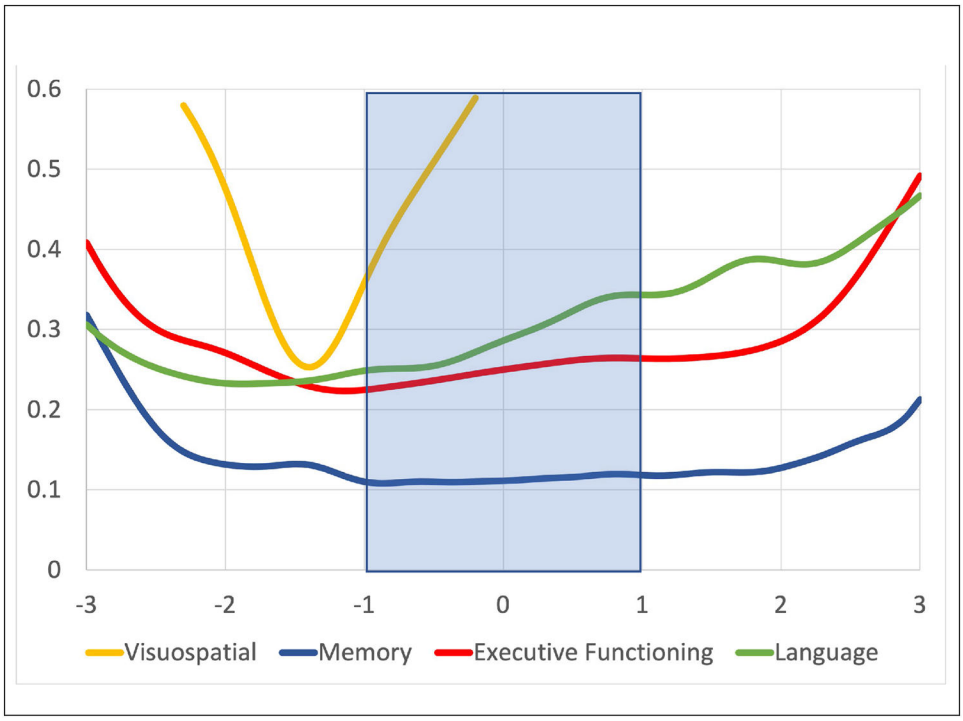

Figure 3.

Standard error of measurement curves for each domain*

* X axis: Ability level. Y axis: Standard error of measurement

Information curves for binary and ordinal scoring for the pentagons item are shown in Figure 4.

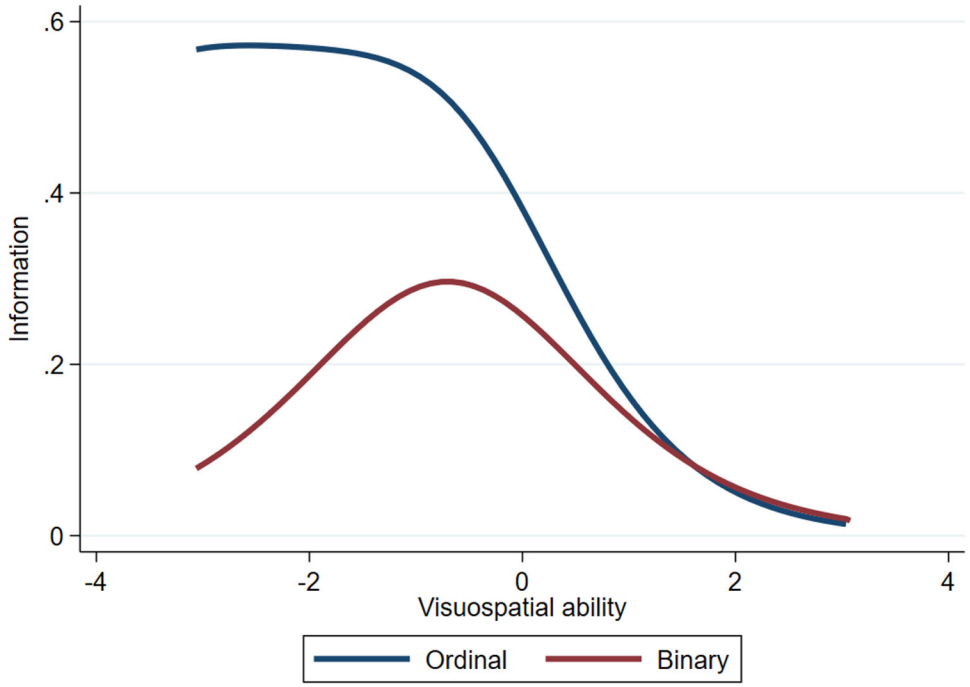

Figure 4.
Item information curves for binary and ordinal scoring of the interlocking pentagons item

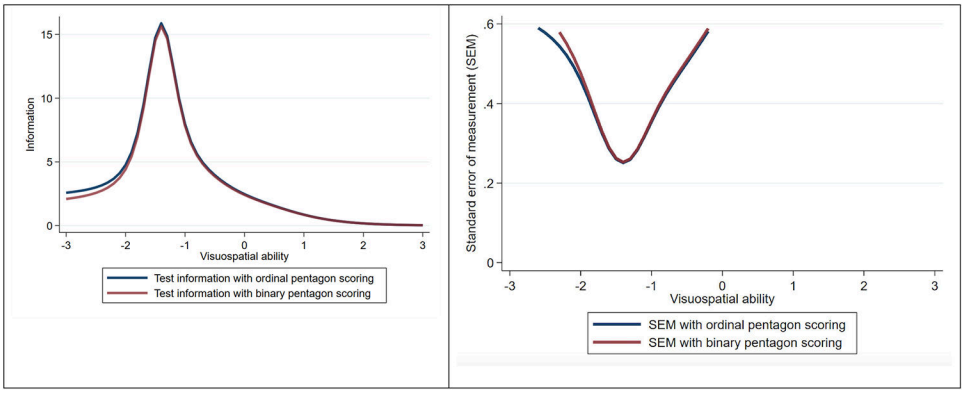

Figure 5.

Test information curves for the visuospatial domain with binary and ordinal scoring for the pentagons item (left panel) and the corresponding test standard error of measurement curves (right panel).

Table 1

Demographic and clinical characteristics of the ADNI study at the most recent study visit (n=3,016).

	Mean (SD) or n (%)
Age, mean (SD)	74.9 (SD = 8.7)
Female, n (%)	1,494 (48.1%)
Education, mean (SD)	16.1 (SD= 2.8)
Self-reported race, n (%)	
White	2,760 (91.5%)
African/American	142 (4.7%)
Others	115 (3.8%)
Cognitive DX at most recent visit, %	
Cognitively Normal	1,155 (38.3%)
MCI	994 (32.0%)
Diagnosed with AD	896 (29.7%)

A few individuals were missing age, sex, education, or race. Clinical diagnosis was missing on 457 individuals.