Published in final edited form as:

J Clin Epidemiol. 2008 October ; 61(10): 1018–27.e9. doi:10.1016/j.jclinepi.2007.11.011.

Item response theory facilitated co-calibrating cognitive tests and reduced bias in estimated rates of decline

Paul K. Crane^{*}, Kaavya Narasimhalu^{*}, and Laura E. Gibbons^{*}

^{*}Department of Medicine, University of Washington. pcrane@u.washington.edu; nkaavya@gmail.com, gibbonsl@u.washington.edu

Dan M. Mungas[†]

[†]Department of Neurology, University of California at Davis. dmmungas@ucdavis.edu.

Sebastien Haneuse[‡] and Eric B. Larson[‡]

[‡]Center for Health Studies, Group Health Cooperative. haneuse.s@ghc.org, larson.e@ghc.org

Lewis Kuller[§]

[§]Department of Epidemiology, University of Pittsburgh. kullerl@edc.pitt.edu

Kathleen Hall^{||}

^{||}Department of Psychiatry, Indiana University - Purdue University in Indianapolis. khall@iupui.edu

Gerald van Belle^{**}

^{**}Departments of Biostatistics and Environmental and Occupational Health Sciences, University of Washington. vanbelle@u.washington.edu

Abstract

Objective—To co-calibrate the Mini-Mental State Examination (MMSE), the Modified Mini-Mental State (3MS), the Cognitive Abilities Screening Instrument (CASI) and the Community Screening Instrument for Dementia (CSI`D') using item response theory (IRT) to compare screening cut-points used to identify cases of dementia from different studies, to compare measurement properties of the tests, and to explore the implications of these measurement properties on longitudinal studies of cognitive functioning over time.

Study design and setting—We used cross-sectional data from three large (n>1000) community-based studies of cognitive functioning in the elderly. We used IRT to co-calibrate the scales and performed simulations of longitudinal studies.

Results—Screening cut-points varied quite widely across studies. The four tests have curvilinear scaling and varied levels of measurement precision, with more measurement error at higher levels of cognitive functioning. In longitudinal simulations IRT scores always performed better than standard scoring, while a strategy to account for varying measurement precision had mixed results.

Conclusion—Co-calibration allows direct comparison of cognitive functioning in studies using any of these four tests. Standard scoring appears to be a poor choice for analysis of longitudinal

Corresponding author and reprint requests: Paul K. Crane, MD MPH, Box 359780, Harborview Medical Center, 325 Ninth Avenue, Seattle, WA 98104. Phone (206) 744-1831. Fax (206) 744-9917. Email pcrane@u.washington.edu.

Author contributions: Conception and design: Crane. Acquisition of data: Larson, Kuller, Hall. Analysis and interpretation of data: Crane, Narasimhalu, Gibbons, Mungas, Haneuse, van Belle. Drafting of manuscript: Crane. Revision of manuscript for critical intellectual content: Narasimhalu, Gibbons, Mungas, Haneuse, Larson, Hall, Kuller, van Belle. Statistical analysis: Crane, Narasimhalu, Gibbons. Obtaining funding: Crane. Administrative, technical, or material support: Crane, Larson, Hall, Kuller. Supervision: Crane, Gibbons. Dr. Crane had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. No author has a conflict of interest relevant to the subject of the manuscript.

cognitive testing data. More research is needed into the implications of varying levels of measurement precision.

Keywords

Cognition; co-calibration; item response theory; psychometrics; longitudinal data analysis; simulation

What is new?

- Co-calibration of widely used cognitive tests facilitates comparisons across studies that use different tests. This is a necessary first step for pooling cognitive data across studies.
- Screening cut-points varied dramatically across studies.
- These tests as usually scored and analyzed have curvilinear scaling, which caused bias in estimated rates of change in cognition over time. Item response theory scoring or some other appropriate strategy should be used rather than standard scores in longitudinal analyses of cognitive testing data.
- These tests have similar measurement precision profiles: reasonable measurement precision at the low end (poor cognitive ability) and poor at the high end (high cognitive ability). This also caused bias in estimated rates of change over time. Further research is needed to identify optimal strategies to deal with varying levels of measurement precision in longitudinal analyses.

Introduction

Tests of global cognitive functioning are widely used in clinical and research studies to screen for and monitor cognitive impairments. The Mini-Mental State Examination (MMSE), developed in 1975, was introduced in one of the most frequently cited papers in the medical literature (1). The MMSE or similar tests have been recommended for clinically identifying patients with mild cognitive impairment (2), though most subjects with mild cognitive impairment will have MMSE scores in the normal range (3). Longer versions of the MMSE have been developed, including the Modified Mini-Mental State (3MS) (4), the Cognitive Abilities Screening Instrument (CASI) (5), and the Community Screening Instrument for Dementia (CSI 'D') (6,7). Some of the large epidemiological studies employing these tests are outlined in Table 1. These studies represent a major societal investment in measuring cognitive functioning in the elderly.

Unfortunately, the different tests are not interchangeable. The MMSE has 20 questions and is scored out of 30 points. The 3MS and the CASI have 40 questions and are scored out of 100 points; the scores are not equivalent. The CSI 'D' has 45 questions and is scored out of 34 points. At present, cross-study analyses can be performed only on studies that used the same test. While a vast amount of data has been collected, until these cognitive tests are co-calibrated with each other, it is as if the tests were written in different languages, and there is no bilingual dictionary.

This situation is similar to the ancient Greek and Egyptian languages before the decipherment of the Rosetta Stone, a granite block inscribed with three scripts discovered in 1799. Scholars immediately recognized one script as Greek, used to record a proclamation. In 1824, Jean-François Champollion demonstrated that the other two scripts recorded the same proclamation in the ancient Egyptian language. Once the different languages and scripts were co-calibrated,

it was possible to develop a bilingual dictionary, laying foundations for our knowledge of ancient Egyptian culture.

The National Institutes of Health has formed a trans-NIH Project on Cognitive and Emotional Health whose Critical Evaluation Study Committee identified 66 papers from 36 longitudinal studies of cognitive functioning with at least 500 subjects (8). The Committee stressed “there is no agreement on the questionnaires used ... making comparisons between studies and combining data from studies difficult.” They advocated development, implementation, and analysis of data collected by a new cognitive and emotional questionnaire for use in subsequent research (8). In effect, this strategy calls for discarding the hundreds of thousands of person-years of data that have already been collected. The item response theory (IRT) co-calibration scheme presented here provides a viable alternative to that strategy. Results can be combined across studies and valid findings can be drawn from existing data once the scales have been co-calibrated.

Even within a study that uses a single test, there may be important challenges to using standard test scoring in longitudinal analyses. The simplest example of standard scoring is a sum score for a test in which each right/wrong item is worth one point, though the same arguments apply for tests with more complicated items and/or more complicated scoring rules. The distribution of item difficulties is not considered in standard scoring. If it happens that a test contains many easy items and few hard items, this fact is not reflected in scores. As we would hope, subjects with greater ability would be expected to have higher scores than subjects with lower ability levels at a single time point. However, with the passage of time, standard scoring of such a test may show strange results for the amount of change. Subjects with lower ability levels have more items whose difficulty level is around their ability level, so their scores are estimated with more precision. Small drops in cognitive ability for such an individual are more likely to be detected by changing from success to failure on several items. Subjects with higher ability levels have fewer items with difficulty levels close to their ability level, so their scores may appear to be more stable over time, at least initially, since changes in cognitive functioning will be less likely to be reflected with changing from success to failure on items. This problem of non-linear measurement properties of standard scoring of global cognitive tests has been discussed previously (9).

An alternative scoring technique is to use IRT. IRT scoring does not assume a pre-specified weighting for the items. Instead, the data are used to determine parameters for the difficulty of each item and, in some models, discrimination (the strength of the relationship between probability of success on the item and the underlying trait or ability measured by the test). IRT can detect the situation discussed above, with several easy items and few hard items. IRT scoring takes the difficulty of test items into account, resulting in a metric with linear scaling properties (10).

An additional potential problem may arise in the analysis of longitudinal data from a test with many easy items and a few hard items. Standard analytic strategies implicitly assume that measurement precision is constant. However, in our example test, measurement precision for people with low ability levels is greater than for people with higher ability levels. Standard approaches do not account for varying levels of measurement precision.

We reviewed the papers identified by the trans-NIH Project on Cognitive and Emotional Health Critical Evaluation Study Committee. None of these papers used IRT or otherwise accounted for non-linearity of cognitive tests. None of the studies summarized in Table 1 account for varying measurement precision at all. Two studies mentioned by the Committee treat measurement error as a constant across the entire measurement spectrum (A47,A48; references preceded by A can be found in Appendix 3). One additional paper mentions measurement error

in the discussion but does not account for measurement error in the analytic plan (A49). The remaining 63 papers do not mention measurement precision at all (A1-A4,A11,A13,A14,A16-A19,A25-A28,A32-A35,A41,A49-A91).

Our primary objective was to co-calibrate the 3MS, CASI, CSI `D', and MMSE using IRT. We illustrate our success by showing on the same co-calibrated scale the cut points used in several studies that employed different tests. Our secondary objective was to discuss measurement properties of the tests and illustrate the potential implications of varying levels of measurement precision on longitudinal studies of cognitive change over time.

Methods

Overview

We present two separate analyses. First, we analyzed real global cognitive test data from 3 large studies to co-calibrate the tests using IRT. Once we co-calibrated the tests, we compared screening cut-points published in the literature on the common co-calibrated scale. We also determined measurement properties of the tests. In the second analysis, we used these results to inform simulated longitudinal studies of cognitive functioning over time. We compared four different scoring strategies in terms of their bias in estimating the true rate of cognitive decline over time.

Analysis 1. Co-calibration of global cognitive functioning tests

Study populations: We used cross-sectional data from 3 studies, the Cardiovascular Health Study (CHS) (n = 4,978), the Adult Changes in Thought Study (ACT) (n = 3,358), and the Indianapolis site from the Indianapolis - Ibadan Dementia Project (Indianapolis) (n = 1,254) (total n = 9,590). Detailed methods from each of these studies have been published (11-14). Local institutional review boards approved each study, and written informed consent was obtained in each study.

CHS enrolled 5,201 individuals aged 65 years or older from four communities between 1989 and 1990; 4,291 participated in the study in 1992-1993. An additional 687 African-American participants were enrolled in 1992-1993. We analyzed 3MS responses from 1992-1993 from these 4,978 individuals.

ACT enrolled 2,554 individuals aged 65 years or older in 1994-1996 from a large health maintenance organization. An additional 811 subjects were enrolled in 2000-2002; 804 had valid CASI scores. We analyzed the most recent CASI results from both cohorts (total n = 3,358).

Indianapolis enrolled 2,147 African-Americans aged 65 years or older in 1992-1993. Of those, 1,254 had CSI `D' data from the second incidence wave of data collection in 1997-1998 and are included here.

Statistical analysis: Initial evaluation of dimensionality employed McDonald's bi-factor model (15). The scales were sufficiently unidimensional to proceed with IRT (Appendix 1).

Co-calibration requires either the same people taking different tests or different tests sharing common items (16); we used common items. We identified anchor items with identical content across tests and ensured that their relationship with the underlying ability tested was the same across study sites. These items were then used to anchor the scales to a common metric.

Candidate anchor item identification: We compared test items to identify those that presented subjects with identical stimuli. If necessary, we re-coded scoring categories. For example,

interlocking pentagons was scored 0 or 1 in the CSI 'D', 0-10 points in the CASI, and 4, 4, and 2 points in the 3MS for the left pentagon, right pentagon, and intersection. Based on scoring rules used in the studies, a score of 1 from the CSI 'D' corresponded to a score of exactly 10 from the CASI, and to scores of exactly 4, 4, and 2 from the 3MS. We re-coded CASI scores as 10 = 1, any other score = 0, and re-coded 3MS scores as (4, 4, and 2) = 1, any other score combination = 0.

Many items were similar but not identical; such items were not considered as candidate anchor items. For example, the CSI 'D' uses the words "boat", "house", and "fish" while the 3MS and CASI use the words "shirt", "brown", and "honesty" for short-term recall.

IRT calibration: Each item from a test has relationships with all of the candidate anchor items from that test (within-test relationships), and the candidate anchor items have relationships with candidate anchor items in the other tests (between-test relationships). We used IRT to parameterize these relationships. We generated a PARSCALE data set (17) containing item responses from all 9,590 subjects; we used Samejima's graded response model for polytomous items (18,19).

Assessment of candidate anchor items for differential item functioning related to study:

Anchor items must have the same relationship with cognitive functioning across study populations. Violation of this assumption is called differential item functioning (DIF), defined as statistical differences across groups (in this case, across studies) in item responses when controlling for the underlying ability measured by the test (20-22). Details of the approach we used to identify items ineligible to be anchor items due to DIF (23,24) are presented in Appendix 2.

Final co-calibration of scores of the four cognitive tests: Samejima's graded response model provides a formula for the probability of each response category for each item for any level of cognitive functioning (18,19,25). We used this formula to determine the most likely response for every item for every cognitive functioning level. We calculated 3MS, CASI, CSI 'D', and MMSE scores using published scoring algorithms (4,5,7). When MMSE item parameters were available from multiple tests (which could happen only if an MMSE candidate anchor item was found to have DIF related to study site), we used parameters from the 3MS. We rescaled IRT scores by multiplying them by 15 (so a standard deviation unit is 15 points) and adding 100 (so average cognitive functioning is 100 points) (26).

Measurement properties of the tests: IRT provides two important summaries of test measurement properties. The *test characteristic curve* is a plot of the most likely score associated with each level of cognitive functioning (10,27). It is useful for assessing whether the relationship between standard scores and the underlying level of cognitive functioning is linear (9). Linear relationships are important for many applications (10). The *test information curve* depicts the measurement precision of the test at each level of cognitive functioning, which may vary. If a cognitive test includes many difficult items it will have high information (good precision) for individuals with above-average levels of cognitive functioning; if it includes few difficult items it will have less information (poor precision) for individuals with above-average levels of cognitive functioning (10,27). The standard error of measurement is proportional to the inverse square root of the information, and is on the same scale as the cognitive functioning level. We used standard formulas (27) to plot test characteristic curves and standard errors of measurement for the MMSE, 3MS, CASI, and CSI 'D'.

Comparison of cut-points used in screening tests: We reviewed the publications listed in Table 1 to determine if a 2-stage sampling design was used and, if so, what screening cut-point was used in each study. We plotted these values on the transformed co-calibrated metric.

Analysis 2: Simulation study of cognitive decline over time—We examined 5 different scenarios for change, each with 250 data sets of 1000 subjects seen at baseline and every 2 years for 8 years (5 data points for each subject in each data set). We drew a random intercept (baseline value) and slope (rate of decline) for each subject in each data set. We varied the mean slope and intercept across five scenarios. In each scenario the standard deviation of the intercept term was 15 and the standard deviation of the slope was 7.5 points over 8 years. This process generated true cognitive abilities for each subject at each time point in each data set. We simulated item responses to a global cognitive test for each subject based on their true cognitive ability at each time point. We chose item parameters for the simulated test to mimic the test characteristic and standard error curves obtained from Analysis 1 (see Figures 2 and 3). Further details are provided in the footnote to Table 4.

We used four different strategies to score the observed item responses: standard total score, naïve IRT score, a single perturbed IRT score in which noise proportional to measurement error is added to each observation, and 10 perturbed IRT scores defined the same way. For IRT scores we used PARSCALE (17) and the graded response model (18,19); we estimated item parameters from the baseline data anew for each run. These parameters were used to compute IRT scores at each follow-up time point. For the perturbed IRT scores we used the point estimate of the score from PARSCALE and added a perturbation term generated by multiplying a random normal $(0, 15^2)$ variable by the standard error of measurement term from PARSCALE.

We fit mixed effects models (28) for each scoring strategy to estimate the rate of change. We defined percent bias as the difference between the actual rates of change for the true scores and the estimated rates of change for each scoring strategy divided by the actual rate of change for the true scores. We split the data sets in half and numbered them from 1-125 to determine the running mean bias, defined for data set n as $(\text{total bias for data sets 1 to } n)/n$. We plotted running mean bias against run number to determine whether the bias estimates were converging to the same number in the two halves of the data.

Results

Analysis 1. Co-calibration of cognitive functioning tests

Table 2 shows demographic characteristics of the study populations. The age and gender distributions of the three study populations were roughly similar, but there were large differences in ethnic and educational distributions.

The items included in the three tests and their final dispositions (anchor items, rejected anchor items due to DIF, or unique) are shown in Appendix Table 4. Two items served as anchors across all three tests, sixteen items anchored the comparison of the 3MS and the CASI, four items anchored the comparison of the 3MS and the CSI 'D', and two items anchored the comparison of the CASI and the CSI 'D'.

Table 3 summarizes the result of the co-calibration. Score comparisons can be made by reading across the rows of the table. For example, scores of 20 on the MMSE correspond to scores of 51 or 52 from the 3MS, scores of 73 or 74 from the CASI, and scores of 23 from the CSI 'D'.

The screening cut-points used to identify subjects with poor cognitive functioning in selected epidemiological studies are shown on the IRT metric in Figure 1. The cut-points used vary dramatically, from a low around $3\frac{1}{2}$ standard deviations below average (for the Chicago Healthy Aging Project and the Kinmen Study) to just about average (for the Women's Health Initiative Memory Study).

In Figure 2 we show test characteristic curves for the MMSE, the CASI, the 3MS, and the CSI 'D'. The test characteristic curve for a scale with linear measurement properties would have a straight line (9). All of these cognitive screening tests produce non-linear test characteristic curves, with steeper slopes at below-normal levels of cognitive functioning and shallow slopes at normal and above-normal levels. All of these tests thus are characterized by having relatively few hard items and more easy items. A test with a curvilinear test characteristic curve is poorly suited for comparing trajectories of cognitive functioning across patients who start from different levels of cognitive ability. A patient who started in a steep portion of the curve would be expected to have a larger change in observed score for a given amount of change in cognitive functioning than a patient who started in a shallow portion of the curve. For longitudinal epidemiological and clinical studies (including those referred to in Table 1), tests with curvilinear test characteristic curves may threaten the validity of study results if standard scores are used (29).

In Figure 3 we show each test's standard error of measurement at each cognitive functioning level. For individuals with cognitive functioning levels below average (below 100), the curves for the different tests are nearly identical. For individuals with cognitive functioning levels above average (above 100) the standard error curves start to rise dramatically. Because the MMSE is a shorter test, its measurement precision is inferior to that of the other tests for individuals with average or above-average cognitive functioning. None of the tests has good measurement precision for these individuals. Between one standard deviation and two standard deviations above average (115-130 on the transformed cognitive functioning scale), each of the tests has a large standard error of measurement of at least one standard deviation (15 points). We performed simulation analyses to illustrate potential problems that may arise from this profile of varying levels of measurement precision.

Analysis 2: Simulation study of changes in cognitive functioning over time

Results from the simulation studies are shown in Table 4. Standard scoring performed poorly in three of the five scenarios (2, 4, and 5). Accounting for curvilinear scaling properties using naïve IRT scores provided more accurate estimates of the rate of change than standard scores in all five scenarios, though the difference was negligible in scenario 1. Further accounting for varying levels of measurement precision using the perturbation strategy proved to be the most accurate strategy for two scenarios (scenarios 2 and 4), but this strategy was not as accurate as naïve IRT scores for the other three scenarios and not as good as standard scores for scenarios 1 and 3. There was no difference in findings when the perturbation strategy was carried out once per data set or 10 times. Running means converged in the two halves of the data at the global means shown in Table 4 (see Appendix Figure 1).

Discussion

We co-calibrated four commonly used tests of global cognitive functioning using IRT. Just as the deciphered Rosetta Stone permitted understanding of the Egyptian language by scholars familiar with Greek, co-calibration permits understanding across studies that use different cognitive tests. Scores from each test can be directly compared to scores on the other tests. This permits comparison of cut-points used in different studies, which we found to vary dramatically across studies. Co-calibration also permits us to compare measurement properties such as curvilinearity of the test characteristic curve and standard errors of measurement of the different tests on the same co-calibrated metric. We found that all four tests had similar curvilinear test characteristic curves, as well as poor measurement precision for high scores. We performed a simulation study to investigate the potential for bias due to curvilinear measurement properties and varying levels of measurement precision in longitudinal studies of cognitive change over time. Simulation results suggest that IRT scores are always an

improvement over standard scores in recovering true rates of change over time. Further accounting for measurement error using a perturbation term strategy had varying effects on the amount of bias in estimating the rate of change over time.

The co-calibration performed here will permit direct comparison of scores across studies that employed different tests. Table 3 should thus prove useful to clinicians and researchers. One application of these co-calibrated scores is the ability to compare cut-points used across studies that used different instruments, which we found varied dramatically. Studies with higher cut-points will have detected a higher proportion of their subjects who have dementia (a higher sensitivity) at the expense of having performed many more evaluations (a lower specificity) than studies with lower cut-points. Because of the differences in cut-points, subjects with dementia identified in studies with widely divergent screening cut-points may not be comparable to each other. Subjects with mild forms of dementia are much more likely to escape detection in a study with a lower cut-point than in a study with a higher cut-point.

The curvilinear test characteristic curves of the cognitive tests (see Figure 2) create formidable impediments to analysis of cognitive changes that attempts to use traditional scoring of data from these tests. Regression and change score approaches to analyzing cognitive trajectories assume that each such trajectory is linear - that is, a change of a few points at the top end of the scale has the same implication for cognitive functioning as a change of the same few points at the bottom end of the scale (10). None of the cognitive tests considered here meets this standard. Changes of a few points at the top end of the scale imply vast differences in cognitive functioning, while changes of a few points at the bottom end of the scale imply tiny differences. Given the curvilinear nature of cognitive functioning as measured by these tests, analyses of changes in cognitive functioning over time should use IRT scoring rather than standard scoring (9,10). These theoretical considerations are supported by our simulation results from Analysis 2, in which using naïve IRT scores was always more accurate than using standard scores. Thus instead of using the standard score equivalents from Table 3, the wisest choice would be to use IRT to obtain scores on the co-calibrated metric. Item parameters are available from the first author for these tests. It should be recalled that none of the studies identified by the Project on Cognitive and Emotional Health used IRT or any other technique appropriate for curvilinear metrics (A1-A4,A11,A13,A14,A16-A19,A25-A28,A32-A35,A40,A41,A47-A91).

Measurement precision also varied dramatically across the cognitive functioning spectrum for the cognitive screening tests that we examined, with no major differences between the tests (Figure 3). For all four tests, measurement precision is much poorer at the higher end of the cognitive functioning spectrum because of small numbers of difficult items. This lack of high-end sensitivity for all of these tests has been noted especially in conjunction with the interest in detecting early cognitive deficits such as mild cognitive impairment (MCI) (2). None of these tests should be solely relied upon for identification of early cognitive deficits, because none of them has much measurement precision at the higher end of the scale. Furthermore, as we show with our simulations, ignoring varying levels of measurement precision in an analysis of change over time may also lead to biased estimates of the rate of change. In two of the five scenarios, our perturbation strategy proved better than naïve IRT scoring in recovering the true rate of change over time. However, in the other three scenarios, naïve IRT scoring was better, and in two of the scenarios, standard scoring itself proved to be superior to the perturbation strategy. We think there are two implications of these findings. The first is that the standard error of measurement curves shown in Figure 3 provide qualitatively different information than the test characteristic curves shown in Figure 2. The second implication of these findings - especially the variability of the findings across scenarios - is that more research is needed before we can provide general guidelines on strategies to account for varying levels of measurement precision. Thus, while measurement error was ignored by almost all of the studies identified by the Project on Cognitive and Emotional Health (A1-A4,A11,A13,A14,A16-A19,A25-

A28,A32-A35,A41,A49-A91), treated as a constant (A47,A48), or mentioned only in the discussion (A40), at present we have nothing better to offer than naïve IRT scores (a strategy employed by none of these studies). This is an active area of our ongoing research.

The perturbation strategy employed in Analysis 2 seems counterintuitive. Rather than relying on the best point estimate of each subject's score at each time point, we actually introduce noise to that point estimate in the form of the perturbation term, which in turn reduces bias in estimating rates of change. The perturbation strategy is directly analogous to the plausible values strategy employed since the 1983-1984 school year in the National Assessment of Educational Progress (NAEP) (30). This strategy is analogous to the multiple imputation framework for missing data (31,32). In essence, every cognitive functioning estimate is treated as missing, and information is drawn from both the point estimate of the score and the certainty of that score. In our simulations and in most longitudinal studies of global cognitive functioning in the elderly, subjects on average move from a cognitive functioning level measured with less measurement precision to a cognitive functioning level measured with more precision. By incorporating knowledge of measurement precision into our analytic strategy, we try to avoid biased estimates of the rate of change. We found no difference between a single perturbed score and 10 perturbed scores, but this was averaged over 250 datasets. In any particular data set, multiple perturbations will likely reduce random fluctuations. Other potential strategies for handling varying levels of measurement precision include multi-level IRT approaches (33-39) in which a measurement model on one level is used to estimate ability, which in turn may be used in a mixed effects model at another level to estimate change over time. Further research is needed to determine optimal strategies for assessing change over time in studies that incorporate tests with widely varying measurement precision.

This study has limitations. The validity of our findings in Analysis 1 hinges on the validity of the anchor items. We reinforced our confidence in the anchor items by retaining only those that had no DIF related to study site. To our knowledge, this is the first co-calibration study to include the step of checking anchor items for DIF related to study site. Demographic characteristics for the subjects in the parent studies were very different from each other, and further studies should be performed to confirm the stability of the anchor items' item parameters across heterogeneous populations. Further targeted data collection would increase our confidence in the item parameters. The CSI `D' was only anchored to the other tests by 8 items. An additional 6 potential anchors for the CSI `D' were found to have DIF. Adding a few items from the CSI `D' to a study that routinely uses the CASI, for example, would dramatically increase the proportion of items that could be used to anchor the two tests. We also did not consider DIF related to anything other than study site in this analysis. We have previously found DIF related to age and education in the CASI (23) and the MMSE (40,41), and would be surprised to not discover similar relationships in the 3MS or the CSI `D'. Analysis 2 is based on modest and realistic amounts of cognitive decline over time, but further simulation work is needed prior to recommending a strategy for handling varying levels of measurement precision.

In summary, we have co-calibrated widely-used screening tests of cognitive functioning. Strategies outlined here can be used to permit data to be combined across studies that used different cognitive tests. Co-calibration can serve as a Rosetta Stone to facilitate understanding across studies that otherwise would have no way to communicate. Standard analytical approaches to longitudinal cognitive testing data appear to be fundamentally flawed because they ignore curvilinear test characteristic curves common to these tests (Figure 2) and varying levels of measurement precision (Figure 3).

Acknowledgements

Drs. Crane and Gibbons and Ms. Narasimhalu were funded by NIH K08 AG 022232 from the National Institute on Aging. Dr. Crane was also supported by a New Investigator Research Grant from the Alzheimer's Association (Crane). Drs. Gibbons and van Belle were supported by grant P50 AG05136 from the National Institute on Aging. The CHS research reported in this article was supported by contracts N01-HC-85079 through N01-HC-85086, N01-HC-35129, N01-HC-15103, N01-HC-55222, and U01 HL080295 from the National Heart, Lung, and Blood Institute, National Institutes of Health, and grants AG15928 and AG 20098 from the National Institute on Aging, National Institutes of Health. A full list of participating CHS investigators and institutions can be found at <http://www.chsnhlbi.org>. The ACT research reported in this article was supported by contract U01-AG-06781 from the National Institute on Aging. The Indianapolis research reported in this article was supported by R01 AG 09956 from the National Institute on Aging. None of these sponsors had any role in the design of the present analyses, selection of these particular studies for co-calibration, data management, data analysis, or interpretation. None was involved in the preparation of the manuscript. The manuscript was reviewed by the Cardiovascular Health Study which includes review by the National Heart, Lung, and Blood Institute, which approved the manuscript.

Appendix 1. Methods and results for the assessment of unidimensionality in the 3MS, CASI, and CSI `D'. (Includes Appendix Tables 1-3)

We used McDonald's bi-factor model (15) to assess whether the items in these tests of global cognitive functioning were sufficiently unidimensional to use the tools of IRT. For both the CASI and the CSI `D', specific domains are mentioned in the original papers describing the scales (5,7). For the 3MS, we used modifications of the CASI domain structure, since both tests were written by Evelyn Teng and test content is largely overlapping between the 3MS and the CASI. Since the MMSE is fully contained within the other tests we did not perform any analyses specifically on the MMSE itself, reasoning that if the 3MS (say) was sufficiently unidimensional to use IRT, any subset of it (including the MMSE) would also be sufficiently unidimensional to use IRT.

In the bi-factor model, each item is modeled to have loadings on two factors: a global cognitive functioning factor (all the items), and a sub-domain factor (such as long-term memory, orientation, attention, etc.). The sub-domains are specified to be mutually uncorrelated and also uncorrelated with the global cognitive functioning factor. The variance for all factors is set to 1.0 to obtain standardized loadings. According to McDonald, the key result is the loadings on the general factor. If all the items have salient loadings (which he defines as standardized loadings >0.30) on the general factor, McDonald states that the items are "sufficiently homogeneous" for applications requiring homogeneity (such as IRT).

We used MPLUS (42) for all factor analyses. We used the tetrachoric/polychoric correlation matrix and employed the weighted least squares with adjustments for mean and variance (WLSMV) estimator appropriate for categorical data.

McDonald does not focus on model fit in his discussion of the bi-factor model (though he emphasizes the importance of model fit elsewhere in his book on test theory). For the 3MS, the Confirmatory Fit Index was low at 0.659, but the Tucker-Lewis Index (TLI) was very good at 0.973, and the root mean squared error of approximation was very good at 0.036.

Appendix Table 1 shows loadings for the 3MS.

Appendix Table 1
Bi-factor model results for the 3MS.

Item	Loading, General Cognitive Function factor	Loading, sub-domain factor	Sub-domain name
Birth year	0.67	0.44	Long term memory
Birth date	0.59	0.57	
Birth month	0.52	0.52	
Birth State	0.86	0.29	
Birth Town	0.98	0.32	
DATE	0.58	0.60	Orientation
MONTH	0.70	0.57	
YEAR	0.82	0.51	
SEASON	0.66	0.26	
WEEK	0.64	0.40	
STATE	0.96	0.09	
COUNTY	0.81	0.15	
CITY	0.91	0.21	
HOSPITAL	0.81	0.14	
SHOES	0.53	0.43	Attention
BROWN	0.76	0.58	
HONESTY	0.79	0.51	
REPEAT	0.76	0.13	
NOIFS	0.54	0.47	
ANDS	0.62	0.64	
BUTS	0.64	0.60	
COUNTSCO	0.90	n/a	Concentration
SPELLSCO	0.63	n/a	
SHOES1	0.51	0.55	Short term memory
BROWN1	0.61	0.60	
HONESTY1	0.54	0.50	
SHOES2	0.56	0.62	
BROWN2	0.64	0.65	
HONESTY2	0.65	0.51	
FOREHEAD	0.67	0.52	Language
CHIN	0.86	0.50	
SHOULDER	0.65	0.47	
ELBOW	0.79	0.54	
KNUCKLE	0.63	0.36	

Item	Loading, General Cognitive Function factor	Loading, sub-domain factor	Sub-domain name
OBEY	0.77	0.30	
WRTWLD	0.72	0.48	
WRTLK	0.72	0.51	
WRTTO	0.73	0.56	
WRTGO	0.71	0.58	
WRTOUT	0.74	0.57	
LEFT	0.39	0.24	
FOLD	0.71	0.27	
HAND	0.38	0.28	
PENT1	0.62	0.56	
PENT2	0.64	0.62	Visual construction
INTER	0.69	0.54	
LEG	0.90	n/a	Fluency
ARMLEG	0.52	0.40	
LAUGHING	0.60	0.62	Abstraction and judgment
EATSLEEP	0.51	0.48	

The standardized loading for each of the items on the general factor is greater than 0.30. There are only two loadings on the general factor in the 0.30s and none in the 0.40s.

Appendix Table 2 shows the standardized loadings for the CASI. Fit statistics for these analyses included a good CFI of 0.922, a good TLI of 0.969, and a good RMSEA of 0.027.

Appendix Table 2
Bi-factor model results for the CASI.

Item	Loading, general cognition factor	Loading, sub-domain factor	Name of sub-domain
Birth place	0.65	0.07	
Birth year	0.90	0.13	
Birth day	0.67	1.29	
Seconds in minute	0.84	0.11	
Direction sun sets	0.51	0.10	Long term memory
Age	0.69	0.23	
Year	0.80	0.28	
Month	0.70	0.65	Orientation

Item	Loading, general cognition factor	Loading, sub-domain factor	Name of sub-domain
Date	0.58	0.57	
Day of week	0.62	0.41	
Season	0.54	0.31	
SPA	0.79	0.36	
SPB	0.64	0.20	
First word repeating	0.34	0.36	
Second word repeating	0.43	0.80	
First phrase repeat	0.43	0.29	
Second phrase repeat	0.40	0.14	Attention
Digits backwards A	0.56	0.26	
Digits backwards B	0.40	0.45	
Digits backwards C	0.41	0.45	
First serial 3 subtract	0.48	0.50	
Second serial 3 subtract	0.56	0.56	
Third-fifth serial 3	0.51	0.55	Concentration
First recall, first word	0.51	0.65	
First recall, second word	0.45	0.72	
First recall, third word	0.44	0.53	
Second recall, first word	0.52	0.57	
Second recall, second word	0.59	0.60	
Second recall, third word	0.59	0.48	
Object recall	0.58	0.18	Short term memory
Follow written command	0.60	0.08	
Write a sentence	0.44	0.18	
Three-step command	0.26	-0.31	
Identify body parts	0.57	0.26	
Identify objects A	0.72	0.55	
Identify objects B	0.77	0.48	Language
Copy pentagons	0.51	N/a	Visual Construction
Animal fluency, 30 seconds	0.95	N/a	Fluency
Similarities	0.50	N/a	
Judgment items	0.38	N/a	Abstraction and judgment

Item	Loading, general cognition factor	Loading, sub-domain factor	Name of sub-domain
------	--	----------------------------	--------------------

The standardized loading for every item with a single exception (three step command had a loading of 0.26) on the general factor was greater than 0.30. Repeating the analysis with that item removed produced nearly identical results for the other items. Similarly, removing the birth day item (whose loading on the secondary factor was >1.0) produced virtually identical results for other items (largest change <0.03).

Appendix Table 3 shows the standardized loadings for the CSI 'D' from the Indianapolis site of the Indianapolis-Ibadan Dementia Study. The CFI was good at 0.933, the TLI was good at 0.923, and the RMSEA was adequate at 0.065 (<0.08 is adequate; <0.05 is considered good fit).

Appendix Table 3
Bi-factor model results from the CSI 'D':

Item	Loading, general cognition factor	Loading, sub-domain factor	Name of sub-domain
REMNAM E 1	0.51	0.56	Short term memory
REMNAM E 2	0.54	0.51	
CHILD1	0.55	0.62	
HOUSEFIR E1	0.57	0.59	
BRAVE1	0.48	0.68	
RESCUE1	0.41	0.74	
MINOR1	0.07	0.74	
WELL1	0.09	0.63	
BRIDGE	0.48	0.04	Abstraction
HAMMER	1.00	-0.02	
CHURCH	0.95	5.87	
BUYMED	0.96	-0.02	
PENCIL	0.71	0.66	Higher cortical functioning
WATCH	0.69	0.72	
CHAIR	0.65	0.67	
SHOES	0.57	0.75	
KNUCKLES	0.63	0.39	
ELBOW	0.68	0.64	
SHOULDE R	0.80	0.54	
IFANDBUT	0.30	0.02	
ANIMALS	3.66	-2.14	
BOAT1	0.86	0.36	
HOUSE1	0.86	0.35	
NODHEAD	0.63	0.05	Praxis

Item	Loading, general cognition factor	Loading, sub-domain factor	Name of sub-domain
WINDDOOR	0.79	0.39	
RTHAND	0.31	0.21	
FOLDS	0.41	0.31	
INLAP	0.58	0.28	
CIRCLES	0.43	0.61	
PENTAGON	0.14	0.87	
NAMECITY	0.78	0.22	
MJRSTREE	0.69	0.51	
LANDMAR K	0.54	0.40	
ADDRESS	0.54	0.69	
MAYOR	0.47	0.40	Orientation to place
MONTH	0.72	0.62	
DAYWEEK	0.70	0.57	
YEAR	0.70	0.44	
SEASON	0.56	0.31	
RAIN	0.66	0.39	
PARTDAY	0.49	0.42	
CIVILWAR	0.59	0.33	Orientation to time

Almost all items had loadings on the general factor >0.30. Two of the story elements - minor and well - had much lower loadings. When we totaled up the story elements and treated them as a single item, model fit was minimally changed and none of the loadings changed very much. The pentagons item had a low loading. Excluding it caused minimal change in loadings for the other items and did not have a significant impact on model fit. Similarly, removing items that had loadings >1.0 had minimal impact on fit statistics or loadings for any of the other items.

We thus concluded that the scales were sufficiently unidimensional to use IRT.

Appendix 2. Detailed methods of differential item functioning (DIF) analyses

We use IRT scores estimated using PARSCALE to evaluate items for DIF. We examine three ordinal logistic regression models for each item for each study comparison (e.g., CHS vs. ACT) (labeled here as “study”) selected for analysis:

$$f(\text{item response}) = \text{cut} + \beta_1 * \theta + \beta_2 * \text{study} + \beta_3 * \theta * \text{study} \quad (\text{model 1})$$

$$f(\text{item response}) = \text{cut} + \beta_1 * \theta + \beta_2 * \text{study} \quad (\text{model 2})$$

$$f(\text{item response}) = \text{cut} + \beta_1 * \theta \quad (\text{model 3})$$

In these models, “cut” is the cutpoint for each level in the proportional odds ordinal logistic regression model, as described by McCullagh and Nelder (43), and “ θ ” is the IRT estimate of cognitive functioning. All DIF analyses are performed using Stata (44).

Two types of DIF are identified in the literature. In items with *non-uniform DIF*, demographic interference between ability level and item responses differs at varying levels of cognitive functioning. In items with *uniform DIF*, this interference is the same across all levels of cognitive functioning. These concepts are analogous to the concepts of effect modification and confounding from epidemiology (23).

To detect non-uniform DIF, we compare the log likelihoods of models 1 and 2 using a X^2 test. We used an α level of 0.05. To detect uniform DIF, the relative difference between the parameters associated with θ (β_I from models 2 and 3) is determined using the formula $|(\beta_{I(model\ 2)} - \beta_{I(model\ 3)}) / \beta_{I(model\ 3)}|$. If the relative difference is large (10% in this study), group membership interferes with the expected relationship between ability and item responses.

If we found that a candidate anchor item had DIF with respect to a comparison between two studies, that item was rejected as an anchor, and treated separately in the two studies. We repeated these steps until DIF-free anchor items had been selected.

Spurious false-positive and false-negative results may occur if the ability score used for DIF detection includes many items with DIF (22). We use an iterative approach to deal with this issue. Once the final set of anchor items was obtained, we confirmed that they had no further DIF related to study site. Stata .ado files for DIF analyses are available for free download; type “ssc install difwithpar” at the Stata prompt.

We used an $\alpha = 0.05$ criterion for non-uniform DIF and a proportional change in β coefficient of 0.10 criterion for uniform DIF for this study.

Appendix 3. Additional references

Appendix Table 4
Description and final disposition of cognitive test items

	3MS	CASI	CSI `D'
1. Anchor items			
1a. For all 3 tests	Year Season	Year Season	Year Season
1b. For 3MS with CASI	Birthday Birthplace Date Registration (3) and 2 nd recall (3) of “shirt”, “brown”, and “honesty” First recall of “honesty” Naming animals (30 seconds, no leg restriction) (10) Spatial orientation (5)	Birthday Birthplace Date Registration (3) and 2 nd recall (3) of “shirt”, “brown”, and “honesty” First recall of “honesty” Naming animals (30 seconds, no leg restriction) (10) Spatial orientation (5)	
1c. For 3MS with CSI `D'	Draw pentagons Identify shoulder Identify elbow		Draw pentagons Identify shoulder Identify elbow

3MS	CASI	CSI `D'
1. Anchor items		
Identify knuckle		Identify knuckle
1d. For CASI with CSI `D'	Month Day	Month Day
2. Rejected candidate anchor items due to differential item functioning		
2a. Rejected for all 3	3 step command (3)	3 step command (3)
2b. Rejected for 3MS with CASI	Repetition of phrase ("he would like to go home") Reading & writing (2) Birth year Similarities (3) 1st recall of "shirt" and "brown" (2) Identify forehead Identify chin	Repetition of phrase ("he would like to go home") Reading & writing (2) Birth year Similarities (3) 1st recall of "shirt" and "brown" (2) Identify forehead Identify chin
2c. Rejected for 3MS with CSI `D'	Repetition of phrase (if and but) Month Day	Repetition of phrase (if and but) (no DIF for CASI/CSI `D') (no DIF for CASI/CSI `D')
2d. Rejected for CASI with CSI `D'	Identify shoulder Identify elbow (No DIF for 3MS/CSI `D') (No DIF for 3MS/CASI)	(no DIF for 3MS/CSI `D') (no DIF for 3MS/CSI `D') (No DIF for 3MS/CSI `D') Spatial orientation (state *, city)
3. Items unique to each test		
Mental Reversal	Age Repetition of phrase ("Yellow circle") Number of minutes/hour Identify wrist Object identification & recall (10) Serial threes (5) Judgment (3) Count backward (3) Direction of sunset	Math functions (8) * Identify pencil, watch , chair, shoes Part of the day * Rain the previous day Describe bridge, hammer, church Examiner's name (2) 1 st recall story elements (6) 2 nd recall story elements (6) * Draw circles General knowledge (when was World War II *, name civil rights leader, name president, name mayor, name governor *) Naming animals (4 legged, 60 seconds) Registration (3) and recall (3) of "boat", "house", and "fish"

3. Items unique to each test

Spatial orientation (street, landmark, address)

* Items with an asterisk are asked but not scored in the standard CSF D'. Items in bold are MMSE items. Numbers in parentheses are standard scoring weights for each item.

Appendix Figure 1.

Running mean bias for three different scoring techniques in estimating rates of change over time.

References

1. Folstein MF, Folstein SE, McHugh PR. "Mini-mental state" A practical method for grading the cognitive state of patients for the clinician. *J Psychiatr Res* 1975;12:189–98. [PubMed: 1202204]
2. Petersen RC, Stevens JC, Ganguli M, Tangalos EG, Cummings JL, DeKosky ST. Practice parameter: early detection of dementia: mild cognitive impairment (an evidence-based review). Report of the

- Quality Standards Subcommittee of the American Academy of Neurology. *Neurology* 2001;56:1133–42. [PubMed: 11342677]
3. Gauthier S, Reisberg B, Zaudig M, et al. Mild cognitive impairment. *Lancet* 2006;367:1262–70. [PubMed: 16631882]
4. Teng EL, Chui HC. The Modified Mini-Mental State (3MS) examination. *J Clin Psychiatry* 1987;48:314–8. [PubMed: 3611032]
5. Teng EL, Hasegawa K, Homma A, et al. The Cognitive Abilities Screening Instrument (CASI): a practical test for cross-cultural epidemiological studies of dementia. *Int Psychogeriatr* 1994;6:45–58. [PubMed: 8054493]discussion 62
6. Hall KS, Gao S, Emsley CL, Ogunniyi AO, Morgan O, Hendrie HC. Community screening interview for dementia (CSI`D'); performance in five disparate study sites. *Int J Geriatr Psychiatry* 2000;15:521–31. [PubMed: 10861918]
7. Hall KS, Hendrie HC, Brittain HM, Norton JA. Development of a dementia screening interview in two distinct languages. *Int J Methods Psych Res* 1993;2:1–28.
8. Hendrie HC, Albert MS, Butters MA, et al. The NIH Cognitive and Emotional Health Project: report of the Critical Evaluation Study Committee. *Alzheimer's & Dementia* 2006;2:12–32.
9. Mungas D, Reed BR. Application of item response theory for development of a global functioning measure of dementia with linear measurement properties. *Stat Med* 2000;19:1631–44. [PubMed: 10844724]
10. Embretson, SE.; Reise, SP. *Item response theory for psychologists*. Mahwah, NJ; Erlbaum: 2000.
11. Kukull WA, Higdon R, Bowen JD, et al. Dementia and Alzheimer disease incidence: a prospective cohort study. *Arch Neurol* 2002;59:1737–46. [PubMed: 12433261]
12. Hendrie HC, Ogunniyi A, Hall KS, et al. Incidence of dementia and Alzheimer disease in 2 communities: Yoruba residing in Ibadan, Nigeria, and African Americans residing in Indianapolis, Indiana. *JAMA* 2001;285:739–47. [PubMed: 11176911]
13. Fitzpatrick AL, Kuller LH, Ives DG, et al. Incidence and prevalence of dementia in the Cardiovascular Health Study. *J Am Geriatr Soc* 2004;52:195–204. [PubMed: 14728627]
14. Fried LP, Borhani NO, Enright P, et al. The Cardiovascular Health Study: design and rationale. *Ann Epidemiol* 1991;1:263–76. [PubMed: 1669507]
15. McDonald, RP. *Test theory: a unified treatment*. Mahwah, NJ; Erlbaum: 1999.
16. McHorney CA, Cohen AS. Equating health status measures with item response theory: illustrations with functional status items. *Med Care* 2000;38:43–59.
17. Muraki, E.; Bock, D. *PARSCALE for Windows*. Scientific Software International; Chicago: 2003.
18. Samejima F. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph* 1969;No. 17
19. Samejima, F. *Handbook of modern item response theory*. van der Linden, WJ.; Hambleton, RK., editors. Springer; NY: 1997. p. 85-100.
20. Millsap RE, Everson HT. Methodology review: statistical approaches for assessing measurement bias. *Applied Psychological Measurement* 1993;17:297–334.
21. Camilli, G.; Shepard, LA. *Methods for identifying biased test items*. Thousand Oaks; Sage: 1994.
22. Holland, PW.; Wainer, H. *Differential item functioning*. Hillsdale, NJ; Erlbaum: 1993.
23. Crane PK, van Belle G, Larson EB. Test bias in a cognitive test: differential item functioning in the CASI. *Stat Med* 2004;23:241–56. [PubMed: 14716726]
24. Crane PK, Hart DL, Gibbons LE, Cook KF. A 37-item shoulder functional status item pool had negligible differential item functioning. *J Clin Epidemiol* 2006;59:478–84. [PubMed: 16632136]
25. Baker, FB.; Kim, S-H. *Item response theory: parameter estimation techniques*. Marcel Dekker; NY: 2004.
26. Mungas D, Reed BR, Kramer JH. Psychometrically matched measures of global cognition, memory, and executive function for assessment of cognitive decline in older persons. *Neuropsychology* 2003;17:380–92. [PubMed: 12959504]
27. Hambleton, RK.; Swaminathan, H.; Rogers, HJ. *Fundamentals of item response theory*. Newbury Park; Sage: 1991.

28. Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics* 1982;38:963–74. [PubMed: 7168798]
29. Gibbons RD, Clark DC, Kupfer DJ. Exactly what does the Hamilton Depression Rating Scale measure? *J Psychiatr Res* 1993;27:259–73. [PubMed: 8295158]
30. Mislevy RJ, Beaton AE, Kaplan B, Sheehan KM. Estimating population characteristics from sparse matrix samples of item responses. *J Educ Meas* 1992;29:133–161.
31. Rubin, DB. Multiple imputation for nonresponse in surveys. Wiley; NY: 1987.
32. Rubin DB. Multiple imputation after 18+ years. *J Am Statist Assoc* 1996;91:473–89.
33. Fox JP. Stochastic EM for estimating the parameters of a multilevel IRT model. *Br J Math Stat Psychol* 2003;56:65–81. [PubMed: 12803822]
34. Fox, JP. Multilevel IRT model assessment. In: van der Ark, LA.; Croon, MA.; Sijtsma, K., editors. *New developments in categorical data analysis for the social and behavioral sciences*. Earlbaum; London: 2004. p. 227–252.
35. Fox JP. Applications of multilevel IRT modeling. *School Effectiveness and School Improvement* 2004;15:261–280.
36. Fox JP. Multilevel IRT using dichotomous and polytomous response data. *Br J Math Stat Psychol* 2005;58:145–72. [PubMed: 15969844]
37. Fox, JP.; Glas, CAW. Modeling measurement error in a structural multilevel model. In: Marcoulides, GA.; Moustaki, I., editors. *Latent variable and latent structure models*. Earlbaum; London: 2002. p. 245–269.
38. Fox JP, Glas CAW. Bayesian modeling of measurement error in predictor variables using item response theory. *Psychometrika* 2003;68:169–191.
39. Fox J-P, Glas CAW. Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika* 2001;66:271–288.
40. Crane PK, Gibbons LE, Jolley L, et al. Differential item functioning related to education and age in the Italian version of the Mini-mental State Examination. *Int Psychogeriatr* 2006;18:505–15. [PubMed: 16478571]
41. Crane PK, Gibbons LE, Jolley L, van Belle G. Differential item functioning analysis with ordinal logistic regression techniques: DIFdetect and difwithpar. *Med Care* 2006;44:S115–S123. [PubMed: 17060818]
42. Muthen, LK.; Muthen, BO. User's guide. LA; Muthen & Muthen: 1998. *Mplus: statistical analysis with latent variables*.
43. McCullagh, P.; Nelder, JA. *Generalized linear models*. Chapman and Hall; London: 1989.
44. StataCorp.. *Stata statistical software: release 8.0*. Stata Corporation; College Station, TX: 2003.
- A1. Ott A, Andersen K, Dewey ME, et al. Effect of smoking on global cognitive function in nondemented elderly. *Neurology* 2004;62:920–4. [PubMed: 15037693]
- A2. Gates GA, Cobb JL, Linn RT, Rees T, Wolf PA, D'Agostino RB. Central auditory dysfunction, cognitive dysfunction, and dementia in older people. *Arch Otolaryngol Head Neck Surg* 1996;122:161–7. [PubMed: 8630210]
- A3. Bassuk SS, Glass TA, Berkman LF. Social disengagement and incident cognitive decline in community-dwelling elderly persons. *Ann Intern Med* 1999;131:165–73. [PubMed: 10428732]
- A4. Lui LY, Stone K, Cauley JA, Hillier T, Yaffe K. Bone loss predicts subsequent cognitive decline in older women: the study of osteoporotic fractures. *J Am Geriatr Soc* 2003;51:38–43. [PubMed: 12534843]
- A5. Evans DA, Bennett DA, Wilson RS, et al. Incidence of Alzheimer disease in a biracial urban community: relation to apolipoprotein E allele status. *Arch Neurol* 2003;60:185–9. [PubMed: 12580702]
- A6. Morris MC, Evans DA, Bienias JL, Tangney CC, Wilson RS. Dietary fat intake and 6-year cognitive change in an older biracial community population. *Neurology* 2004;62:1573–9. [PubMed: 15136684]
- A7. Elkind MS, Cheng J, Boden-Albala B, Paik MC, Sacco RL. Elevated white blood cell count and carotid plaque thickness : the Northern Manhattan Stroke Study. *Stroke* 2001;32:842–9. [PubMed: 11283380]

- A8. Maggi S, Zucchetto M, Grigoletto F, et al. The Italian Longitudinal Study on Aging (ILSA): design and methods. *Aging (Milano)* 1994;6:464–73. [PubMed: 7748921]
- A9. Solfrizzi V, Colacicco AM, D'Introno A, et al. Dietary intake of unsaturated fatty acids and age-related cognitive decline: A 8.5-year follow-up of the Italian Longitudinal Study on Aging. *Neurobiol Aging*. 2005
- A10. Black SA, Espino DV, Mahurin R, et al. The influence of noncognitive factors on the Mini-Mental State Examination in older Mexican-Americans: findings from the Hispanic EPESE. Established Population for the Epidemiologic Study of the Elderly. *J Clin Epidemiol* 1999;52:1095–102. [PubMed: 10527004]
- A11. Winnock M, Letenneur L, Jacqmin-Gadda H, Dallonville J, Amouyel P, Dartigues JF. Longitudinal analysis of the effect of apolipoprotein E epsilon4 and education on cognitive performance in elderly subjects: the PAQUID study. *J Neurol Neurosurg Psychiatry* 2002;72:794–7. [PubMed: 12023428]
- A12. Reynolds CA, Finkel D, Gatz M, Pedersen NL. Sources of influence on rate of cognitive change over time in Swedish twins: an application of latent growth models. *Exp Aging Res* 2002;28:407–33. [PubMed: 12227921]
- A13. Guo Z, Fratiglioni L, Winblad B, Viitanen M. Blood pressure and performance on the Mini-Mental State Examination in the very old. Cross-sectional and longitudinal data from the Kungsholmen Project. *Am J Epidemiol* 1997;145:1106–13. [PubMed: 9199540]
- A14. Yip AG, Brayne C, Easton D, Rubinsztein DC. Apolipoprotein E4 is only a weak predictor of dementia and cognitive decline in the general population. *J Med Genet* 2002;39:639–43. [PubMed: 12205106]
- A15. Stewart R, Johnson J, Richards M, Brayne C, Mann A. The distribution of Mini-Mental State Examination scores in an older UK African-Caribbean population compared to MRC CFA study norms. *Int J Geriatr Psychiatry* 2002;17:745–51. [PubMed: 12211125]
- A16. Geerlings MI, Schoevers RA, Beekman AT, et al. Depression and risk of cognitive decline and Alzheimer's disease. Results of two prospective community-based studies in The Netherlands. *Br J Psychiatry* 2000;176:568–75. [PubMed: 10974964]
- A17. Dik MG, Jonker C, Comijs HC, et al. Memory complaints and APOE-epsilon4 accelerate cognitive decline in cognitively normal elderly. *Neurology* 2001;57:2217–22. [PubMed: 11756600]
- A18. Slooter AJ, van Duijn CM, Bots ML, et al. Apolipoprotein E genotype, atherosclerosis, and cognitive decline: the Rotterdam Study. *J Neural Transm Suppl* 1998;53:17–29. [PubMed: 9700643]
- A19. Kalmijn S, Launer LJ, Lindemans J, Bots ML, Hofman A, Breteler MM. Total homocysteine and cognitive decline in a community-based sample of elderly subjects: the Rotterdam Study. *Am J Epidemiol* 1999;150:283–9. [PubMed: 10430233]
- A20. Zelinski EM, Kennison RF. The Long Beach Longitudinal Study: evaluation of longitudinal effects of aging on memory and cognition. *Home Health Care Serv Q* 2001;19:45–55. [PubMed: 11436406]
- A21. McCann JJ, Gilley DW, Hebert LE, Beckett LA, Evans DA. Concordance between direct observation and staff rating of behavior in nursing home residents with Alzheimer's disease. *J Gerontol B Psychol Sci Soc Sci* 1997;52:P63–72. [PubMed: 9060981]
- A22. Leveille SG, Guralnik JM, Ferrucci L, Corti MC, Kasper J, Fried LP. Black/white differences in the relationship between MMSE scores and disability: the Women's Health and Aging Study. *J Gerontol B Psychol Sci Soc Sci* 1998;53:P201–8. [PubMed: 9602835]
- A23. Canadian study of health and aging: study methods and prevalence of dementia. *CMAJ* 1994;150:899–913. [PubMed: 8131123]
- A24. Rapp SR, Espeland MA, Hogan P, Jones BN, Dugan E. Baseline experience with Modified Mini Mental State Exam: The Women's Health Initiative Memory Study (WHIMS). *Aging Ment Health* 2003;7:217–23. [PubMed: 12775404]
- A25. Rapp SR, Espeland MA, Shumaker SA, et al. Effect of estrogen plus progestin on global cognitive function in postmenopausal women: the Women's Health Initiative Memory Study: a randomized controlled trial. *JAMA* 2003;289:2663–72. [PubMed: 12771113]
- A26. Haan MN, Shemanski L, Jagust WJ, Manolio TA, Kuller L. The role of APOE epsilon4 in modulating effects of other risk factors for cognitive decline in elderly persons. *JAMA* 1999;282:40–6. [PubMed: 10404910]

- A27. Yaffe K, Haan M, Byers A, Tangen C, Kuller L. Estrogen use, APOE, and cognitive decline: evidence of gene-environment interaction. *Neurology* 2000;54:1949–54. [PubMed: 10822435]
- A28. Kuller LH, Shemanski L, Manolio T, et al. Relationship between ApoE, MRI findings, and cognitive function in the Cardiovascular Health Study. *Stroke* 1998;29:388–98. [PubMed: 9472879]
- A29. Tschanz JT, Welsh-Bohmer KA, Plassman BL, Norton MC, Wyse BW, Breitner JC. An adaptation of the Modified Mini-Mental State Examination: analysis of demographic influences and normative data: the Cache County Study. *Neuropsychiatry Neuropsychol Behav Neurol* 2002;15:28–38. [PubMed: 11877549]
- A30. Khachaturian AS, Gallo JJ, Breitner JC. Performance characteristics of a two-stage dementia screen in a population sample. *J Clin Epidemiol* 2000;53:531–40. [PubMed: 10812327]
- A31. Rankin MW, Clemons TE, McBee WL. Correlation analysis of the in-clinic and telephone batteries from the AREDS cognitive function ancillary study. AREDS Report No. 15. *Ophthalmic Epidemiol* 2005;12:271–7. [PubMed: 16033748]
- A32. Yaffe K, Lindquist K, Penninx BW, et al. Inflammatory markers and cognition in well-functioning African-American and white elders. *Neurology* 2003;61:76–80. [PubMed: 12847160]
- A33. Yaffe K, Barrett-Connor E, Lin F, Grady D. Serum lipoprotein levels, statin use, and cognitive function in older women. *Arch Neurol* 2002;59:378–84. [PubMed: 11890840]
- A34. Petrovitch H, White L, Masaki KH, et al. Influence of myocardial infarction, coronary artery bypass surgery, and stroke on cognitive impairment in late life. *Am J Cardiol* 1998;81:1017–21. [PubMed: 9576163]
- A35. Foley D, Monjan A, Masaki K, et al. Daytime sleepiness is associated with 3-year incident dementia and cognitive decline in older Japanese-American men. *J Am Geriatr Soc* 2001;49:1628–32. [PubMed: 11843995]
- A36. Launer LJ, Ross GW, Petrovitch H, et al. Midlife blood pressure and dementia: the Honolulu-Asia aging study. *Neurobiol Aging* 2000;21:49–55. [PubMed: 10794848]
- A37. Kukull WA, Higdon R, Bowen JD, et al. Dementia and Alzheimer disease incidence: a prospective cohort study. *Arch Neurol* 2002;59:1737–46. [PubMed: 12433261]
- A38. Larson EB, Wang L, Bowen JD, et al. Exercise is associated with reduced risk for incident dementia among persons 65 years of age and older. *Ann Intern Med* 2006;144:73–81. [PubMed: 16418406]
- A39. Yamada M, Sasaki H, Kasagi F, et al. Study of cognitive function among the Adult Health Study (AHS) population in Hiroshima and Nagasaki. *Radiat Res* 2002;158:236–40. [PubMed: 12105994]
- A40. Graves AB, Rajaram L, Bowen JD, McCormick WC, McCurry SM, Larson EB. Cognitive decline and Japanese culture in a cohort of older Japanese Americans in King County, WA: the Kame Project. *J Gerontol B Psychol Sci Soc Sci* 1999;54:S154–61. [PubMed: 10363046]
- A41. Rice MM, Graves AB, McCurry SM, et al. Postmenopausal estrogen and estrogen-progestin use and 2-year rate of cognitive change in a cohort of older Japanese American women: The Kame Project. *Arch Intern Med* 2000;160:1641–9. [PubMed: 10847257]
- A42. Liu HC, Teng EL, Lin KN, et al. Performance on the cognitive abilities screening instrument at different stages of Alzheimer's disease. *Dement Geriatr Cogn Disord* 2002;13:244–8. [PubMed: 12006735]
- A43. Liu HC, Chou P, Lin KN, et al. Assessing cognitive abilities and dementia in a predominantly illiterate population of older individuals in Kinmen. *Psychol Med* 1994;24:763–70. [PubMed: 7991758]
- A44. Ogunniyi A, Baiyewu O, Gureje O, et al. Epidemiology of dementia in Nigeria: results from the Indianapolis-Ibadan study. *Eur J Neurol* 2000;7:485–90. [PubMed: 11054131]
- A45. Hall KS, Gao S, Emsley CL, Ogunniyi AO, Morgan O, Hendrie HC. Community screening interview for dementia (CSI 'D'); performance in five disparate study sites. *Int J Geriatr Psychiatry* 2000;15:521–31. [PubMed: 10861918]
- A46. Prince M, Acosta D, Chiu H, Scazufca M, Varghese M. Dementia diagnosis in developing countries: a cross-cultural validation study. *Lancet* 2003;361:909–17. [PubMed: 12648969]
- A47. Aartsen MJ, Smits CH, van Tilburg T, Knipscheer KC, Deeg DJ. Activity in older adults: cause or consequence of cognitive functioning? A longitudinal study on everyday activities and cognitive performance in older adults. *J Gerontol B Psychol Sci Soc Sci* 2002;57:P153–62. [PubMed: 11867663]

- A48. Dik MG, Deeg DJ, Bouter LM, Corder EH, Kok A, Jonker C. Stroke and apolipoprotein E epsilon4 are independent risk factors for cognitive decline: A population-based study. *Stroke* 2000;31:2431–6. [PubMed: 11022076]
- A49. Yaffe K, Barnes D, Nevitt M, Lui LY, Covinsky K. A prospective study of physical activity and cognitive decline in elderly women: women who walk. *Arch Intern Med* 2001;161:1703–8. [PubMed: 11485502]
- A50. Yaffe K, Browner W, Cauley J, Launer L, Harris T. Association between bone mineral density and cognitive decline in older women. *J Am Geriatr Soc* 1999;47:1176–82. [PubMed: 10522949]
- A51. Yaffe K, Cauley J, Sands L, Browner W. Apolipoprotein E phenotype and cognitive decline in a prospective study of elderly community women. *Arch Neurol* 1997;54:1110–4. [PubMed: 9311354]
- A52. Yaffe K, Grady D, Pressman A, Cummings S. Serum estrogen levels, cognitive performance, and risk of cognitive decline in older community women. *J Am Geriatr Soc* 1998;46:816–21. [PubMed: 9670866]
- A53. Aartsen MJ, Martin M, Zimprich D. Gender differences in level and change in cognitive functioning. Results from the Longitudinal Aging Study Amsterdam. *Gerontology* 2004;50:35–8. [PubMed: 14654725]
- A54. Abbott RD, White LR, Ross GW, et al. Height as a marker of childhood development and late-life cognitive function: the Honolulu-Asia Aging Study. *Pediatrics* 1998;102:602–9. [PubMed: 9738183]
- A55. Albert MS, Jones K, Savage CR, et al. Predictors of cognitive change in older persons: MacArthur studies of successful aging. *Psychol Aging* 1995;10:578–89. [PubMed: 8749585]
- A56. Barnes LL, Wilson RS, Schneider JA, Bienias JL, Evans DA, Bennett DA. Gender, cognitive decline, and risk of AD in older persons. *Neurology* 2003;60:1777–81. [PubMed: 12796530]
- A57. Bretsky P, Guralnik JM, Launer L, Albert M, Seeman TE. The role of APOE-epsilon4 in longitudinal cognitive decline: MacArthur Studies of Successful Aging. *Neurology* 2003;60:1077–81. [PubMed: 12682309]
- A58. Chyou PH, White LR, Yano K, et al. Pulmonary function measures as predictors and correlates of cognitive functioning in later life. *Am J Epidemiol* 1996;143:750–6. [PubMed: 8610684]
- A59. Comijs HC, Jonker C, Beekman AT, Deeg DJ. The association between depressive symptoms and cognitive decline in community-dwelling elderly persons. *Int J Geriatr Psychiatry* 2001;16:361–7. [PubMed: 11333422]
- A60. Elias PK, D'Agostino RB, Elias MF, Wolf PA. Blood pressure, hypertension, and age as risk factors for poor cognitive performance. *Exp Aging Res* 1995;21:393–417. [PubMed: 8595804]
- A61. Elias MF, Elias PK, Sullivan LM, Wolf PA, D'Agostino RB. Lower cognitive function in the presence of obesity and hypertension: the Framingham heart study. *Int J Obes Relat Metab Disord* 2003;27:260–8. [PubMed: 12587008]
- A62. Elias PK, Elias MF, D'Agostino RB, Silbershatz H, Wolf PA. Alcohol consumption and cognitive performance in the Framingham Heart Study. *Am J Epidemiol* 1999;150:580–9. [PubMed: 10489997]
- A63. Elias MF, D'Agostino RB, Elias PK, Wolf PA. Neuropsychological test performance, cognitive functioning, blood pressure, and age: the Framingham Heart Study. *Exp Aging Res* 1995;21:369–91. [PubMed: 8595803]
- A64. Fillenbaum GG, Landerman LR, Blazer DG, Saunders AM, Harris TB, Launer LJ. The relationship of APOE genotype to cognitive functioning in older African-American and Caucasian community residents. *J Am Geriatr Soc* 2001;49:1148–55. [PubMed: 11559372]
- A65. Galanis DJ, Joseph C, Masaki KH, Petrovitch H, Ross GW, White L. A longitudinal study of drinking and cognitive performance in elderly Japanese American men: the Honolulu-Asia Aging Study. *Am J Public Health* 2000;90:1254–9. [PubMed: 10937006]
- A66. Galanis DJ, Petrovitch H, Launer LJ, Harris TB, Foley DJ, White LR. Smoking history in middle age and subsequent cognitive performance in elderly Japanese-American men. The Honolulu-Asia Aging Study. *Am J Epidemiol* 1997;145:507–15. [PubMed: 9063340]
- A67. Glynn RJ, Beckett LA, Hebert LE, Morris MC, Scherr PA, Evans DA. Current and remote blood pressure and cognitive decline. *Jama* 1999;281:438–45. [PubMed: 9952204]

- A68. Graves AB, Bowen JD, Rajaram L, et al. Impaired olfaction as a marker for cognitive decline: interaction with apolipoprotein E epsilon4 status. *Neurology* 1999;53:1480–7. [PubMed: 10534255]
- A69. Gregg EW, Yaffe K, Cauley JA, et al. Is diabetes associated with cognitive impairment and cognitive decline among older women? Study of Osteoporotic Fractures Research Group. *Arch Intern Med* 2000;160:174–80. [PubMed: 10647755]
- A70. Grodstein F, Chen J, Hankinson SE. Cataract extraction and cognitive function in older women. *Epidemiology* 2003;14:493–7. [PubMed: 12843777]
- A71. Grodstein F, Chen J, Pollen DA, et al. Postmenopausal hormone therapy and cognitive function in healthy older women. *J Am Geriatr Soc* 2000;48:746–52. [PubMed: 10894312]
- A72. Grodstein F, Chen J, Willett WC. High-dose antioxidant supplements and cognitive function in community-dwelling elderly women. *Am J Clin Nutr* 2003;77:975–84. [PubMed: 12663300]
- A73. Grodstein F, Chen J, Wilson RS, Manson JE. Type 2 diabetes and cognitive function in community-dwelling elderly women. *Diabetes Care* 2001;24:1060–5. [PubMed: 11375371]
- A74. Jonker C, Comijs HC, Smit JH. Does aspirin or other NSAIDs reduce the risk of cognitive decline in elderly persons? Results from a population-based study. *Neurobiol Aging* 2003;24:583–8. [PubMed: 12714115]
- A75. Hee Kang J, Grodstein F. Regular use of nonsteroidal anti-inflammatory drugs and cognitive function in aging women. *Neurology* 2003;60:1591–7. [PubMed: 12771247]
- A76. Launer LJ, Masaki K, Petrovitch H, Foley D, Havlik RJ. The association between midlife blood pressure levels and late-life cognitive function. The Honolulu-Asia Aging Study. *JAMA* 1995;274:1846–51. [PubMed: 7500533]
- A77. Lee S, Kawachi I, Berkman LF, Grodstein F. Education, other socioeconomic indicators, and cognitive function. *Am J Epidemiol* 2003;157:712–20. [PubMed: 12697575]
- A78. Lee S, Kawachi I, Grodstein F. Does caregiving stress affect cognitive function in older women? *J Nerv Ment Dis* 2004;192:51–7. [PubMed: 14718776]
- A79. Logroscino G, Kang JH, Grodstein F. Prospective study of type 2 diabetes and cognitive decline in women aged 70–81 years. *BMJ* 2004;328:548. [PubMed: 14980984]
- A80. Masaki KH, Losonczy KG, Izmirlian G, et al. Association of vitamin E and C supplement use with cognitive function and dementia in elderly men. *Neurology* 2000;54:1265–72. [PubMed: 10746596]
- A81. Matthews K, Cauley J, Yaffe K, Zmuda JM. Estrogen replacement therapy and cognitive decline in older community women. *J Am Geriatr Soc* 1999;47:518–23. [PubMed: 10323642]
- A82. Meyer PM, Powell LH, Wilson RS, et al. A population-based longitudinal study of cognitive functioning in the menopausal transition. *Neurology* 2003;61:801–6. [PubMed: 14504324]
- A83. Morris MC, Evans DA, Bienias JL, et al. Dietary folate and vitamin B12 intake and cognitive decline among community-dwelling older persons. *Arch Neurol* 2005;62:641–5. [PubMed: 15824266]
- A84. Peila R, White LR, Petrovich H, et al. Joint effect of the APOE gene and midlife systolic blood pressure on late-life cognitive impairment: the Honolulu-Asia aging study. *Stroke* 2001;32:2882–9. [PubMed: 11739991]
- A85. Rozzini R, Ferrucci L, Losonczy K, Havlik RJ, Guralnik JM. Protective effect of chronic NSAID use on cognitive decline in older persons. *J Am Geriatr Soc* 1996;44:1025–9. [PubMed: 8790225]
- A86. Schmand B, Smit J, Lindeboom J, et al. Low education is a genuine risk factor for accelerated memory decline and dementia. *J Clin Epidemiol* 1997;50:1025–33. [PubMed: 9363037]
- A87. Schwartz BS, Stewart WF, Bolla KI, et al. Past adult lead exposure is associated with longitudinal decline in cognitive function. *Neurology* 2000;55:1144–50. [PubMed: 11071492]
- A88. Seeman T, McAvay G, Merrill S, Albert M, Rodin J. Self-efficacy beliefs and change in cognitive performance: MacArthur Studies of Successful Aging. *Psychol Aging* 1996;11:538–51. [PubMed: 8893321]
- A89. Seeman TE, Lusignolo TM, Albert M, Berkman L. Social relationships, social support, and patterns of cognitive aging in healthy, high-functioning older adults: MacArthur studies of successful aging. *Health Psychol* 2001;20:243–55. [PubMed: 11515736]

- A90. Vermeer SE, Prins ND, den Heijer T, Hofman A, Koudstaal PJ, Breteler MM. Silent brain infarcts and the risk of dementia and cognitive decline. *N Engl J Med* 2003;348:1215–22. [PubMed: 12660385]
- A91. White L, Katzman R, Losonczy K, et al. Association of education with incidence of cognitive impairment in three established populations for epidemiologic studies of the elderly. *J Clin Epidemiol* 1994;47:363–74. [PubMed: 7730861]

Figure 1.

Screening cut-points used in selected studies. Red = MMSE, blue = 3MS, black = CASI, green = CSI 'D'. The x axis is the re-scaled IRT score, with a mean of 100 and a standard deviation of 15. Screening cutpoints were abstracted from the sources cited in the figure. Thus there is variability in screening cutpoints across studies that use the same instrument (identified in the figure by ink color). The unique contribution here is the ability to compare cutpoints across studies that used different tests, such as the MMSE cutpoint used by the Framingham study and the CASI cutpoint used by the ACT study. The cutpoints used in the different studies varied by nearly 4 standard deviations, a huge variation.

Figure 2.

Test characteristic curves for tests of global cognitive functioning. Red = MMSE, blue = 3MS, black = CASI, green = CSI 'D'. The x axis is the re-scaled IRT score, with a mean of 100 and a standard deviation of 15.

Figure 2a. 3MS

Figure 2b. CASI

Figure 2c. CSI 'D'

Figure 2d. MMSE

Figure 3. Standard error of measurement for tests of global cognitive functioning. Red = MMSE, blue = 3MS, black = CASI, green = CSI 'D'. The x axis is the re-scaled IRT score, with a mean of 100 and a standard deviation of 15.

Table 1

Large studies that have employed a cognitive functioning test*.

Mini-Mental State Examination (MMSE)	n
EuroDEM (A1)	31,031
Framingham Heart Study (A2)	10,850
Established Populations for the Epidemiological Study of the Elderly (A3)	10,294
Study of Osteoporotic Fractures (A4)	9,704
Chicago Health and Aging Project (A5,A6)	3,718
Northern Manhattan Study (A7)	3,298
Italian Longitudinal Study on Aging (A8,A9)	3,120
Hispanic EPESE (A10)	3,050
Personnes Agees Quid (A1,A11)	2,792
Swedish Adoption / Twin Study of Aging (A12)	2,000
Kungsholmen Project (A13)	1,736
Medical Research Council Cognitive Function and Aging (A14,A15)	1,463
Longitudinal Aging Study Amsterdam (A16,A17)	1,224
Rotterdam Study (A18,A19)	619
A Longitudinal Study of Cognition in Adults (A20)	600
Longitudinal Study of Daycare in Alzheimer's Disease (A21)	517
Women's Health and Aging Study (A22)	436
Subtotal, MMSE	86,452
Modified Mini-Mental State Examination (3MS)	n
Canadian Study on Health and Aging (A23)	10,263
Women's Health Initiative Memory Study (A24,A25)	7,480
Cardiovascular Health Study (A26-A28)	5,888
Cache County Memory Study (A29,A30)	5,092
Age-Related Eye Disease Study (A31)	4,757
Health Aging and Body Composition Study (A32)	3,075
Heart and Estrogen/Progestin Replacement Study (A33)	1,037
Subtotal, 3MS	37,592
Cognitive Abilities Screening Instrument (CASI)	n
Honolulu Asia Aging Study (A34-A36)	3,734
Adult Changes in Thought (A37,A38)	3,358
Adult Health Study in Hiroshima and Nagasaki (A39)	3,113
KAME (A40,A41)	1,991
Taipei Veteran's Hospital Study (A42)	475
Study in Kinmen, Taiwan (A43)	455
Subtotal, CASI	13,126

Community Screening Instrument for Dementia (CSI `D')	n
Indianapolis- Ibadan Study (A44,A45)	8,530
10/66 Project (A46)	2,885
Subtotal, CSI `D'	11,415
Grand total of subjects in all studies	148,585

*
References preceded by A (e.g., A1) refer to references available in Appendix

Table 2
Demographic characteristics of subjects in the parent studies.

	Cardiovascular Health Study	Adult Changes in Thought Study	Indianapolis site of the Indianapolis-Ibadan Dementia Study	Total
Age, y				
65-70	132 (3%)	297 (9%)	123 (9%)	552
70-75	2,074 (41%)	701 (21%)	418 (33%)	3,193
75-80	1,460 (29%)	971 (29%)	341 (27%)	2,772
80-85	884 (18%)	804 (24%)	200 (16%)	1,888
85+	428 (9%)	585 (17%)	172 (14%)	1,185
Sex		*		
Female	2,916 (59%)	1,979 (59%)	867 (69%)	5,762
Male	2,062 (41%)	1,378 (41%)	387 (31%)	3,827
Race / ethnicity		*		
White	4,072 (81%)	3,034 (90%)	0	7,106
Black	874 (18%)	151 (5%)	1,254 (100%)	2,279
Other	32 (1%)	172 (5%)	0	204
Education, y	†	*	‡	
0-5	226 (5%)	25 (1%)	154 (12%)	405
6-11	1,145 (23%)	396 (12%)	605 (48%)	2,146
12	1,344 (27%)	864 (26%)	291 (23%)	2,699
13-16	460 (9%)	1,449 (43%)	198 (16%)	2,107
17+	1,794 (36%)	623 (19%)	0	2,417

* Data missing for 1 subject

† Data missing for 9 subjects

‡ Data missing for 6 subjects

Table 3

Co-calibrated scores from four global cognitive screening tests*

MMSE	3MS	CASI	CSFD'	IRT Score
	18	50	13.0	47.5
	20	51	14.5	
10	24	53		
	25	57	15.0	
11	27	58	16.0	55.0
12	30			
	31		17.5	
	32	59		
13	33		18.5	
	34	61		
	35		19.5	
14	39	62		
15	40		20.5	62.5
	41	63		
16	42	66		
17	44	67	21	
	45	69		
18	46	70	22	
	47	71		
	49	72		
19	50			
	51	73	23	
20	52			
21	53	74		
22	54			

MMSE	3MS	CASI	CSFD'	IRT Score
			24	
	55		25	
	56			
	57			70.0
	59		26	
	60			
23	61			
	64	77	26	
24	65	78	27	
	66			
	67	79		
	68			
25	69		28	
	70	80		
			28.5	77.5
26	71	81		
		82		
	72			
	73	83		
	74	84		
	75	85	29.5	85.0
27	76	86		
	77			
	78			
	79	87		
	80			

MMSE	3MS	CASI	CSFD'	IRT Score
	81			
	82	88		
	83			
	84			
	85	89		
28	86	90		92.5
	87	91	30.5	
29	88		31.5	
	89	92		
	92	94	32	100.0
	94	95		
	96	96		
30	98	97		107.5
	99	98	33	115.0
	100	99		
		100		

* Each cell indicates ranges of scores on other tests that are equivalent to that score. For example, 20 points on the MMSE corresponds to 51 or 52 points on the 3MS, 73 or 74 points on the CASI, 23 on the CSI 'D', and about 67 or so on the IRT metric. The IRT score is scaled such that the mean score in non-demented individuals is 100 and the standard deviation (corresponding to 1 logit unit) is 15. We provide ½ logit (7.5 points) increments for the IRT score scale.

Table 4

Simulation study results*

Scenario number	Average intercept	Average decline over 8 years	% bias, standard scores	% bias, naïve IRT scores	% bias, single perturbed IRT score	% bias, 10 perturbed IRT scores
1	107.5	15.0	-0.2%	0.0%	-7.7%	-7.9%
2	100.0	7.5	+17.0%	+8.7%	+0.5%	+0.8%
3	100.0	15.0	+0.8%	-0.1%	-6.0%	-6.3%
4	92.5	7.5	+13.0%	+5.2%	-1.1%	-0.8%
5	85.0	7.5	+7.5%	+1.9%	-2.9%	-3.0%

* Each scenario is defined by an average intercept and an average slope (average decline over 8 years). Standard deviations for the random intercepts were all 15 and for the random slopes were all 7.5 points over 8 years. These values were then used to generate true ability levels at each time point for each subject for each simulated data set (250 data sets per scenario). The true ability levels were used to generate item responses. These item responses were then scored using 4 strategies (see methods section for more details). Bias was defined as the difference between the slope estimated using the scoring technique and the slope from the true ability levels; percent bias was defined as this value divided by the slope from true ability levels x 100%.