Published in final edited form as:

J Aging Health. 2015 December ; 27(8): 1392–1414. doi:10.1177/0898264315583054.

Harmonizing Measures of Cognitive Performance Across International Surveys of Aging Using Item Response Theory

Kitty S. Chan, Ph.D.^{1,*}, Alden L. Gross, Ph.D., M.H.S.², Liliana E. Pezzin, Ph.D., J.D.³, Jason Brandt, Ph.D.⁴, and Judith D. Kasper, Ph.D.¹

¹Department of Health Policy and Management, Johns Hopkins Bloomberg School of Public Health

²Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health

³Department of Medicine and Center for Patient Care and Outcomes Research, Medical College of Wisconsin

⁴Departments of Psychiatry & Behavioral Sciences and Neurology, Johns Hopkins University School of Medicine, The Copper Ridge Institute, Sykesville, MD

Abstract

Objective—To harmonize measures of cognitive performance using item response theory (IRT) across two international aging studies.

Methods—Data for persons 65 from the Health and Retirement Study (HRS, N=9,471) and the English Longitudinal Study of Aging (ELSA, N=5,444). Cognitive performance measures varied (HRS fielded 25, ELSA 13); 9 were in common. Measurement precision was examined for IRT scores based on: 1) common items; 2) common items adjusted for differential item functioning (DIF); 3) DIF-adjusted all items.

Results—Three common items (day of date, immediate word recall, and delayed word recall) demonstrated DIF by survey. Adding survey-specific items improved precision, but mainly for HRS respondents at lower cognitive levels.

Discussion—IRT offers a feasible strategy for harmonizing cognitive performance measures across other surveys and for other multi-item constructs of interest in studies of aging. Practical implications depend on sample distribution and the difficulty mix of in-common and survey-specific items.

Keywords

Item Response Theory; Cognitive Performance; Surveys

*Corresponding Author: Kitty S. Chan, Ph.D., Associate Professor, Department of Health Policy and Management, Johns Hopkins Bloomberg School of Public Health, 624 N. Broadway #633, Baltimore, MD 21205; Telephone: 410-614-4043; Fax: 410-955-0470; kchan10@jhu.edu.

BACKGROUND

Cross-national comparisons provide a window on the aging experience across varying societal contexts (Schoeni & Ofstedal, 2010). With the proportion of persons aged 60 and older worldwide projected to increase from 11% in 2007 to 22% by 2050 (United Nations, 2007; Kinsella & He, 2009), cross-national research on social, economic, and cultural variability in aging is more relevant than ever. A unique contribution of cross-national studies is the opportunity to identify aspects of the disablement process that are modifiable through policy or interventions, including behavior change and modifications to living environments (Figueras & McKee, 2012).

Longitudinal studies suggest that a significant proportion of older people experience cognitive decline (Yaffe et al., 2009), and that cognitive capacity, or the global ability to think and process information, is predictive of functioning in self-care and other activities of daily living (Njegovan, Hing, Mitchell, & Molnar, 2001). Prior research has shown that cognitive skills, such as memory, reasoning, processing speed, may be modifiable by training individuals on strategies to improve cognitive performance (Gross & Rebok, 2011; Willis et al., 2006; Ball et al., 2002). For example, training older adults in mnemonic techniques of rehearsal, association, categorization or imagery has improved performance on memory tests (Gross et al., 2012). Recognition that cognitive ability plays an important role in functioning (e.g., Gross & Rebok, 2011) has led to inclusion of cognitive measures in population-based studies of older people. One of the earliest to include such measures was the Asset and Health Dynamics of the Oldest Old Survey (AHEAD) (Herzog & Wallace, 1997; Herzog & Rodgers, 1999), but cognitive assessments are now part of assessments in several large international studies of aging. Among these are the Health and Retirement Study (HRS) and the English Longitudinal Study on Aging (ELSA), which have a shared focus on health and economic issues related to aging. Pooling existing data across these major surveys, done appropriately, can foster novel opportunities for international research into the aging experience.

Two concerns often arise in pooling data to conduct cross-national research—whether items that are identically worded are being interpreted and responded to similarly across surveys and how to analyze a construct of interest when the number and content of items differ across surveys. For cognitive assessments, the logistical challenges of creating and norming cognitive assessments that are comparable cross-culturally yet appropriate in different contexts are well recognized (Hendrie, 2006; Ferraro, 2002; Nell, 1999). For example, a project to harmonize the Cambridge Cognitive Examination (a neuropsychological test battery used in dementia diagnosis; Roth, Huppert, Mountjoy & Tym, 1999) across seven European countries, used an intensive iterative consensus process to achieve cultural and translational comparability (Verhey et al., 2003). This type of cross-national standardization is costly and time consuming to implement. Vignettes offer another methodology for harmonizing measures across international surveys (King, Murray, Salomon & Tandon, 2004; Bago d'Uva, O'Donnell & van Doorslaer, 2008; Salomon, Tandon & Murray, 2004). In this approach, respondents are asked to rate a common set of vignettes, known as anchoring vignettes, and then provide a self-assessment along the trait of interest. As the underlying trait level in the anchoring vignettes is constant across different raters, comparing

an individual's ratings on the anchoring vignettes along with his or her self-rating provides a means to adjust for differences in how individuals approach the rating process. However, as vignette assessment data has to be collected for each item of interest, this approach to harmonization is not applicable after primary data collection is completed.

Varying objectives and breadth of issues being covered by different surveys often limit the set of items included to measure constructs like cognitive performance, resulting in use of different items and different number of items, and a limited set of common items across surveys. This holds true even for HRS and ELSA where greater coordination of content has been achieved than is typical. Faced with varying items across data sets, researchers often exclude non-common items and collapse categories to achieve a set of comparable items that can be pooled (Pluijm, 2005). However, this approach leads to loss of information since it ignores the measurement contribution of excluded items.

Item response theory (IRT) offers two major advantages for harmonizing constructs, such as cognitive performance, that are assessed through multiple items. First, IRT can leverage common items across surveys to align scores along the same scale and generate comparable scores while retaining the additional information from available survey-specific items. This can potentially enhance score precision, particularly in the range of performance assessed by these survey-specific items. Second, IRT allows differentially functioning items (DIF; Hambleton, Swaminathan & Rogers, 1991; Holland & Wainer, 1993), perhaps due to cultural differences in interpretation, to be identified and accounted for during scoring. Empirical studies that ignore this information during scoring may introduce unnecessary measurement error or conflate actual group differences with measurement artifacts, potentially limiting the validity of their findings.

In this study, we investigated the value of IRT-based strategies for harmonizing measures of general cognitive performance across international surveys on aging using data from HRS and ELSA. We compared measurement consequences of using an IRT score: 1) based solely on the set of nine items common to both surveys; 2) using the common set of nine items, but adjusting for differentially functioning items; and 3) using *all* available items from each survey, with adjustment for items that show DIF. The main hypothesis being tested is that IRT scores based on all available items from each survey, adjusted for DIF, will have better measurement precision than IRT scores based on the common set only, whether adjusted for DIF or not.

METHODS

Data Sources

Data are drawn from the HRS and ELSA. These international surveys are both longitudinal and share focus on the social, economic and health aspects of the lives of people 50 and older. The HRS is nationally representative of the U.S. and has been ongoing since 1992 (Juster & Suzman, 1995). The ELSA is representative of the U.K. and is one of a family of surveys patterned after the HRS that now extends to many countries (Banks, Marmot, Oldfield, & Smith, 2006). In this study, we used closely aligned years of data with HRS data from 2002 and ELSA data from March 2002 – March 2003 (Wave 1; Release 2).

Our analyses focus on participants 65 or older who were administered cognitive assessments in each survey. Final sample sizes for our analyses were 9,471 in HRS and 5,444 in ELSA. Age, sex, and education of the sample by survey are presented in Table 1. The dichotomous education variable uses items specific to each survey: HRS provides items on completed education and degrees; ELSA provides a 7-level categorical variable (500 cases classified as “foreign/other” were excluded).

Measures

Nine measures were common to both surveys. Four of these items assessed *orientation* to date (i.e., today’s date, month, year and day of week). Respondents were asked “What is today’s date?” Credit was given for the correct date, month, year and day of week separately. Probes were used for components not reported spontaneously. In addition, three numeracy items, one related to disease prevalence, another to savings and the third on lottery winnings as well as two word recall items, immediate and delayed, were included in both HRS and ELSA.

Each survey also fielded unique items. The HRS fielded 16 cognitive items that were not found in ELSA. These items included counting backwards (from 20 and from 86), serial subtraction from 100 by 7, defining five words from the Wechsler Adult Intelligence Scale (WAIS-R), naming the president and vice-president, and recognizing and naming 2 familiar items (cactus and scissors). The four items administered only in ELSA included two numeracy items. One involved calculating a half price discount. The other involved determining the original cost of a used car priced at two-thirds of its new car price. Naming animals and letter cancelation were the two remaining ELSA-only items. Further details on these assessments are available from the documentation for each survey. (<http://hrsonline.isr.umich.edu>; <http://www.esds.ac.uk/longitudinal/access/elsa/15050.asp>).

An important IRT assumption is that a set of items measures a single, or unidimensional, construct (Embretson & Reise, 2000; Stout, 1990). We used exploratory factor analysis to examine this assumption for the set of cognitive assessments in our samples. This analysis was conducted separately for HRS and ELSA as contemporary restrictions in missing data estimation prevent testing for dimensionality for HRS and ELSA items together (Curran et al., 2008). At present, no definitive criterion for determining unidimensionality exists. However, “sufficient” unidimensionality for IRT analysis (McHorney & Cohen, 2000) may be demonstrated if proportion of the variance explained by the first factor is 20% (Reckase, 1979) and if the ratio of eigenvalues between the first and second factor is 4 (Reeve et al., 2007). Finally, strong factor loadings (>0.40) observed for all items on the first factor provide support of sufficient unidimensionality for valid IRT modeling. As a sensitivity analysis, we also implemented bifactor models to examine whether study results would substantially change if excess correlation among the numeracy items were accounted for. Bifactor models were implemented using Mplus (version 7.11, Muthen & Muthen, 1998–2008).

Score Derivation

Using the cognitive measures available from both surveys, we created three alternative sets of IRT scores based on: 1) common items set; 2) DIF-adjusted common items set; 3) and DIF-adjusted all items set. The first score set was generated using IRT-estimated parameters for the nine items that were fielded in the two surveys, assuming no DIF was present. The second set uses the same common items, but DIF was evaluated and adjusted for during IRT modeling and scoring. The final set of scores adjusted for identified DIF items and added survey-specific items to the item set for parameter estimation and scoring. For all IRT modeling and scoring, we used the HRS sample as the reference group. Parameter and score estimates were scaled to the HRS (with HRS group mean set to 0; each unit of the scale = 1 standard deviation of the HRS sample).

Estimating Item Parameters—The first step in generating each IRT score set is the estimation of item parameters for both binary (e.g., correct, incorrect) and ordinal (e.g., partial credit for WAIS vocabulary items: 2 for completely correct, 1 for partially correct; quartiles for naming animals and letter cancellation; score ranging between 0 and 7 for immediate or delayed word recall) items. We collapsed scores >7 on the two word recall items to 7 and created quartiles based on the distribution of the letter cancellation and naming animal tests to facilitate IRT modeling. We used the graded response model (GRM, Samejima, 1969), which can accommodate both binary and ordinal items with different numbers of categories, to model item characteristics and to generate scores. The GRM estimates one discrimination (a) and $k-1$ boundary location ($b_1 \dots b_{k-1}$) parameters, where k = number of response categories, for each item. The a parameter reflects the ability of an item to discriminate among persons with different levels of underlying cognitive performance. Higher a values indicate better discrimination. For binary items, the GRM is equivalent to the 2-parameter logistic model and the item location parameter (b) is the point on the cognitive performance scale where the probability of responding correctly to the item is 50%. For ordinal assessments, $k-1$ boundary parameters are estimated for k response categories. For example, the letter cancellation item is categorized in quartiles: the 3 boundary parameters were based on the first quartile relative to all others (b_1), the first two quartiles relative to the last two (b_2), and the first three quartiles relative to the last (b_3). We implemented IRT models using *Multilog* (Thissen, Chen & Bock, 2003).

DIF identification—Likelihood ratio (LR) difference tests were used to test whether item parameters functioned differently by survey. As part of these tests, we first identified a set of “anchor items,” or items that do not demonstrate DIF, among the nine common items (Teresi et al., 2007). Identification of anchor items involved iterative LR tests to identify and exclude items that show DIF. In these LR tests, an assumption is made that all items other than the item being tested serve as adequate anchors in the initial round. Subsequent LR tests are performed only within the set of preliminary “anchor items” identified by the previous LR test. Additional items identified with DIF are excluded from the anchor set. This process is repeated until the set of anchor items include no items demonstrating DIF.

Using the final set of anchor items, we tested for a difference in slope and location parameters by survey for each non-anchor item. To complement the statistical testing for

DIF, we examined the magnitude of the DIF for each item by comparing the item characteristic curves (ICC) for each survey. The ICCs are estimated from the model in which item parameters showing DIF are freely estimated for each survey. These curves plot the probability of endorsing the item over the range of underlying cognitive performance. Differences in these curves for the two surveys reveal the magnitude and direction of the DIF at the item-level. Non-overlapping ICCs by survey indicate DIF; coincident curves reflect absence of DIF.

Given the large samples in our study, we considered both results from statistical LR tests and the magnitude of the DIF identified through graphic analysis in determining whether the item should be modeled separately by survey before scoring. Specifically, if DIF was identified after the Benjamini-Hochberg adjustment for multiple comparisons (Thissen, Steinberg, & Kuang, 2002; Benjamini & Hochberg, 1995), we generated graphic displays of the ICC curves for each group to illustrate the nature of the DIF over the entire range of cognitive performance. For significant DIF items, we also produced expected score difference along the latent trait. To facilitate comparison between binary and ordinal items with different score range, we standardized the score range by dividing by the item score range. A between group difference of 0.16 in the expected score for a binary item (e.g., 1=correct, 0=incorrect) would have a scaled difference of 0.16; a 0.89 difference for an ordinal item with a score range of 3 would have a scaled difference of 0.30. These score differences were subsequently used to evaluate the importance of the DIF.

To ensure that only important DIF was mapped, any item with scaled difference of 0.10 (Perkins, Stump, Monahan & McHorney, 2006) at any point along the latent trait was selected for further evaluation using the standardization methods described by Dorans and Kulik (2006). This approach evaluates the overall impact of the score difference across the range of latent trait to determine if the item should be modeled for DIF. The sample size for ELSA at each score level served as the weight to calculate the standardized p-difference (STD PDIF), which can range between -1.0 and 1.0. We determined items with STD PDIF values between -0.05 and 0.05 to have negligible DIF as recommended (Dorans and Kulik, 2006).

IRT Scoring

Common items set scores: For these scores, we estimated only one set of parameters for each item and assumed no DIF existed. Using these parameters, scores were estimated for each respondent based on their responses to the nine common items.

DIF-adjusted common items set scores: These scores account for DIF identified among the nine common items. If an item did not demonstrate DIF, only one set of parameters is estimated for the item but items with DIF is modeled separately for HRS and ELSA respondents. Scores estimated for each respondent using this set of parameters is adjusted for DIF.

DIF-adjusted all items scores: These scores are based on parameters estimated using all available cognitive assessments from the two surveys, with separate parameters estimated for items showing DIF.

Analysis

We compared the standard error associated with score estimates at each point along the cognitive performance spectrum among the scoring methods. The size of the standard error is influenced by the discrimination and the number of items located within a region of the underlying trait. Higher discrimination and more items located at a given score reduce standard error for that score (since items measure best at their location described by the b parameter). We hypothesized that adding survey-specific items would improve measurement precision and thus lower standard error of the score estimates.

The practical impact of the different scoring strategies cannot be inferred directly from the standard error functions, as these functions do not account for the distribution of cognitive performance in the HRS and ELSA samples. Specifically, if most survey respondents are located in a region of the trait where the differences in standard errors by scoring method are small, minimal differences in the standard error for the overall sample would be observed. To examine overall impact, we compared, for each sample, the average standard errors across the different scoring methods.

RESULTS

Age and gender distribution of HRS and ELSA respondents were similar, although a slightly higher percentage of HRS respondents were aged 85 and older and female. Differences in education were more pronounced, with HRS respondents having greater educational attainment than ELSA respondents (Table 1).

Unidimensionality

Table 2 presents the results from the exploratory factor analysis. The goal is to determine if sufficient unidimensionality exists for valid IRT modeling. For both HRS and ELSA, loadings on the first factor were 0.40 for all items. In addition, the proportion of variance explained by the first factor was 70% for HRS and 80% for ELSA and the ratio of the eigenvalue of the first factor to the second factor was 5.32 and 4.93. These results all exceed suggested criteria (McHorney & Cohen, 2000; Reckase, 1979; Reeve et al., 2007) and indicate that IRT modeling for these items is appropriate.

Item Parameters and DIF Findings

Table 3 presents the discrimination (a) and location (b) parameters for items that were unique to each survey or that did not show DIF across surveys. Item discrimination varied, ranging from 0.70 (WAIS Vocabulary) to 3.08 (Serial Subtraction by 7: 4th task). Furthermore, most of the location parameters are negative. This indicates that measurement of cognitive performance in these surveys is in general better in the more impaired range, since easier items have location parameters of lower numerical value than harder items. Finally, 16 items were fielded solely in the HRS. Therefore, we expect measurement precision to be greater for this sample given the information contributed by these extra items.

The pattern of location parameters observed is as expected. For example, in the serial subtraction tasks (in which the respondent is asked to subtract 7 from 100 and continue to subtract 7 from the prior answer), the easiest task is the first subtraction ($b = -0.99$) and the hardest is the last subtraction ($b = -0.18$). Naming the president was easier ($b = -2.44$) than naming the vice president ($b = -0.73$). Counting backwards from 20 is also easier ($b = -2.13$) than counting back from 86 ($b = -1.43$).

Of the 9 common items, three demonstrated DIF across the two surveys (Table 4). Day of date and delayed word recall demonstrated discrimination and location DIF, while the immediate word recall demonstrated DIF only for location. For all three items, the location parameter values were higher (less negative or more positive) for ELSA respondents suggesting that these items are more challenging for ELSA respondents compared to HRS respondents with the same level of cognitive performance.

Score Comparisons

Figure 1 presents the standard error functions for the three scoring methods for both a unidimensional model and a bifactor model. Compared to the two common items scores, the scores based on all available items had consistently smaller standard errors (or greater measurement precision), with the greatest difference at the lower end of cognitive performance between 0 and -2.5 , the region of the trait where most of the survey-specific items are located (Table 3). The standard error functions were comparable for the two common items scores, although the two scores alternated having higher standard errors across different cognitive performance levels. The standard error patterns for the three scores were generally similar in the bi-factor model, suggesting that excess covariation among numeracy items had limited influence on the measurement precision results.

Impact of DIF Adjustment and Including All Items on Measurement Precision

The average standard errors of the three scores for HRS and ELSA respondents are presented in Table 5. For HRS respondents, average standard errors become progressively smaller from the common items scores to the DIF-adjusted common items scores to the DIF-adjusted all items scores. In contrast, the average standard error was modestly smaller for the common items scores compared to the other two DIF-adjusted scores for ELSA respondents. Most results were comparable whether the unidimensional or bifactor model was used, except for the all items DIF-adjusted score in ELSA respondents. For this score, average standard errors were slightly larger than those for the common items score with the unidimensional model, but were modestly smaller when the bifactor model was used.

The difference in the behavior of the three scores between HRS and ELSA samples appears to be related to the number and location of survey-specific items, the shifts in item location resulting from DIF adjustment, and the underlying distribution of cognitive performance of each sample (Figure 2). In the ELSA sample, DIF adjustment shifted item locations higher for the common items, resulting in higher standard errors at lower cognitive performance levels but lower standard errors at higher cognitive performance levels for the two DIF-adjusted scores. Furthermore, the four ELSA-specific items primarily contribute to measurement precision at higher cognitive performance levels (Table 3). How DIF

adjustment and addition of ELSA-specific items affects measurement precision depends on the level of cognitive performance. ELSA respondents in region A of the cognitive performance trait, shown in Figure 2, have the smallest average standard errors with the common items scores, modestly larger standard errors for all items DIF-adjusted scores and the largest standard errors for the common items DIF-adjusted scores. The situation differs for ELSA respondents with higher cognitive performance in region B of the trait, for whom the all items DIF-adjusted scores produced the smallest standard errors. As a large proportion of ELSA respondents are at lower levels of cognitive performance where the common items score had the smallest standard errors, the average standard errors for the ELSA sample were lowest for these scores (Table 5).

Although the pattern of standard error functions for HRS respondents differs, the same issues appear to influence the average standard errors for the HRS sample reported in Table 5. The two common items scores were generally similar, although the DIF-adjusted scores had slightly smaller standard errors. However, the addition of HRS-specific items substantially reduced standard errors for the all items DIF-adjusted scores, particularly at lower levels of cognitive performance (trait <0.0), consistent with the item locations of most HRS-specific items (Table 3). From Figure 2, it is apparent that the practical impact of different scores again depend on the region of the trait HRS respondents are located in. For HRS respondents located approximately -1.0 on the trait (region A), the all items DIF-adjusted scores had substantially smaller standard errors than the common item scores (with and without DIF adjustment). In contrast, the standard errors for all three scores were very similar for HRS respondents located near 0.5 (region B). A large proportion of HRS respondents had trait levels <0.5 where the all items DIF-adjusted scores performed best. These findings from Figure 2 appear to account for the pattern of average standard errors for the HRS sample in Table 5.

DISCUSSION

Our findings demonstrated that IRT methods can be effectively utilized for harmonizing cognitive performance assessments across two major international surveys. Our study provided insights on the effects of this strategy on score comparability and measurement precision. First, differential item function was observed for cognitive measures, highlighting the often unrecognized role that measurement non-equivalence can play in international group comparisons, even when common test items are used. In particular, substantial DIF was found for the two word recall items, which have been included in a number of cross-country comparisons of cognitive performance (Rohwedder & Willis, 2010; Oksuzyan et al. 2010; Skirbekk, Loichinger & Weber, 2012). Assuming measurement equivalence when the same items or set of items are used can conflate true group differences with measurement artifacts in how groups respond to measures. Adjusting for DIF has different implications for measurement precision, depending on the level of cognitive performance respondents possess. However, accounting for differentially functioning items before pooling HRS and ELSA data for comparative studies is still important to ensure score validity.

Second, our findings showed that the number and item location of available survey-specific items interact with the distribution of the underlying trait to influence the effect of using all

available measures for estimating scores. For example, in the HRS, adding the 16 survey-specific items greatly improved measurement in the lower range of the cognitive performance trait where the location parameters for these items indicate these items are most informative. Since nearly half of the HRS sample in our study were in the lower cognitive performance levels, using the all-items DIF-adjusted scores substantially lowered the average standard error for the entire sample. This resulted in greater overall measurement precision for HRS respondents. In contrast, ELSA contributed only 4 items with location parameters in the upper range of the cognitive performance trait. Since only a small proportion of ELSA respondents were in this region, adding these items did not reduce the average standard errors for the overall sample compared with using the common items scores. In fact, location parameters for the three DIF items shifted toward higher cognitive performance (i.e., these items were more difficult for the ELSA respondents). This improved measurement at upper cognitive performance levels, while reducing measurement precision at lower cognitive levels. However, since most ELSA respondents were in the lower cognitive performance regions, the average standard errors for the DIF-adjusted common items scores were higher than that for the common items scores without DIF adjustment. The situation was similar for the all items DIF-adjusted scores in the ELSA sample. The addition of the four ELSA-specific items improved measurement at the upper cognitive performance region. However, as only a small proportion of ELSA respondents have cognitive performance at the higher levels where the ELSA-specific items performed best, average standard errors for the all items DIF-adjusted scores was similar to that for the DIF-adjusted common scores. Regardless of average standard error of scores observed for each sample, the DIF-adjustment correctly apportions error to appropriate parts of the range.

The consequence of DIF adjustment and adding survey-specific items will differ for individuals at different levels of the trait even within the same sample. For HRS respondents, measurement precision for individuals at the lower end of cognitive performance would be substantially improved with the addition of the HRS-specific items, while scores for HRS respondents with high cognitive performance levels would not differ much regardless of which score was used. In contrast, while using the four ELSA-specific items would improve score precision for ELSA respondents with higher cognitive performance, it would not have the same effect for ELSA respondents at lower cognitive performance levels.

Our IRT analysis also offers insights on the measurement properties of cognitive items included in two major surveys on aging. For the most part, the common items demonstrated relatively good discrimination. However, both common and survey-specific items provided the most information in the lower range of cognitive performance. Few cognitive performance measures assessed the less impaired range of functioning (as an item that only very high functioning individuals would respond to correctly would do). This suggests that the measures being fielded by these surveys do better at discriminating among individuals at lower cognitive performance levels than individuals at higher cognitive performance levels. From a clinical or policy perspective, it may be appropriate to ensure good measurement properties for individuals with more severe cognitive impairment. However, the current set of measures is less useful for studying individuals with good to excellent cognition.

Our study has several limitations. First, findings are applicable only for the cognitive items from the surveys examined in the study. The practical value of the three scoring strategies may change under different sample distributions and balance of common and survey-specific items with different item parameters. Adding more difficult items which discriminate among people at less impaired levels of cognitive performance, for example, could improve the precision of the scale among community-living persons. Second, while evidence of differential item functioning suggests that pooling data for the common cognitive items from HRS and ELSA without accounting for DIF would produce non-comparable scores, our analysis does not provide an explanation for the observed DIF. Specifically, although the DIF observed suggests response differences between U.S. and U.K. populations, whether the cause is cultural or due to methodologic differences cannot be determined. Finally, to improve our ability to link across the two surveys, we analyzed a broad set of cognitive assessments, including several numeracy measures, as a unidimensional trait. Therefore, the latent trait in our study may not reflect the more complex structure of cognitive performance reported in other studies (Herzog & Wallace, 1997; McArdle, Fisher & Kadlec, 2007). However, our analyses indicate that the items we used have sufficient unidimensionality for IRT modeling and are adequate for investigating the value of an IRT approach to linking cognitive measures across surveys. Furthermore, findings from the bi-factor model that account for the numeracy items and the unidimensional model were similar, suggesting that dimensionality issues did not substantially affect study findings and conclusions.

International surveys represent an important opportunity for cross-national investigations into issues associated with aging and cognitive performance. Our study demonstrated the feasibility and value of using IRT methods to improve comparability and precision of cognitive performance scores by accounting for DIF and utilizing all measures available in each survey. This approach may be useful for harmonizing cognitive measures across other nationally and internationally representative surveys, including surveys conducted in other languages, such as the Mexican Health and Aging Study (MHAS), one of the sister surveys of the HRS. Future investigations should also examine the contexts, such as the number of common items and the distribution of common and unique items, where these methods would produce the greatest measurement benefits. However, our study provides the methodological underpinnings for conducting valid cross-national research on cognitive performance and the needs of the cognitively impaired.

References

- Bago d'Uva T, O'Donnell O, van Doorslaer E. Differential health reporting by education level and its impact on the measurement of health inequalities among older Europeans. *Int J Epidemiol*. 2008; 37:1375–1383. [PubMed: 18676985]
- Banks J, Marmot M, Oldfield Z, Smith J. Disease and disadvantage in the United States and in England. *Journal of the American Medical Association*. 2006 May; 295(17):2037–2045. [PubMed: 16670412]
- Ball K, Berch DB, Helmers KF, Jobe JB, Leveck MD, Marsiske M, Morris JN, Rebok GW, Smith DM, Tennstedt SL, Unverzagt FW, Willis SL. for the ACTIVE Study Group. Effects of cognitive training interventions with older adults: A randomized controlled trial. *JAMA*. 2002; 288:2271–2281. [PubMed: 12425704]
- Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*. 1995; 57(1):289–300.

- Curran PJ, Hussong AM, Cai L, Huang W, Chassin L, Sher KJ, Zucker RA. Pooling data from multiple longitudinal studies: the role of item response theory in integrative data analysis. *Dev Psychol.* 2008; 44(2):365–380. [PubMed: 18331129]
- Dorans NJ, Kulik E. Differential Item Functioning on the Mini-Mental State Examination: An Application of the Mantel-Haenszel and Standardization Procedures. *Medical Care.* 2006; 44(11):S107–S114. <http://dx.doi.org/10.1097/01.mlr.0000245182.36914.4a>. [PubMed: 17060817]
- Embretson, SE.; Reise, SP. Item response theory for psychologists. Mahwah, NJ: Erlbaum; 2000.
- Ferraro, FR. Minority and Cross-cultural Aspects of Neuropsychological Assessment. Taylor and Francis; 2002.
- Figueras, J.; McKee, M. Health Systems, Health, Wealth and Societal Well-being: Assessing the case for investing in health systems. Berkshire, England: McGraw Hill Open University Press; 2012.
- Gross AL, Parisi JM, Spira AP, Kueider AM, Ko JY, Saczynski JS, Samus QM, Rebok GW. Memory training interventions for older adults: A meta-analysis. *Aging Ment Health.* 2012; 16(6): 722–734. [PubMed: 22423647]
- Gross AL, Rebok GW. Memory training and strategy use in older adults: Results from the active study. *Psychology and Aging.* 2011
- Hambleton, RK.; Swaminathan, H.; Rogers, HJ. Fundamentals of Item Response Theory. Newbury Park: Sage Publications; 1991.
- Hendrie HC. Lessons learned from international comparative crosscultural studies on dementia. *The American Journal of Geriatric Psychiatry : Official Journal of the American Association for Geriatric Psychiatry.* 2006; 14(6):480–488. [PubMed: 16731716]
- Herzog, AR.; Rodgers, WL. Cognitive performance measures in survey research in older adults. In: Schwartz, N.; Park, DC.; Knauper, B.; Sudman, S., editors. *Aging, cognition, and self-reports.* Philadelphia, PA: Psychology Press; 1999. p. 327-340.
- Herzog AR, Wallace RB. Measures of cognitive functioning in the AHEAD study. *Journals of Gerontology, Series B: Psychological Sciences and Social Sciences.* 1997; 52:P37–P48.
- Holland, PW.; Wainer, H. Differential Item Functioning. Hillsdale, N.J.: Lawrence Erlbaum Assoc. Inc.; 1993.
- Juster FT, Suzman R. An overview of the Health and Retirement Study. *The Journal of Human Resources.* 1995; 30(Suppl.):S7–S56.
- King G, Murray CJL, Salomon JA, Tandon A. Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review.* 2004; 98:191–207.
- Kinsella, K.; He, W. U.S. Census Bureau, International Population Reports, P95/09-1, *An Aging World: 2008.* Washington, DC: U.S. Government Printing Office; 2009.
- McArdle JJ, Fisher GG, Kadlec KM. Latent variable analyses of age trends of cognition in the Health and Retirement Study, 1992–2004. *Psychology and Aging.* 2007; 22(3):525–545. [PubMed: 17874952]
- McHorney CA, Cohen AS. Equating Health Status Measures with Item Response Theory: Illustrations with Functional Status Items. *Med Care.* 2000; 38(9, Suppl II):II43–II59. [PubMed: 10982089]
- Muthén, LK.; Muthén, BO. *Mplus User's Guide.* Seventh. Los Angeles, CA: Muthén & Muthén; 1998–2012.
- Nell, V. Cross-cultural neuropsychological assessment: Theory and practice. Lawrence Erlbaum Associates; 1999.
- Njegovan V, Hing MM, Mitchell SL, Molnar FJ. The hierarchy of functional loss associated with cognitive decline in older persons. *The Journals of Gerontology. Series A, Biological Sciences and Medical Sciences.* 2001; 56(10):M638–M643.
- Oksuzyan A, Crimmins E, Saito Y, O'Rand A, Vaupel JW, Christensen K. Cross-national comparison of sex differences in health and mortality in Denmark, Japan, and the U.S. *Eur J Epidemiol.* 2010; 25:471–480. [PubMed: 20495953]
- Perkins AJ, Stump TE, Monahan PO, McHorney CA. Assessment of differential item functioning for demographic comparisons in the MOS SF-36 health survey. *Quality of Life Research.* 2006; 15(3): 331–348. [PubMed: 16547771]

- Pluijm SM. A harmonized measure of activities of daily living was a reliable and valid instrument for comparing disability in older people across countries. *Journal of Clinical Epidemiology*. 2005; 58:1015–1023. [PubMed: 16168347]
- Reckase M. Unifactor latent trait models applied to multifactor tests: Results and implications. *J Educ Stat*. 1979; 4:207–230.
- Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA, Thissen D, Revicki DA, Weiss DJ, Hambleton RK, Liu H, Gershon R, Reise SP, Lai JS, Cella D. on behalf of the PROMIS Cooperative Group. Psychometric Evaluation and Calibration of Health-Related Quality of Life Item Banks: Plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical Care*. 2007; 45:S22–S31. [PubMed: 17443115]
- Rohwedder S, Willis RJ. Mental retirement. *Journal of Economic Perspectives*. 2010; 24(1):119–138. [PubMed: 20975927]
- Roth, M.; Huppert, FA.; Mountjoy, CQ.; Tym, E. *The revised Cambridge Examination for Mental Disorders of the Elderly*. Second. Cambridge: Cambridge University Press; 1999.
- Salomon JA, Tandon A, Murray CJL. Comparability of self-rated health: Cross sectional multi-country survey using anchoring vignettes. *British Medical Journal*. 2004; 328:258–261. [PubMed: 14742348]
- Samejima F. Estimation of latent trait ability using a response pattern of graded scores. *Psychometrika*. Monograph Supplement. No. 17. 1969
- Schoeni RF, Ofstedal MB. Key themes in research on the demography of aging. *Demography*. 2010; (47-Supplement):S5–S15. <http://dx.doi.org/10.1353/dem.2010.0001>. [PubMed: 21302420]
- Skirbekk V, Loichinger E, Weber D. Variation in cognitive functioning as a refined approach to comparing aging across countries. *Proceedings of the National Academy of Sciences*. 2012; 109(3):770–774.
- Stout WF. A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika*. 1990; 55:293–325.
- Teresi JA, Ocepek-Welikson K, Kleinman M, Cook KF, Crane PK, Gibbons LE, Morales LS, Orlando-Edelen M, Cella D. Evaluating measurement equivalence using the item response theory log-likelihood ratio (IRTLR) method to assess differential item functioning (DIF): Applications (with illustrations) to measures of physical functioning ability and general distress. *Quality of Life Research*. 2007; 16:43–68. [PubMed: 17484039]
- Thissen, D.; Chen, W.; Bock, D. *Multilog (computer program)*. version 7.0 for windows. Chicago: Scientific Software Inc.; 2003.
- Thissen D, Steinberg L, Kuang D. Quick and easy implementation of the Benjamin-Hochberg Procedure for Controlling the False Positive Rate in Multiple Comparisons. *Journal of Educational and Behavior Statistic*. 2002; 27:77–83.
- United Nations. *World Population Ageing*. United Nations (Sales #E.07.XIII.5, ISBN 978-92-1-151432-2); 2007.
- Verhey FRJ, Huppert FA, Korten ECCM, Houx P, DeVugt M, Van Lang M, DeDeyn PP, Saerens J, Neri M, De Vreese L, Pena-Casanova J, Bohm P, Stoppe G, Fleischmann U, Wallin A, Hellstrom P, Middelkoop H, Bollen W, Kliinkenberg EL, Derix MMA, Jolles J. Cross-national comparisons of the Cambridge Cognitive Examination- revised: the CAMCOG-R. *Age and Ageing*. 2003; 32:534–540. [PubMed: 12958004]
- Willis SL, Tennstedt SL, Marsiske M, Ball K, Elias J, Koepke KM, Morris JN, Rebok GW, Unverzagt FW, Stoddard AM, Wright E. for the ACTIVE Study Group. Long-term Effects of Cognitive Training on Everyday Functional Outcomes in Older Adults. *JAMA*. 2006; 296(23):2805–2814. [PubMed: 17179457]
- Yaffe K, Fiocco AJ, Lindquist K, Vittinghoff E, Simonsick EM, Newman AB, Satterfield S, Rosano C, Rubin SM, Ayonayon HN, Harris RTB. Predictors of maintaining cognitive function in older adults: the Health ABC study. *Neurology*. 2009; 72(23):2029–2035. [PubMed: 19506226]

(a) Unidimensional IRT Model

(b) Bifactor Model

Figure 1.
Standard Error by Scoring Method and IRT Model

Figure 2.

Cognitive Performance Trait Distribution Effects on Sample Standard Error

Note: For each survey, top graph reflects standard error of the three scores across different cognitive performance levels (x-axis, higher value indicates greater cognitive performance). Bottom graph provides the distribution of the sample.

Table 1

Sample Characteristics Overall and by Survey (Unweighted)

Characteristic	HRS	ELSA
Sample Size	9471	5444
Age Group (%)		
65–69	31.54	31.25
70–74	24.44	26.78
75–79	18.83	19.80
80–85	14.90	14.35
85+	10.29	7.83
Gender (%)		
Male	40.59	44.55
Education (%)		
Secondary/high school or less	63.77	74.87
Beyond secondary/high school	36.21	15.50

Table 2

Dimensionality of Cognitive Items in HRS and ELSA

Item Content	HRS		ELSA	
	Factor 1	Factor 2	Factor 3	Factor 2
Day (of date)	0.51	0.37	-0.21	0.55
Month (of date)	0.65	0.52	-0.24	0.84
Year (of date)	0.77	0.35	-0.13	0.85
Day of week	0.61	0.48	-0.22	0.81
Numeracy – disease	0.70	-0.21	0.12	0.73
Numeracy – savings	0.54	-0.26	0.12	0.53
Numeracy - lottery	0.65	-0.25	0.01	0.63
Word Recall (Immediate)	0.56	0.25	0.01	0.67
Word Recall (Delayed)	0.57	0.29	-0.02	0.64
Count back –20	0.70	0.00	-0.02	--
Count back – 86	0.67	-0.04	-0.03	--
Tool to cut paper	0.53	0.25	0.21	--
Name of prickly plant	0.68	0.09	0.24	--
Name of president (last name)	0.73	0.31	0.06	--
Name of vice president (last name)	0.63	0.14	-0.00	--
WAIS vocab	0.58	-0.03	0.41	--
WAIS vocab	0.41	0.00	0.36	--
WAIS vocab	0.45	-0.04	0.38	--
WAIS vocab	0.40	0.00	0.38	--
WAIS vocab	0.48	-0.09	0.38	--
1st serial 7 subtraction	0.83	-0.36	-0.11	--
2 nd serial 7 subtraction	0.74	-0.31	-0.20	--
3 rd serial 7 subtraction	0.73	-0.34	-0.27	--
4 th serial 7 subtraction	0.78	-0.39	-0.26	--
5 th serial 7 subtraction	0.77	-0.37	-0.24	--
Numeracy -- halfprice	--	--	--	0.88
				0.15

Item Content	HRS		ELSA	
	Factor 1	Factor 2	Factor 3	Factor 2
Numeracy – twothirds	--	--	--	0.59
Naming animals	--	--	--	0.65
Letter cancelation	--	--	--	0.53
Eigenvalue *	10.16	1.91	1.28	6.28
Proportion Explained by Factor	0.70	0.13	0.09	0.80
Ratio of Eigenvalue between adjacent factors	5.32	1.49	--	4.93
				--

*Factor retained (Eigenvalue >1.0); exploratory principal factor analysis

Table 3

Item Parameters for Survey-Specific Items and Items that show no DIF

Item Content	Survey(s) Fielded	Item Discrimination	Item Location		
			<u>B1</u>	<u>B2</u>	<u>B3</u>
Orientation – Day of Week	HRS & ELSA	1.45	-2.57		
Orientation – Month (of date)	HRS & ELSA	1.57	-2.50		
Orientation – Year (of date)	HRS & ELSA	1.88	-2.14		
Numeracy – Disease	HRS & ELSA	1.63	-0.49		
Numeracy – Savings	HRS & ELSA	1.39	2.52		
Numeracy – Lottery	HRS & ELSA	1.46	0.63		
Count back – 20	HRS	1.69	-2.13		
Count back – 86	HRS	1.47	-1.43		
Name tool to cut paper	HRS	1.22	-4.03		
Name of prickly plant	HRS	1.48	-1.89		
Name President (last name)	HRS	1.62	-2.44		
Name Vice President (last name)	HRS	1.19	-0.73		
WAIS vocabulary* (repair/conceal)	HRS	1.18	-2.56	-1.91	
WAIS vocabulary*	HRS	0.70	-3.76	0.44	
WAIS vocabulary*	HRS	0.82	-0.97	0.89	
WAIS vocabulary*	HRS	0.73	-1.13	1.95	
WAIS vocabulary*	HRS	1.09	1.28	2.96	
Serial Subtraction by 7: 1 st task	HRS	2.88	-0.99		
Serial Subtraction by 7: 2 nd task	HRS	2.27	-0.31		
Serial Subtraction by 7: 3 rd task	HRS	2.38	-0.20		
Serial Subtraction by 7: 4 th task	HRS	3.08	-0.25		
Serial Subtraction by 7: 5 th task	HRS	2.95	-0.18		
Numeracy – Half Price	ELSA	2.11	-1.60		
Numeracy – Two Thirds	ELSA	1.22	0.55		

Item Content	Survey(s) Fielded	Item Discrimination	Item Location		
			B1	B2	B3
Letter Cancellation	ELSA	1.59	-0.81	0.29	1.55
Naming animals	ELSA	1.00	-0.90	0.48	1.93

*Wechsler Adult Intelligence Scale" (word from word list #1/word from word list #2). Word list 1 or 2 randomly assigned to respondent; A (column 3) refers to item discrimination, a higher value of A reflects a stronger relationship of the item to cognitive performance; B (columns 4-6) indicates the item location or where the item (or item category) measures best on the cognitive performance trait, e.g., a lower B value indicates the item taps an "easier" functioning task.

Table 4

DIF in Common Cognitive Items from the Final All Items Model

Item	DIF	Survey	A	B1	B2	B3	B4	B5	B6	B7
Orientation – Day of date	A & B	HRS	0.87	-2.00						
		ELSA	0.75	-1.29						
Word Recall Immediate	B	HRS	1.51	-2.90	-2.62	-1.99	-1.19	-0.34	0.49	1.34
		ELSA	1.51	-2.67	-2.03	-1.52	-0.77	0.04	0.96	1.97
Word Recall Delayed	A & B	HRS	1.24	-2.02	-1.63	-1.03	-0.30	0.50	1.34	2.19
		ELSA	1.93	-0.87	-0.54	-0.07	0.50	1.19	1.90	2.73

A (column 4) refers to item discrimination, a higher value of A reflects a stronger relationship of the item to cognitive performance; B (columns 5–11) indicates the item location or where the item (or item category) measures best on the cognitive performance trait, e.g., a lower B value indicates the item taps an “easier” functioning task.

Table 5

Average Standard Errors for HRS and ELSA Respondents by Scoring Approaches

Survey	# Items in Survey	Common Items with no DIF adjustment	Common Items DIF- Adjusted	All Available Items DIF- adjusted
Unidimensional Model				
HRS	25	0.422	0.407	0.363
ELSA	13	0.410	0.433	0.418
Bifactor Model				
HRS	25	0.424	0.411	0.371
ELSA	13	0.412	0.431	0.403