Journals of Gerontology: Psychological Sciences
cite as: *J Gerontol B Psychol Sci Soc Sci*, 2021, Vol. 76, No. 9, 1767–1776
doi:10.1093/geronb/gbaa205
Advance Access publication November 28, 2020

Research Article

You Say Tomato, I Say Radish: Can Brief Cognitive Assessments in the U.S. Health Retirement Study Be Harmonized With Its International Partner Studies?

Lindsay C. Kobayashi, PhD,[1,2,*] Alden L. Gross, PhD,[3] Laura E. Gibbons, PhD,[4] Doug Tommet, MS,[5] R. Elizabeth Sanders, BA,[4] Seo-Eun Choi, PhD,[4] Shubhabrata Mukherjee, PhD,[4] Maria Glymour, ScD,[6] Jennifer J. Manly, PhD,[7] Lisa F. Berkman, PhD,[2] Paul K. Crane, MD,[4] Dan M. Mungas, PhD,[8] and Richard N. Jones, ScD[5,]

[1]Center for Social Epidemiology and Population Health, Department of Epidemiology, University of Michigan School of Public Health, Ann Arbor. [2]Harvard Center for Population and Development Studies, Harvard T. H. Chan School of Public Health, Cambridge, Massachusetts. [3]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health and Johns Hopkins University Center on Aging and Health, Baltimore, Maryland. [4]Department of Medicine, School of Medicine, University of Washington, Seattle. [5]Department of Psychiatry and Human Behavior, Alpert Medical School, Brown University, Providence, Rhode Island. [6]Department of Epidemiology and Biostatistics, University of California, San Francisco. [7]Department of Neurology and the Taubman Institute for Research on Alzheimer's Disease and the Aging Brain, Columbia University, New York, New York. [8]Department of Neurology, University of California, Davis, Sacramento.

*Address correspondence to: Lindsay C. Kobayashi, PhD, Center for Social Epidemiology and Population Health, Department of Epidemiology, University of Michigan School of Public Health, 1415 Washington Heights, Ann Arbor, MI 48109. E-mail: lkob@umich.edu

Received: June 10, 2020; Editorial Decision Date: November 10, 2020

Decision Editor: Shevaun Neupert, PhD, FGSA

Abstract

Objectives: To characterize the extent to which brief cognitive assessments administered in the population-representative U.S. Health and Retirement Study (HRS) and its International Partner Studies can be considered to be measuring a single, unidimensional latent cognitive function construct.
Methods: Cognitive function assessments were administered in face-to-face interviews in 12 studies in 26 countries (N = 155,690), including the U.S. HRS and selected International Partner Studies. We used the time point of the first cognitive assessment for each study to minimize differential practice effects across studies and documented cognitive test item coverage across studies. Using confirmatory factor analysis models, we estimated single-factor general cognitive function models and bifactor models representing memory-specific and nonmemory-specific cognitive domains for each study. We evaluated model fits and factor loadings across studies.
Results: Despite relatively sparse and inconsistent cognitive item coverage across studies, all studies had some cognitive test items in common with other studies. In all studies, the bifactor models with a memory-specific domain fit better than single-factor general cognitive function models. The data fit the models at reasonable thresholds for single-factor models in 6 of the 12 studies and for the bifactor models in all 12 of the 12 studies.
Discussion: The cognitive assessments in the U.S. HRS and its International Partner Studies reflect comparable underlying cognitive constructs. We discuss the assumptions underlying our methods, present alternatives, and future directions for cross-national harmonization of cognitive aging data.

Keywords: Cognitive function, Health survey, International comparison, Item response theory, Statistical harmonization

Population aging is a social, economic, and public health concern worldwide. In 2015, the global prevalence of dementia was estimated at 46 million people globally, and this figure is projected to nearly triple to 131.5 million by 2050 (Prince et al., 2015). Research into the drivers and outcomes of aging-related cognitive impairments and dementias is essential for their prevention, early detection, and appropriate management within families and communities and for health care systems. From a global perspective, dementia cases in low- and middle-income countries with large and rapidly aging populations are projected to rise in coming years to represent approximately 70% of the worldwide share by 2050 (Prince et al., 2015). Thus, the future of cognitive aging research requires a global perspective to address the changing worldwide distribution of cognitive impairments and dementias (Tollman et al., 2016). However, there has been little cross-national comparative research involving cognitive aging outcomes, and no research using cognitive assessments that have been statistically harmonized using modern psychometric methods to ensure consistency in measurement across countries at differing levels of economic development.

In recent years, several countries have launched nationally representative cohort studies of aging modeled after the U.S. Health and Retirement Study (HRS), a nationally representative cohort of American adults aged 50 and older who have been interviewed biennially since 1992 (Sonnega et al., 2014). The basic aspects of study design and questionnaire measures in these HRS International Partner Studies (IPS) are intended to be harmonized to facilitate cross-country comparisons. Some well-known examples of the HRS IPS include the English Longitudinal Study of Ageing (ELSA), the Chinese Health and Retirement Study (CHARLS), the Mexican Health and Aging Study (MHAS), and the Study on Global Ageing and Adult Health (SAGE; Kowal et al., 2012; Steptoe et al., 2013; Zhao et al., 2014). The HRS IPS all include brief cognitive batteries in their study interviews that assess multiple domains of cognitive function that are sensitive to aging-related changes. These studies use commonly administered measures, appropriate to each country, such as immediate and delayed recall of 10 words, the ability to count backward from 20 to 1 or serially backward from 100 by 7s, and the ability to correctly state the current date, day, month, year, and the country's president (or equivalent; Ofstedal et al., 2005). The cognitive batteries are intended to characterize individuals across a broad range of cognitive ability and especially those with mild cognitive impairment or probable dementia.

The original HRS battery has been translated and culturally adapted in various ways in each of the IPS. The content and number of cognitive test items vary across studies, as several have removed, substituted, or added items. For example, the ELSA cognitive battery added a verbal fluency item involving naming animals (Banks et al., 2006), whereas the object recognition item using a cactus was removed from the cognitive battery in South Africa, where cacti are not indigenous (Gómez-Olivé et al., 2018). Other adaptations were to help account for low population levels of literacy or numeracy in certain contexts, such as the backward 20 to 1 count being changed to a forward 1 to 20 count in South Africa (Gómez-Olivé et al., 2018). Translations of word recall lists also may introduce heterogeneity in cross-national batteries, as the interpretation, difficulty, lexical frequency, and visual imagery of words may differ across different languages. Furthermore, the HRS and each International Partner Study have different baseline years, with the HRS beginning in 1992 and the others as recently as 2015 (Table 1). In any given calendar year, the number of previously administered cognitive batteries varies across studies. Consequently, different study populations have differing degrees of practice with the batteries in any particular year, which may be important because the practice can improve scores on cognitive tests (Vivot et al., 2016). Practice effects in cognitive testing can introduce substantial artificial differences in average cognitive performance between study populations depending on the degree of previous practice (Bartels et al., 2010; Jones, 2015; Vivot et al., 2016). When comparing cognitive data across studies, it may be important to include samples with comparable prior cognitive testing experience, to minimize effects of differential practice effects on cross-national comparisons.

While there is a precedent for pairwise co-calibration of cognitive measures across studies (Chan et al., 2015), there has not yet been an attempt to simultaneously harmonize the cognitive measures from multiple studies. Item response theory (IRT) can be used to calibrate cognitive data when the number and content of cognitive items differ (Chan et al., 2015; Crane et al., 2008; Gross et al., 2015). IRT assumes the presence of an underlying latest trait that is sufficient to characterize a person's response, and that there is at least one "anchor" item that performs identically across study populations for which the cognitive data are being harmonized. As long as these assumptions hold, IRT can be used to statistically harmonize cognitive battery data from different studies even when batteries are not identical across studies (Gibbons et al., 2014). A multi-country statistical harmonization analysis using IRT methods would help identify the limits of the comparability of standard cognitive measures across diverse country contexts and elucidate opportunities for cross-national comparative research on the prevalence, distribution, and mechanisms of cognitive aging outcomes.

We therefore leveraged an IRT-based latent variable approach to investigate the degrees to which: (a) cognitive function data from the U.S. HRS and selected IPS measure a single, unidimensional latent cognitive construct, and (b) individual cognitive test items relate to the underlying latent cognitive function constructs equivalently across the HRS and IPS. Knowledge of the underlying latent cognitive function structure across these studies will allow future researchers to relate exposures of interest to the specific

Table 1. Included Studies, U.S. Health and Retirement Study and Selected International Partner Studies

Country	Study	Year	*N*
China	China Health and Retirement Longitudinal Study (CHARLS)	2011	16,043
Costa Rica	Costa Rican Longevity and Healthy Aging Study (CRELES)	2006	2,026
England	English Longitudinal Study of Ageing (ELSA)	2002	11,778
South Africa	Heath and Aging in Africa: A Longitudinal Study of an INDEPTH Community in South Africa (HAALSI)	2015	4,927
United States	Health and Retirement Study, Children of the Depression Birth Cohort (HRS CODA)	1998	2,187
Indonesia	Indonesia Family Life Survey (IFLS)	2007	21,603
South Korea	Korean Longitudinal Study of Aging (KLoSA)	2006	10,041
India	Longitudinal Aging Study in India (LASI) Pilot	2010	1,619
Mexico	Mexican Health and Aging Study (MHAS)	2012	5,457
China	Study on Global Ageing and Adult Health (SAGE; China)	2010	14,280
Ghana	Study on Global Ageing and Adult Health (SAGE; Ghana)	2010	5,096
India	Study on Global Ageing and Adult Health (SAGE; India)	2010	11,228
Mexico	Study on Global Ageing and Adult Health (SAGE; Mexico)	2010	2,596
Russian Federation	Study on Global Ageing and Adult Health (SAGE; Russian Federation)	2010	4,323
South Africa	Study on Global Ageing and Adult Health (SAGE; South Africa)	2010	4,050
Europe[a]	Survey of Health, Ageing, and Retirement in Europe (SHARE)	2004	29,932
Ireland	The Irish Longitudinal Study on Ageing (TILDA)	2010	8,504
Total			155,690

[a] Includes Austria, Belgium, the Czech Republic, Denmark, France, Germany, Greece, Ireland, Italy, the Netherlands, Poland, Spain, Sweden, and Switzerland.

cognitive domains represented in these studies and will facilitate the future statistical harmonization of cognitive measures across studies. We aimed to address methodological challenges to cross-national comparisons of cognitive function data in older populations using the best available psychometric methods, with the intention of supporting the future statistical harmonization of cognitive assessments across the HRS IPS.

Method

Study Samples

We used publicly available data from 12 HRS IPS, representing 26 countries (N = 155,690; Table 1).

The HRS was selected as the reference sample. An important consideration prior to statistical harmonization is that cognitive test performance in the full HRS cohort in any recent year may be subject to strong practice effects, as cognitive assessments have been administered biennially to this study population since 1992. To compare test performance in an experienced HRS cohort to test performance in test-naïve samples from other national studies could introduce bias. Importantly for our purpose, the HRS includes "refresher" subsamples added to the lower end of the study age range at most biennial visits to account for the aging structure of the original cohort. The HRS also includes specific birth cohort subsamples added during the 1990s to broaden the total study age range, as the original 1992 sample only included adults born from 1931 to 1941 (aged 51–61 years at enrollment). The use of an HRS subsample at its first study interview would overcome practice effects in cognitive test performance that may be introduced by using the full sample at any non-inaugural interview.

We therefore restricted the HRS reference sample to the first-time interview of a subsample called the "Children of the Depression" (CODA), a birth cohort of n = 2,320 adults born from 1924 to 1930 who were enrolled into the HRS in 1998 (aged 68–74 at study entry) to account for a missing age band of adults in the HRS at that time. Other more recent test-naive "refresher" subsamples were not utilized for this analysis, as their relatively young and narrow age range at study entry (51–56 years) resulted in insufficient age overlap compared with IPS, and a restricted distribution of cognitive scores.

We ensured as best possible that study populations of all IPS were naïve to the cognitive tests included in the study by selecting interview waves that represented the first cognitive assessment occasion for each study population.

Cognitive Measures

The cognitive function items were taken from short screening instruments administered in face-to-face interviews to participants in each study. The cognitive test items included in the HRS IPS were carefully selected by the study investigators to be brief assessments of cognitive functions that are important for day-to-day function and are sensitive to aging-related changes, such as orientation to time and place, attention and calculation, spatial orientation, immediate and delayed episodic memory, fluency, and learning and following instructions. These assessments primarily reflect “fluid” cognitive functions, as opposed to “crystallized” cognitive functions that are relatively stable with age, such as acquired knowledge long-term memories. Individual cognitive test items and their overlap across studies are presented in Supplementary Table 1.

Statistical Approach

We used IRT-based models estimated within a structural equation modeling approach. The assumption that we test with this modeling strategy is that the HRS IPS cognitive batteries represent a single, unidimensional general fluid cognitive function factor (Gross et al., 2014; Jones et al., 2019; McArdle et al., 2009; Strauss & Fritsch, 2004; Wouters et al., 2010). This assumption holds if the observed factor structure is comparable across different studies and if model fit statistics are adequate in each study. Formally, this is called configural invariance (Bontempo & Hofer, 2007), and we tested this separately for single-factor models of all cognitive items and bifactor models with memory-specific and non-memory-specific subfactors. Figure 1 presents the generalized structures of a unidimensional model (Panel A) and a bifactor model (Panel B) for cognitive function. We incorporated memory-specific subfactors because previous research and theory on cognitive aging suggest that episodic memory items strongly covary with one another and are one of the most sensitive cognitive domains to aging-related changes (Salthouse, 2001, 2009). We interpret the estimated factors as representing the covariance across their included individual cognitive items, which may be influenced by an individual’s level of cognitive ability as well as the difficulty and discrimination of the individual test items.

We estimated the standardized factor loadings to indicate the strength of the relationship between each cognitive test item and its underlying cognitive factor, for both single-factor and bifactor models, across all included studies. We estimated an additional memory subfactor for ELSA (England) and two more for The Irish Longitudinal Study on Ageing (TILDA; Ireland), due to their inclusion of additional memory items. The additional memory subfactor in ELSA incorporated two prospective memory items. In TILDA, one of the additional subfactors incorporated two picture memory items, and the other incorporated two trails tasks. We assessed the fit of all models using the comparative fit index (CFI), root mean square error of approximation (RMSEA), standardized root mean squared residual (SRMR), and number of large residuals. The CFI and RMSEA incorporate model complexity, rewarding more parsimonious models. The RMSEA is sample size dependent and may penalize small samples and models with few degrees of freedom (Kenny et al., 2015). The SRMR is an absolute measure of model fit that suggests the value of the mean residual for the model-implied correlation matrix, given the observed correlation matrix. Residuals were calculated as the differences between Fisher’s z-transformed observed and model-implied correlation coefficients. We primarily relied on the SRMR and the number of large residuals to evaluate model fit, supported by the CFI and RMSEA. We used the following approximate thresholds to indicate satisfactory model fit: CFI >0.95, RMSEA <0.06, and SRMR <0.08 (Hu & Bentler, 1999). We considered residuals more than 0.5 to be “large,” according to Cohen’s thresholds for effect sizes (q) (Cohen, 1988). All analyses were conducted using Mplus statistical software (Muthén & Muthén, 2017).

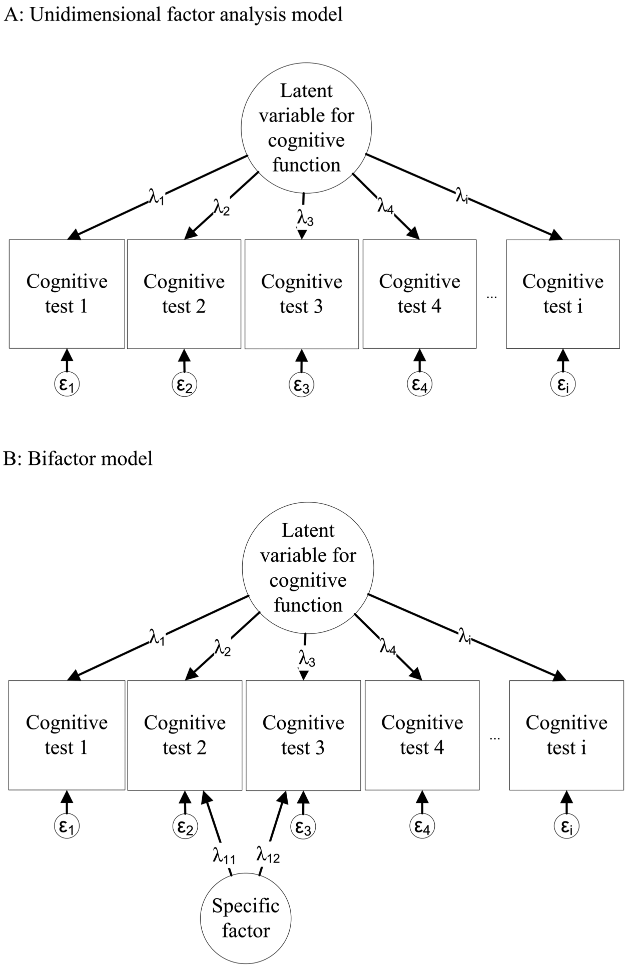

Figure 1. (A) A unidimensional factor analysis model in which common variation across observed cognitive tests is attributable to one latent variable, representing a cognitive function. (B) A bifactor model, in which a specific subfactor is added to account for covariation between Tests 2 and 3 that is greater than predicted from a unidimensional model. Observed test scores are shown in squares, while latent or unobserved variables are shown in circles. Arrows from latent variables to observed indicators represent factor loadings or relationships between the underlying latent variable and its observed indicators.

Cognitive test item handling

Test items were treated as categorical in confirmatory factor analysis models, which is consistent with IRT and uses polychoric correlations among test items. Immediate and delayed recall lists of N words (N is 10 in most studies, but

8 or 3 in others) were treated as categorical variables with values from 0 to 10 (or 0–5). Because Mplus statistical software does not allow categorical variables with more than 10 categories, we recode 10-item word recall lists (which have 11 categories) by collapsing responses of 5 and 6 words recalled together. This reduces measurement precision at the middle of the ability distribution, in order to preserve precision at the tails of the distribution (i.e., the scale minimum and maximum). Animal fluency was coded as the number of mistakes subtracted from the total number of animals named and then recoded in binned categories (0–1, 2–5, 6–9, 10–13, 14–7, 18–21, 22–25, and 26–29).

We estimated the single-factor and bifactor models using a robust weighted least squares (WLSMV) estimator. The WLSMV estimator makes no distributional assumptions about the observed data, allowing estimation with continuous and categorical independent variables. The mean is held at zero and variance at one for the latent variables; otherwise, there are no parameter constraints. The WLSMV estimator assumes that data are missing completely at random (MCAR). Individuals with missing data on all cognitive test items were excluded from this study. Otherwise, item-level missingness in cognitive test items did not preclude inclusion into this study, as missingness was not differential by the level of cognitive functioning or other factors not in the models.

Results

Sample Characteristics Across Studies

The mean (*SD*) ages of the included samples ranged from 38.3 (16.7) years in the Indonesia Family Life Survey (IFLS) to 77.2 (9.7) years in the Costa Rican Longevity and Healthy Aging Study (CRELES; Table 2). Mean (*SD*) age in the HRS CODA sample was 70.4 (3.1) years. Just more than half of the participants were female in each study, which is representative of the older populations of most countries, given longer female than male life expectancies internationally (Table 2).

Cognitive Function Test Item Inclusion Across Studies

Across all studies, the most frequently included test items were the ability to state the current day of the week (10 of the 12 studies), immediate recall of 10 words (9 of the 12 studies), and delayed recall of 10 words (8 of the 12 studies). All studies had at least four items in common with at least one other study, although many studies also included unique items that were not in any other study (Supplementary Table 1).

Single-Factor and Bifactor Model Fits Across Studies

The summary model fit statistics for the single-factor and bifactor model structures within each study are summarized in Table 3. We summarize model fit statistics (CFI, RMSEA, SRMR, number of large residuals) for single-factor general cognitive function models, as well as bifactor models with separate memory-specific and non-memory-specific cognitive domains (Table 3). Detailed tables with standardized factor loadings for each study are given in Supplementary Table 2.

The single-factor models fit the data at acceptable model fit thresholds for 6 of the 12 included HRS IPS, indicating

Table 2. Sample Characteristics, U.S. Health and Retirement Study and Selected International Partner Studies

	Age			
Study	Mean (*SD*)	Interquartile range	Range	Female, *N* (%)
CHARLS	59.0 (10.0)	51–65	22–101	8,536 (53%)
CRELES	77.2 (9.7)	69–84	62–111	1,288 (54%)
ELSA	64.1 (10.9)	55–72	20–90	6,764 (56%)
HAALSI	61.7 (13.1)	52–71	40–111	2,714 (54%)
HRS CODA	70.4 (3.1)	69–72	34–81	1,378 (59%)
IFLS	38.3 (16.7)	24–50	14–97	11,413 (53%)
KLoSA	61.7 (11.1)	52–70	45–105	5,791 (56%)
LASI Pilot	58.5 (11.5)	50–65	22–96	333 (20%)
MHAS	55.1 (6.2)	52–59	21–112	3,071 (56%)
SAGE (China)	60.3 (11.8)	53–69	18–99	7,665 (54%)
SAGE (Ghana)	60.1 (14.1)	52–70	18–114	2,411 (47%)
SAGE (India)	50.0 (16.6)	37–62	18–106	6,881 (61%)
SAGE (Mexico)	63.7 (14.2)	58–74	22–105	1,614 (62%)
SAGE (Russian Federation)	62.3 (13.0)	54–72	18–100	2,783 (64%)
SAGE (South Africa)	60.3 (12.4)	53–68	18–113	2,328 (58%)
SHARE	63.9 (10.6)	55–71	25–103	16,921 (56%)
TILDA	63.0 (9.4)	55–70	49–80	4,724 (56%)

Table 3. Single-Factor and Bifactor Model Fit Statistics, U.S. Health and Retirement Study and Selected International Partner Studies

			Single-factor model fits				Bifactor model fits			
Study	Number of participants	Number of cognitive items	CFI	RMSEA	SRMR	Number of large residuals	CFI	RMSEA	SRMR	Number of large residuals
CHARLS	16,043	9	0.903	0.108	0.100	0	0.976	0.055	0.063	0
CRELES	2,026	8	0.981	0.036	0.040	0	—	—	—	—
ELSA	11,778	15	0.945	0.053	0.094	1	0.971	0.039	0.085	0
HAALSI	4,927	7	0.957	0.183	0.090	0	0.999	0.031	0.013	0
HRS CODA	2,187	12	0.959	0.052	0.108	0	0.992	0.024	0.080	0
IFLS	21,603	4	0.999	0.035	0.024	0	1.000	0.005	0.002	0
KLoSA	10,041	12	0.980	0.047	0.066	1	0.981	0.046	0.065	1
LASI Pilot	1,619	11	0.913	0.143	0.090	1	0.967	0.089	0.068	0
MHAS	5,457	12	0.842	0.114	0.082	1	0.924	0.082	0.058	0
SAGE (Ghana)	5,096	6	0.904	0.239	0.055	0	0.970	0.164	0.035	0
SAGE (China)	14,280	6	0.978	0.152	0.031	0	0.991	0.121	0.020	0
SAGE (Mexico)	2,596	6	0.933	0.218	0.046	0	0.983	0.135	0.021	0
SAGE (Russian Fed.)	4,323	6	0.951	0.272	0.046	0	0.989	0.159	0.029	0
SAGE (South Africa)	4,050	6	0.923	0.229	0.044	0	0.970	0.177	0.028	0
SAGE (India)	11,228	6	0.935	0.208	0.042	0	0.995	0.070	0.012	0
SHARE	29,932	10	0.967	0.064	0.078	0	0.983	0.046	0.060	0
TILDA	8,504	15	0.930	0.076	0.075	0	0.987	0.034	0.054	0

Notes: CFI = confirmatory fit index; RMSEA = root mean square error of approximation; SRMR = standardized root mean squared residual. Bifactor models for ELSA and TILDA include additional memory bifactors. A bifactor model was not estimated for CRELES, due to sparse memory item coverage.

that a single, unidimensional latent general cognitive function construct is measured across all of these studies (Table 3). These studies were CRELES (Costa Rica), IFLS (Indonesia), Korean Longitudinal Study of Aging (South Korea), Survey of Health, Ageing, and Retirement in Europe (Europe), the SAGE family of studies, and TILDA (Ireland). Single-factor models did not fit the data well for the cognitive assessments in the CHARLS (China), ELSA (England), Health and Ageing in Africa: a Longitudinal Study of an INDEPTH Community in South Africa (HAALSI; South Africa), HRS CODA (the United States), Longitudinal Aging Study in India (LASI) Pilot (India), and MHAS (Mexico; Table 3). The bifactor models with separate memory-specific and non-memory-specific subfactors (with additional memory-specific subfactors for ELSA and TILDA) fit the data better than the single-factor models for all of the HRS IPS (Table 3). The model fit statistics for the more complex bifactor models became acceptable for the CHARLS (China), ELSA (England), HAALSI (South Africa), HRS CODA (the United States), LASI Pilot (India), and MHAS (Mexico), such that these models fit the data in all included studies at acceptable fit thresholds (Table 3).

Discussion

We observed that only half of the cognitive batteries included in 12 U.S. HRS IPS reflected a single, unidimensional latent general cognitive function factor. However, bifactor models with separate memory-specific and non-memory-specific domains were of a better fit to the data for all included studies. These results indicate that latent variable approaches to cross-national cognitive harmonization should incorporate memory-specific factors, and that further statistical harmonization of cognitive batteries across the HRS IPS should be possible. Although relatively simple latent variable models fit the data well in this study, common cognitive test item coverage was sparse, and we could not derive factors for specific non-memory cognitive domains, such as processing speed, fluency, or executive function. More comprehensive cognitive batteries, such as the Harmonized Cognitive Assessment Protocol (HCAP), may be needed to harmonize data for multiple specific cognitive domains. Ultimately, this report should help facilitate the statistical harmonization of cognitive aging data from around the world to promote, as best possible, cross-national comparisons of cognitive aging outcomes in a rapidly aging world.

Assumptions and Alternatives

Harmonizing cognitive function data across countries, languages, and cultures involves some intractable heterogeneities in data. Here, we highlight the item coverage and model fit statistics for single-factor and bifactor models of the HRS IPS cognitive function batteries and discuss the necessary assumptions and potential strategies for their statistical harmonization. An important immediate future direction for statistical harmonization research is to investigate differential item functioning (DIF) of the included cognitive batteries. DIF could be introduced by heterogeneity in test and study sample characteristics across studies (Teresi et al., 2000). For example, country-level differences in interpretation of test items, language of test items, use of aids during the study interview, and alternate forms of items (e.g., alternate word lists) could contribute to differences in cognitive test item performance across studies, independently of true differences in underlying cognitive function across older populations.

Within study populations, DIF in common cognitive batteries has been observed according to gender, education, race/ethnicity, and urban–rural residence (Crane et al., 2004; Goel & Gross, 2019; Jones, 2006; Jones & Gallo, 2002). Between study populations in different countries, differences in language and cultural interpretation of cognitive test items across countries could further contribute to DIF across studies. Chan et al. (2015) previously observed DIF between the ELSA (England) and HRS (the United States) cognitive batteries, whereby cognitive test items were more difficult for ELSA than HRS respondents, given equal levels of underlying cognitive performance. Their study used data collected in 2002 from both studies, which was the first year of ELSA but the 10th year of the HRS, so differential practice effects could have biased their results. While outside of the scope of the present analysis, the identification and characterization of the many different possible sources of DIF across the HRS IPS cognitive batteries are an important future research area.

To minimize potential bias due to practice effects in cognitive testing, we selected time points from all studies that represented the first cognitive function assessment for as many sample members as possible. Hence, we restricted the HRS to the CODA sample, as it was the largest and oldest test-naïve HRS sample with the most comprehensive test item coverage. The intention is for the CODA sample to serve as the reference sample in future IRT-based statistical harmonization modeling. Our use of the CODA sample also meant that we did not make use of newer HRS items, such as those assessing numeracy, which is a trade-off of our decision to prioritize minimizing practice effects in cognitive testing. CODA also had a restricted age range (68–74 years) and sample size relative to the full HRS sample.

Alternatively, we could have chosen a different HRS subsample for inclusion, such as (a) a synthetic cohort of first-time HRS enrollees (although most first-time enrollees were aged 51–56 years and had relatively low variability in their cognitive function scores), (b) the HRS/ADAMS sample (although this sample had variable prior cognitive testing exposure), or (c) all HRS data to-date (maximizes sample size and included age range, but with variable prior cognitive testing exposure). We also could have chosen a sample from a different IPS as the reference sample. Future analyses that statistically harmonize the HRS IPS data could conduct sensitivity analyses to quantify the implications of selecting these and other samples as the calibration reference sample.

Limitations

The limitations of this study stem partly from the required assumptions, as discussed in the previous section. Another potential limitation of our approach is that the WLSMV estimator in Mplus assumes that missing data are MCAR. Although this is a potentially unrealistic assumption, the majority of studies were missing less than 5% of observations across the cognitive test items. The only studies missing more than 5% of observations for any cognitive test items were CHARLS (5%–8% missing across test items), SAGE Russian Federation and South Africa (9% and 10% missing for delayed recall, respectively), and ELSA and SHARE (both missing 19%–40% missing on the three numeracy items, but nearly complete on all other items). To improve model fit and reduce any potential bias due to missing data patterns, items with more than 5% missing observations could have been dropped, but at the cost of reducing common item coverage across studies. Alternatively, maximum likelihood estimation with robust standard errors (MLR estimator in Mplus) could be used. This estimator is more computationally intensive than the WLSMV estimator, but it allows for missing cognitive function data to be missing at random (i.e., as a function of the observed data). More importantly, the maximum likelihood estimators do not produce limited-information model fit statistics (e.g., CFI, RMSEA) that serve as a common language for communicating the adequacy of model fit. We believe that for this study, a principally descriptive study and assessment of configural invariance of the cognitive assessment batteries in the HRS and IPS, the limited-information estimators are reasonable. However, future statistical harmonization analyses should use maximum likelihood estimation or other estimator that involves less restrictive assumptions about the missing data mechanism.

Future Directions for Cross-National Harmonization of Cognitive Aging Data

This report presents the prestatistical harmonization methodology and basic model fit and factor structure assessment of the cognitive batteries included in the U.S. HRS and 11 selected IPS. The immediate future direction leading from this work is to empirically evaluate the equivalence of factor loadings across countries (formally, this is known as metric invariance) and evaluate and account for any observed DIF by country or demographic characteristics of study samples in the performance of the cognitive batteries across the HRS IPS (Bontempo & Hofer, 2007; Chan et al., 2015). This next step in cross-national statistical harmonization would allow for the extraction of parameter estimates, such as item discrimination and location on the common latent cognitive function variable, facilitating valid cross-national comparisons of the older population distribution and predictors of cognitive function. There are several important research questions about global cognitive aging that could be answered with harmonized cross-national data on cognitive function. For just one example, harmonized data would allow valid comparisons of the magnitudes of social inequalities in cognitive function, impairment, and dementia across countries. Previous research has investigated the roles of structural factors such as economic development and gender equity indicators in cross-national differences in cognitive outcomes (Bonsang et al., 2017; Skirbekk et al., 2012; Weber et al., 2014), but this important area of inquiry remains limited without statistically harmonized outcome measures.

Conclusions

Across all 12 U.S. HRS IPS, a relatively simple bifactor model with an episodic memory factor and non-memory-specific factor fit reasonably well. Two of these studies included additional memory subfactors to account for additional memory items in their cognitive batteries. Future cross-national studies that harmonize cognitive aging data using latent variable approaches should incorporate memory-specific subfactors, where appropriate. More comprehensive and consistent measurement of multiple cognitive domains across countries may be needed in order to facilitate better cross-national comparisons of the distribution and determinants of cognitive aging outcomes from a global perspective. A promising future direction is the HCAP that has recently been implemented in several U.S. HRS IPS to improve the consistency and quality of cognitive function assessments across diverse country contexts. Future research using the HCAP and other harmonized cognitive measures should carefully consider the assumptions and statistical co-calibrations necessary to conduct valid cross-national comparisons of cognitive aging outcomes.

Supplementary Material

Supplementary data are available at *The Journals of Gerontology, Series B: Psychological Sciences and Social Sciences* online.

Funding

This work was supported by the National Institute of Aging at the National Institutes of Health (R01AG051170 to R. N. Jones; K01AG050699 to A. L. Gross; R56AG05066-01 and P01AG041710-06 to L. F. Berkman; K25AG055620 to S. Mukherjee; L. E. Gibbons' effort was supported by P50AG05136 to Thomas J. Grabowski; L. C. Kobayashi's effort was partially supported by P30AG012846 to Vicki A. Freedman). L. C. Kobayashi was supported by the David E. Bell Fellowship at the Harvard Center for Population and Development Studies and as a Junior Faculty Fellow of the Michigan Center on the Demography of Aging.

Conflict of Interest

None declared.

Data Availability

Data for this study are made publicly available by the U.S. Health and Retirement Study and its International Partner Studies on their individual websites, as well as the Gateway to Global Aging (https://g2aging.org) for most of the studies. The present analysis of these data was not preregistered.

References

Banks, J., Breeze, E., Cheshire, H., Cox, K., Demakakos, P., Emmerson, C., Gardener, E., Gjonça, E., Guralnik, J. M., Hacker, E., Huppert, F. A., Kumari, M., Lang, I., Leicester, A., Lessof, C., Maisey, S., Marmot, M., McWilliams, B., Melzer, D., … Zaninotto, P. (2006). Cognitive function. In J. Banks, E. Breeze, C. Lessof, & J. Nazroo (Eds.), *Retirement, health and relationships of the older population in England: The 2004 English Longitudinal Study of Ageing (Wave 2)* (Issue July, pp. 217–242). The Institute for Fiscal Studies.

Bartels, C., Wegrzyn, M., Wiedl, A., Ackermann, V., & Ehrenreich, H. (2010). Practice effects in healthy adults: A longitudinal study on frequent repetitive cognitive testing. *BMC Neuroscience*, **11**, 118. doi:10.1186/1471-2202-11-118

Bonsang, E., Skirbekk, V., & Staudinger, U. M. (2017). As you sow, so shall you reap: Gender-role attitudes and late-life cognition. *Psychological Science*, **28**(9), 1201–1213. doi:10.1177/0956797617708634

Bontempo, D., & Hofer, S. (2007). Assessing factorial invariance in cross-sectional and longitudinal studies. In A. D. Ong & M. H. M. van Dulmen (Eds.), *Oxford handbook of methods in positive psychology* (1st ed., pp. 153–175). Oxford University Press.

Chan, K. S., Gross, A. L., Pezzin, L. E., Brandt, J., & Kasper, J. D. (2015). Harmonizing measures of cognitive performance across international surveys of aging using item response theory. *Journal of Aging and Health*, **27**(8), 1392–1414. doi:10.1177/0898264315583054

Cohen, J. (1988). Differences in correlation coefficients. In J. Cohen (Ed.), *Statistical power analysis for the behavioral sciences* (2nd ed., pp. 109–144). Lawrence Erlbaum Associates.

Crane, P. K., van Belle, G., & Larson, E. B. (2004). Test bias in a cognitive test: Differential item functioning in the CASI. *Statistics in Medicine*, **23**(2), 241–256. doi:10.1002/sim.1713

Crane, P. K., Narasimhalu, K., Gibbons, L. E., Mungas, D. M., Haneuse, S., Larson, E. B., Kuller, L., Hall, K., & van Belle, G. (2008). Item response theory facilitated cocalibrating cognitive tests and reduced bias in estimated rates of decline. *Journal of Clinical Epidemiology*, **61**(10), 1018–1027.e9. doi:10.1016/j.jclinepi.2007.11.011

Gibbons, R. D., Perraillon, M. C., & Kim, J. B. (2014). Item response theory approaches to harmonization and research synthesis. *Health Services & Outcomes Research Methodology*, **14**(4), 213–231. doi:10.1007/s10742-014-0125-x

Goel, A., & Gross, A. (2019). Differential item functioning in the cognitive screener used in the Longitudinal Aging Study in India. *International Psychogeriatrics*, **31**(9), 1331–1341. doi:10.1017/S1041610218001746

Gómez-Olivé, F. X., Montana, L., Wagner, R. G., Kabudula, C. W., Rohr, J. K., Kahn, K., Bärnighausen, T., Collinson, M., Canning, D., Gaziano, T., Salomon, J. A., Payne, C. F., Wade, A., Tollman, S. M., & Berkman, L. (2018). Cohort profile: Health and Ageing in Africa: A Longitudinal Study of an INDEPTH Community in South Africa (HAALSI). *International Journal of Epidemiology*, **47**(3), 689j–690j. doi:10.1093/ije/dyx247

Gross, A. L., Jones, R. N., Fong, T. G., Tommet, D., & Inouye, S. K. (2014). Calibration and validation of an innovative approach for estimating general cognitive performance. *Neuroepidemiology*, **42**(3), 144–153. doi:10.1159/000357647

Gross, A. L., Power, M. C., Albert, M. S., Deal, J. A., Gottesman, R. F., Griswold, M., Wruck, L. M., Mosley, T. H. Jr, Coresh, J., Sharrett, A. R., & Bandeen-Roche, K. (2015). Application of latent variable methods to the study of cognitive decline when tests change over time. *Epidemiology (Cambridge, Mass.)*, **26**(6), 878–887. doi:10.1097/EDE.0000000000000379

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, **6**(1), 1–55. doi:10.1080/10705519909540118

Jones, R. N. (2006). Identification of measurement differences between English and Spanish language versions of the Mini-Mental State Examination. Detecting differential item functioning using MIMIC modeling. *Medical Care*, **44**(11 Suppl. 3), S124–S133. doi:10.1097/01.mlr.0000245250.50114.0f

Jones, R. N. (2015). Practice and retest effects in longitudinal studies of cognitive functioning. *Alzheimer's & Dementia (Amst)*, **1**(1), 101–102. doi:10.1016/j.dadm.2015.02.002

Jones, R. N., & Gallo, J. J. (2002). Education and sex differences in the mini-mental state examination: Effects of differential item functioning. *The Journals of Gerontology, Series B: Psychological Sciences and Social Sciences*, **57**(6), 548–558. doi:10.1093/geronb/57.6.p548

Jones, R. N., Rudolph, J. L., Inouye, S. K., Yang, F. M., Fong, T. G., Milberg, W. P., Tommet, D., Metzger, E. D., Cupples, L. A., & Marcantonio, E. R. (2019). Development of a unidimensional composite measure of neuropsychological functioning in older cardiac surgery patients with good measurement precision. *Journal of Clinical and Experimental Neuropsychology*, **32**(10), 1041–1049. doi:10.1038/jid.2014.371

Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2015). The performance of RMSEA in models with small degrees of freedom. *Sociological Methods and Research*, **44**(3), 486–507. doi:10.1177/0049124114543236

Kowal, P., Chatterji, S., Naidoo, N., Biritwum, R., Fan, W., Lopez Ridaura, R., Maximova, T., Arokiasamy, P., Phaswana-Mafuya, N., Williams, S., Snodgrass, J. J., Minicuci, N., D'Este, C., Peltzer, K., & Boerma, J. T.; SAGE Collaborators. (2012). Data resource profile: The World Health Organization Study on global AGEing and adult health (SAGE). *International Journal of Epidemiology*, **41**(6), 1639–1649. doi:10.1093/ije/dys210

McArdle, J. J., Grimm, K. J., Hamagami, F., Bowles, R. P., & Meredith, W. (2009). Modeling life-span growth curves of cognition using longitudinal data with multiple samples and changing scales of measurement. *Psychological Methods*, **14**(2), 126–149. doi:10.1037/a0015857

Muthén, L., & Muthén, B. (2017). Examples: Confirmatory factor analysis and structural equation modeling. In L. Muthén & B. Muthén (Eds.), *MPlus user's guide* (pp. 55–112). Muthén & Muthén.

Ofstedal, M. B., Fisher, G., Herzog, A.; HRS Health Working Group. (2005). *Documentation of cognitive functioning measures in the Health and Retirement Study*. Institute for Social Research, University of Michigan. HRS Documentation Report DR-006. http://hrsonline.isr.umich.edu/sitedocs/userg/dr-006.pdf

Prince, M., Wimo, A., Guerchet, M., Ali, G. -C., Wu, Y. -T., Prina, M.; Alzheimer's Disease International. (2015). *World Alzheimer Report 2015: The Global Impact of Dementia: An analysis of prevalence, incidence, cost, and trends*. Alzheimer's Disease International. https://www.alzint.org/u/WorldAlzheimerReport2015.pdf

Salthouse, T. A. (2001). Structural models of the relations between age and measures of cognitive functioning. *Intelligence*, **29**(2), 93–115. doi:10.1016/S0160-2896(00)00040-4

Salthouse, T. A. (2009). When does age-related cognitive decline begin? *Neurobiology of Aging*, **30**(4), 507–514. doi:10.1016/j.neurobiolaging.2008.09.023

Skirbekk, V., Loichinger, E., & Weber, D. (2012). Variation in cognitive functioning as a refined approach to comparing aging across countries. *Proceedings of the National Academy of Sciences of the United States of America*, **109**(3), 770–774. doi:10.1073/pnas.1112173109

Sonnega, A., Faul, J. D., Ofstedal, M. B., Langa, K. M., Phillips, J. W., & Weir, D. R. (2014). Cohort profile: The Health and Retirement Study (HRS). *International Journal of Epidemiology*, **43**(2), 576–585. doi:10.1093/ije/dyu067

Steptoe, A., Breeze, E., Banks, J., & Nazroo, J. (2013). Cohort profile: The English Longitudinal Study of Ageing. *International Journal of Epidemiology*, **42**(6), 1640–1648. doi:10.1093/ije/dys168

Strauss, M. E., & Fritsch, T. (2004). Factor structure of the CERAD neuropsychological battery. *Journal of the International Neuropsychological Society*, **10**(4), 559–565. doi:10.1017/S1355617704104098

Teresi, J. A., Kleinman, M., & Ocepek-Welikson, K. (2000). Modern psychometric methods for detection of differential item functioning: Application to cognitive assessment measures. *Statistics in Medicine*, **19**(11–12), 1651–1683. doi:10.1002/(sici)1097-0258(20000615/30)19:11/12<1651::aid-sim453>3.0.co;2-h

Tollman, S. M., Norris, S. A., & Berkman, L. F. (2016). Commentary: The value of life course epidemiology in low- and middle-income countries: An ageing perspective. *International Journal of Epidemiology*, **45**(4), 997–999. doi:10.1093/ije/dyw109

Vivot, A., Power, M. C., Glymour, M. M., Mayeda, E. R., Benitez, A., Spiro, A. 3rd, Manly, J. J., Proust-Lima, C., Dufouil, C., & Gross, A. L. (2016). Jump, hop, or skip: Modeling practice effects in studies of determinants of cognitive change in older adults. *American Journal of Epidemiology*, **183**(4), 302–314. doi:10.1093/aje/kwv212

Weber, D., Skirbekk, V., Freund, I., & Herlitz, A. (2014). The changing face of cognitive gender differences in Europe. *Proceedings of the National Academy of Sciences of the United States of America*, **111**(32), 11673–11678. doi:10.1073/pnas.1319538111

Wouters, H., van Gool, W. A., Schmand, B., Zwinderman, A. H., & Lindeboom, R. (2010). Three sides of the same coin: Measuring global cognitive impairment with the MMSE, ADAS-cog and CAMCOG. *International Journal of Geriatric Psychiatry*, **25**(8), 770–779. doi:10.1002/gps.2402

Zhao, Y., Hu, Y., Smith, J. P., Strauss, J., & Yang, G. (2014). Cohort profile: The China Health and Retirement Longitudinal Study (CHARLS). *International Journal of Epidemiology*, **43**(1), 61–68. doi:10.1093/ije/dys203