Received: 16 November 2022 | Revised: 29 March 2023 | Accepted: 10 April 2023

DOI: 10.1002/dad2.12438

Alzheimer's & Dementia
Diagnosis, Assessment & Disease Monitoring

RESEARCH ARTICLE

Harmonization of cognitive screening tools for dementia across diverse samples: A simulation study

Brandon E. Gavett[1] | **Sindana D. Ilango**[2] | **Rebecca Koscik**[3,4,5] | **Yue Ma**[4,6] | **Benjamin Helfand**[7,8] | **Chloe W. Eng**[9] | **Alden Gross**[10] | **Emily H. Trittschuh**[11,12] | **Richard N. Jones**[8,13] | **Dan Mungas**[14]

[1]School of Psychological Science, University of Western Australia, Perth, Western Australia, Australia

[2]Department of Epidemiology, University of Washington School of Public Health, Seattle, Washington, USA

[3]Wisconsin Alzheimer's Institute, University of Wisconsin School of Medicine and Public Health, Madison, Wisconsin, USA

[4]Wisconsin Alzheimer's Disease Research Center, University of Wisconsin School of Medicine and Public Health, Madison, Wisconsin, USA

[5]Department of Medicine, University of Wisconsin School of Medicine and Public Health, Madison, Wisconsin, USA

[6]Division of Geriatrics and Gerontology, Department of Medicine, University of Wisconsin School of Medicine and Public Health, Madison, Wisconsin, USA

[7]Department of Emergency Medicine, University of Massachusetts Medical School, Worcester, Massachusetts, USA

[8]Departments of Psychiatry and Human Behavior and Neurology, Warren Alpert Medical School, Brown University, Providence, Rhode Island, USA

[9]Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, California, USA

[10]Department of Epidemiology, Johns Hopkins Bloomberg School Public Health, Baltimore, Maryland, USA

[11]VA Puget Sound Health Care System, Geriatric Research Education and Clinical Care, Seattle, Washington, USA

[12]Department of Psychiatry and Behavioral Sciences, University of Washington School of Medicine, Seattle, Washington, USA

[13]Department of Neurology, Brown University Warren Alpert Medical School, Providence, Rhode Island, USA

[14]Department of Neurology, University of California, Sacramento, California, USA

Correspondence
Brandon E. Gavett, School of Psychological Science, University of Western Australia, 35 Stirling Highway (M304), Perth, WA 6009, Australia.
Email: brandon.gavett@uwa.edu.au

Funding information
National Institute on Aging, Grant/Award Numbers: R13 AG030995, R01 AG027161, R01 AG051170; National Institute of Environmental Health Sciences, Grant/Award Number: T32 ES015459

Abstract

Introduction: Research focusing on cognitive aging and dementia is a global endeavor. However, cross-national differences in cognition are embedded in other sociocultural differences, precluding direct comparisons of test scores. Such comparisons can be facilitated by co-calibration using item response theory (IRT). The goal of this study was to explore, using simulation, the necessary conditions for accurate harmonization of cognitive data.

Method: Neuropsychological test scores from the US Health and Retirement Study (HRS) and the Mexican Health and Aging Study (MHAS) were subjected to IRT analysis to estimate item parameters and sample means and standard deviations. These estimates were used to generate simulated item response patterns under 10 scenarios that adjusted the quality and quantity of linking items used in harmonization. IRT-derived factor scores were compared to the known population values to assess bias, efficiency, accuracy, and reliability of the harmonized data.

Results: The current configuration of HRS and MHAS data was not suitable for harmonization, as poor linking item quality led to large bias in both cohorts. Scenarios with more numerous and higher quality linking items led to less biased and more accurate harmonization.

Discussion: Linking items must possess low measurement error across the range of latent ability for co-calibration to be successful.

KEYWORDS
bias, cognition, computer simulation, global health, item response theory

HIGHLIGHTS

- We developed a statistical simulation platform to evaluate the degree to which cross-sample harmonization accuracy varies as a function of the quality and quantity of linking items.
- Two large studies of aging—one in Mexico and one in the United States—use three common items to measure cognition.
- These three common items have weak correspondence with the ability being measured and are all low in difficulty.
- Harmonized scores derived from the three common linking items will provide biased and inaccurate estimates of cognitive ability.
- Harmonization accuracy is greatest when linking items vary in difficulty and are strongly related to the ability being measured.

1 | BACKGROUND

Neuropsychological tests are used to estimate abilities in cognitive domains such as memory, language, visuospatial abilities, and executive function. In longitudinal research on aging and dementia, these tests can be utilized to monitor cognitive status over time with aims of identifying risk factors for neurodegenerative outcomes, such as cognitive decline and Alzheimer's disease. There are a multitude of these types of studies worldwide (see, e.g., https://iadrp.nia.nih.gov/, https://hrs.isr.umich.edu/about/international-sister-studies, https://g2aging.org/). Although such studies tend to have similar goals, they also tend to rely upon assessment methods that are culturally and linguistically familiar to the population of interest. For example, the Health and Retirement Study (HRS) and the Mexican Health and Aging Study (MHAS) are two sister longitudinal studies of aging in the United States and Mexico, respectively, with the overarching goal to identify determinants of healthy aging.[1,2] Although the cognitive tests used in these two studies are similar, they are not identical. Differences in cognitive assessment methods pose challenges for making comparisons of group means across studies, as differences in test items can either obscure true differences in cognitive ability or cause the illusory appearance of group differences.[3]

Combining information from multiple studies with harmonization methods can allow researchers to evaluate questions related to cross-country/cultural differences and expand research to questions that require more statistical power and increased generalizability (e.g., genome wide association studies). Harmonization methods can be applied to compare and combine studies that are not directly comparable (e.g., differences in tests administered, procedures, and study populations). Also referred to as co-calibration, harmonization encompasses a broad set of methods used to combine instruments administered with different procedures or in different populations to yield one or more summary factor scores that can be used in subsequent analyses that pool data across studies.[3–5] One statistical approach to data harmonization uses item response theory (IRT). IRT is an application of latent variable models that summarizes individual responses in relation to item and test characteristics and can be a valuable tool for pooling psychometric data.[6] For example, when two batteries share one or more linking items – identical items present in both batteries – IRT can be used to estimate traits that are assessed by both batteries. The resulting factor score(s) are said to be harmonized across the batteries. Although methods of harmonization using IRT co-calibration are well-documented and growing in usage,[7–9] there is still a paucity of research evidence to guide them, such as the quantity and quality of the linking items shared by both studies required for adequate harmonization.

The aim of this study is to leverage simulation methods to better understand how best to harmonize cognitive tests across studies using IRT-based methods. To date, the published literature on simulation of

cognitive harmonization is scant.[10] In this study, we apply simulation methods to address this knowledge gap, and specifically, to evaluate the accuracy of IRT harmonization methods with cognitive data from HRS and MHAS, by manipulating the quantity and quality of linking items. Although this study utilizes the HRS and MHAS samples, our main objective – to understand the conditions under which cognitive aging data can be successfully harmonized – is not restricted to these cohorts but is intended to be applicable to cognitive aging research more broadly. We expect that these findings can contribute to understanding how test item overlap relates to harmonization quality, which can inform cross-national and other cross-group research and improve the design of future cognitive test batteries.

RESEARCH IN CONTEXT

1. **Systematic Review**: A literature review was performed using PubMed and Scopus; this was augmented by examining the reference lists from relevant published articles. Few cognitive aging studies have used statistical methods to equate test scores across cross-national samples or other diverse groups.
2. **Interpretation**: Successful harmonization of neuropsychological outcome data across groups requires common (linking) items to be present in both samples. The current results show that a small number of linking items may be adequate for harmonization, but only if they are very high in discrimination and able to capture a broad range of the ability being measured.
3. **Future Directions**: The methods described here can be used to evaluate the degree to which future efforts to harmonize cognitive data across diverse groups may be biased or unreliable, or to design prospective collection of research outcomes that are amenable to harmonization.

2 | METHODS

2.1 | Participants

The population parameters used to simulate data for this study were derived from two large adult cohorts. The University of Michigan's HRS (https://hrs.isr.umich.edu/) is a longitudinal study of aging, which consists of a representative United States sample of adults aged 50 years and older and their spouses. The initial cohort was selected in 1992 (and spouses invited) with subsequent studies recruiting new subjects in waves to cover the continuum of aging over time and by generation.[1,11–13] The MHAS (http://www.mhasweb.org/) began in 2001 and is a longitudinal study of those age 50+ with emphasis placed on recruitment of subjects from urban and rural Mexico.[14–17] All participants in HRS and MHAS provided informed consent to participate in their respective studies.

To obtain realistic values for the population parameters used to simulate data, we examined archival data from 19,311 HRS participants and 26,018 MHAS participants (total $N = 45{,}329$). Pre-statistical harmonization[18] of these data was performed as part of the Psychometric Integrative Technology for Cognitive Health Research (PITCH) study.

2.2 | Materials

Cognitive assessment in HRS and MHAS was performed using instruments taken from or modeled after the Telephone Interview of Cognitive Status[19] and Cross-Cultural Cognitive Examination.[20] Test scores in both HRS and MHAS include orientation to the day, month, and year (all dichotomous). We refer to these as the natural linking items. Test scores exclusive to HRS include orientation to date, president, and vice president (all dichotomous); backward counting (dichotomous); verbal naming (two dichotomous items); serial subtraction (0–5 scale); and 10-item immediate and delayed list recall tasks (recoded to a 0–9 scale). Test scores exclusive to MHAS include eight-item word list learning and delayed recall tasks (0–8 scale); visual scanning (recoded to a 0–9 scale), and figure copy and recall tasks (0–2 scale).

2.3 | Procedure

This study was performed using statistical simulation. All data processing and analysis were performed in R version 4.1.2.[21] All harmonization and data simulations were performed using the *mirt* package version 1.34.[22] The code used to simulate and analyze these data can be found at https://github.com/dmungas/harmonization-simulation. Our analyses proceeded in five steps:

Step 1: Selecting population parameters

In step 1, we applied a unidimensional graded response model[23] to the combined HRS + MHAS sample data to estimate the threshold and discrimination parameters for each of the cognitive items. Resulting factor scores from this model were used to estimate the means (M_{HRS}, M_{MHAS}) and standard deviations (SD_{HRS}, SD_{MHAS}) of global cognition in the two samples. The resulting sample means and SD were used to set the population parameters in the simulation studies, as described below.

Step 2. Generating harmonization scenarios

We generated 10 harmonization scenarios that manipulated the linking items shared by the two groups and patterns of available data for each group. These scenarios, which are summarized in Table 1, were designed to represent multiple combinations of linking item quantity (the number of linking items) and quality (the item difficulty, represented by threshold parameters) to achieve a comprehensive understanding of how these variations influence the adequacy of IRT-based harmonization.

For Scenarios 1–8, we used the actual item threshold and discrimination values that were estimated in step 1. For Scenario 9, we manually changed the thresholds to values that were balanced

TABLE 1 Ten simulation scenarios manipulating linking items and non-linking items.

Condition number	Items																	Condition description	
	Orientation to day	Orientation to month	Orientation to year	Orientation to date	President	Vice-President	Backwards counting	Verbal naming—*scissors*	Verbal naming—*cactus*	Serial 7s	Immediate 10-word recall	Delayed 10-word recall	8-word recall, 3 trials[d]	Delayed 8-word recall	Visual scanning	Figure copy	Figure recall		
1	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	1.	All items are linking items.
2	L	L	L	G1	G1	G1	G1	G1	G1	G1	G1	G1	G2	G2	G2	G2	G2	2.	Three natural linking items: *conservative linking scenario.*
3	L	L	L	G1	G1	G1	G1	G1	G1	G1	L	G1	G2	G2	G2	G2	G2	3.	Four linking items: *liberal linking scenario.*
4	L	L	L	L	L	L	L	L	L	L	L	L	–	–	–	–	–	4.	G1 (HRS) items only (all 12 HRS items are linking items).
5	–	–	–	–	–	–	–	–	–	–	–	–	L	L	L	L	L	5.	G2 (MHAS) items only (all 5 MHAS items are linking items).
6	L	G1	G1	G1	G1	G1	G1	G1	G1	G1	G1	G1	G2	G2	G2	G2	G2	6.	Only one (dichotomous) linking item of low difficulty.
7	G1	G1	G1	G1	G1	G1	G1	G1	G1	G1	L	G1	G2	G2	G2	G2	G2	7.	Only one (polytomous) linking item, with thresholds of a range of difficulty.
8	G2	G2	G2	G1	G1	G1	G1	G1	G1	G1	L	G1	G2	G2	G2	G2	G2	8.	Only one (polytomous) linking item, with thresholds of a range of difficulty, natural linking items not simulated in group 1.
9	L[a]	L[b]	L	G1	G1	G1	G1	G1	G1	G1	G1	G1	G2	G2	G2	G2	G2	9.	Three natural linking items with balanced thresholds.
10	L[ac]	L[bc]	L	G1	G1	G1	G1	G1	G1	G1	G1	G1	G2	G2	G2	G2	G2	10.	Three natural linking items with balanced thresholds and boosted discrimination.

Abbreviations: G1, item data simulated in Group 1 only; G2, item data simulated in Group 2 only; HRS, Health and Retirement Study; L, linking item (item data simulated in both groups); MHAS, Mexican Health and Aging Study; – item data not simulated in either group; SD, standard deviation.

[a]Threshold parameter changed from sample-estimated value (−2.343) to 0.

[b]Threshold parameter changed from sample-estimated value (−2.724) to 2.180.

[c]Discrimination parameter changed from sample-estimated value to 4.0.

[d]Each learning trial treated as a separate item.

across the ability spectrum (i.e., item thresholds were changed from sample-estimated values of -2.180, -2.433, and -2.724 to -2.180, 0, and +2.180). For Scenario 10, we made the same alterations to the threshold parameters and also boosted the linking item discrimination parameters to values of 4.0 (very high discrimination). These changes were made to systematically evaluate the differences between balanced thresholds/high discrimination (Scenario 10), balanced thresholds/low discrimination (Scenario 9), and imbalanced thresholds/low discrimination (Scenario 2).

To summarize, Scenario 1 represents the optimal harmonization condition, where all items are available as linking items. Scenario 2 represents the true status of the empirical data, with three linking items shared between HRS and MHAS. Scenarios 3–8 explore the effects of changing the number of linking items, from as few as one (Scenarios 6–8) to as many as twelve (Scenario 4). Scenarios 9 and 10 seek to determine whether – under conditions similar to Scenario 2 (three natural linking items) – the limitations of having a small number of linking items can be mitigated by more desirable psychometric properties of those linking items.

Step 3. Simulating item responses

In step 3, we used the IRT parameter estimates derived in step 1 to simulate item responses in Groups 1 and 2. The specific data simulated was dependent upon each scenario's item availability (Table 1). In all scenarios, we simulated 500 data sets using the *simdata* function in the *mirt* package. The sample size for both groups was set to $n = 500$ (total $N = 1000$) and no missing data was generated. We treated Group 1 as the reference for comparison purposes. Population-level SD $\sigma\theta_{G1}$ and $\sigma\theta_{G2}$ were set to the sample-estimated values of SD_{HRS} and SD_{MHAS}, respectively. The *simdata* function works as follows: for a given simulation, an individual respondent's true ability level θ_i was randomly sampled from a normal distribution with the specified population parameters $\theta_{iG1}\sim N(\mu_{\theta G1},\sigma^2{}_{\theta G2})$ and $\theta_{iG2}\sim N(\mu\theta_{G2},\sigma^2{}_{\theta G2})$. These θ_i values were then used to randomly generate item response patterns given the item threshold and discrimination parameters obtained in step 1.

Step 4. Harmonizing simulated data

In step 4, we harmonized the simulated Group 1 and Group 2 data with one another. We began by fitting a graded response model to the Group 1 simulated item responses to obtain parameter estimates for each item (step 4a). We then re-ran the graded response model in the combined sample (Group 1 + Group 2) with constraints on the linking item parameters fixed to the estimates derived from Group 1; the mean and standard deviation of the latent variable were freely estimated (step 4b). Finally, in step 4c, the resulting discrimination and threshold parameters were fixed to their population estimates derived in step 4b, while the metric of the latent trait was fixed to the sample-estimated mean and SD in the combined sample. Factor scores were then estimated using the expected a posteriori method.[24] These harmonized factor scores provide an estimate of a respondent's true ability level (θ_i) and are the primary focus of our analysis.

Step 5. Evaluating harmonized factor scores

In step 5, we evaluated the extent to which the population means $\mu_{\theta G1}$, $\mu_{\theta G2}$, $\mu_{\theta CG}$ (the mean population θ of the Combined Group) and individual θ_i values were recovered by the harmonized factor scores in each scenario. Our criterion measures include bias, empirical standard error (ESE), root mean square error (RMSE), and the correlation coefficient (r). Bias provides a measure of the degree to which an estimated sample mean is consistently above or below the population mean; lower absolute values are better. The ESE is a measure of the variability in the mean harmonized factor scores; lower ESE values indicate greater efficiency in recovering the population parameters. RMSE is a measure of the accuracy with which the simulated respondents' harmonized factor scores reproduced the individual data-generating θ_i values; lower RMSE values indicate greater accuracy, with values < 0.3 preferred.[25] Finally, r is a measure of the strength of the linear association between the true data-generating θ_i values and the harmonized ability estimates, χ_i; values > 0.90 are preferred.

3 | RESULTS

The parameter estimates derived from the combined PITCH data set are shown in the electronic supplement. Most of the dichotomous items, including the three natural linking items, were of low difficulty. Of these natural linking items, discrimination was lowest – but acceptable – for Orientation to Day, and highest – and quite good – for Orientation to Year. The polytomous memory items were highly discriminatory and covered the range of ability well, as can be seen from their threshold parameter estimates (see electronic supplement). Estimated factor scores based on this IRT model were HRS, M = 0.011 (SD = 0.920); MHAS, M = -0.319 (SD = 0.996).

The simulation results, comparing the 10 harmonization scenarios on the chosen outcome measures, are shown in Tables 2–4. Table 2 shows the results in the combined sample, whereas Tables 3 and 4 show the results for Groups 1 and 2, respectively. Figure 1 depicts the distributions of the sample factor score means, derived from all 500 simulations, separated by scenario and group.

When interpreting these results, it is useful to consider Scenario 1, where all items were available as linking items, as representing the most optimal conditions for co-calibration. This scenario led to excellent harmonization, as evidenced by low bias ($\leq$0.004 in all groups), high efficiency (ESE $\leq$ 0.063 in all groups), high accuracy (RMSE $\leq$ 0.257 in all groups), and high reliability ($r \geq 0.966$ in all groups).

Poor harmonization was achieved in Scenarios 2 (three natural linking items: conservative linking scenario), 6 (only one [dichotomous] linking item of low difficulty), and 9 (three natural linking items with balanced thresholds). These three scenarios produced factor scores that were more biased, less precise, less accurate, and less reliable than the other scenarios. However, it should be noted that, even in these poorly performing scenarios, the correlations (r) between the

TABLE 2 Simulation results for the combined group (population parameters: $\mu_{\theta CG} = -0.165$, $\sigma_{\theta CG} = 0.971$).

Scenario	M	SD	Bias (%)	ESE	RMSE	*r*
1	−0.163	1.013	0.002 (1.1%)	0.032	0.255	0.969
2	−0.178	0.976	−0.013 (−7.8%)	0.090	0.380	0.932
3	−0.161	0.989	0.004 (2.2%)	0.038	0.326	0.946
4	−0.161	0.979	0.004 (2.4%)	0.034	0.355	0.935
5	−0.171	1.003	−0.005 (−3.3%)	0.034	0.318	0.950
6	−0.165	0.979	0.000 (−0.1%)	0.116	0.408	0.927
7	−0.155	0.987	0.010 (6.1%)	0.037	0.328	0.945
8	−0.166	0.990	−0.001 (−0.8%)	0.040	0.331	0.945
9	−0.161	0.986	0.004 (2.7%)	0.056	0.363	0.935
10	−0.163	0.999	0.003 (1.6%)	0.036	0.314	0.951

Abbreviations: ESE, empirical standard error; G1, item data simulated in Group 1 only; G2, item data simulated in Group 2 only; RMSE, root mean square error; SD, standard deviation.

TABLE 3 Quality of harmonization for recovering the Group 1 population mean (population parameters: $\mu_{\theta G1} = -0.011$, $\sigma_{\theta G1} = 0.920$).

Scenario	M_{G1}	SD_{G1}	Bias	ESE	RMSE	*r*
1	−0.011	0.972	−0.000	0.002	0.257	0.966
2	−0.133	0.947	−0.122	0.083	0.395	0.928
3	−0.036	0.947	−0.025	0.010	0.356	0.930
4	−0.019	0.941	−0.008	0.005	0.356	0.929
5	−0.017	0.957	−0.006	0.003	0.319	0.944
6	−0.148	0.966	−0.137	0.109	0.418	0.928
7	−0.034	0.948	−0.023	0.009	0.356	0.930
8	−0.044	0.950	−0.033	0.013	0.364	0.927
9	−0.104	0.970	−0.093	0.037	0.374	0.931
10	−0.042	0.971	−0.031	0.010	0.323	0.946

Note: Bias reflects the difference between M_{G1} and $\mu_{\theta G1}$.
Abbreviations: ESE, empirical standard error; RMSE, root mean square error; SD, standard deviation.

TABLE 4 Quality of harmonization for recovering the Group 2 population mean (population parameters: $\mu_{\theta G2} = -0.319$, $\sigma_{\theta G2} = -0.996$).

Scenario	M_{G2}	SD_{G2}	Bias	ESE	RMSE	*r*
1	−0.315	1.029	0.004	0.063	0.254	0.970
2	−0.223	1.002	0.096	0.099	0.363	0.947
3	−0.287	1.014	0.033	0.067	0.293	0.960
4	−0.303	0.995	0.016	0.065	0.354	0.938
5	−0.324	1.023	-0.005	0.065	0.317	0.953
6	−0.182	0.992	0.137	0.123	0.393	0.945
7	−0.277	1.009	0.043	0.066	0.297	0.958
8	−0.288	1.013	0.031	0.069	0.293	0.960
9	−0.217	1.000	0.102	0.075	0.350	0.947
10	−0.283	1.012	0.036	0.063	0.305	0.957

Note: Bias reflects the quantity of the difference between M_{G2} and $\mu_{\theta G2}$.
Abbreviations. ESE, empirical standard error; RMSE, root mean square error; SD, standard deviation.

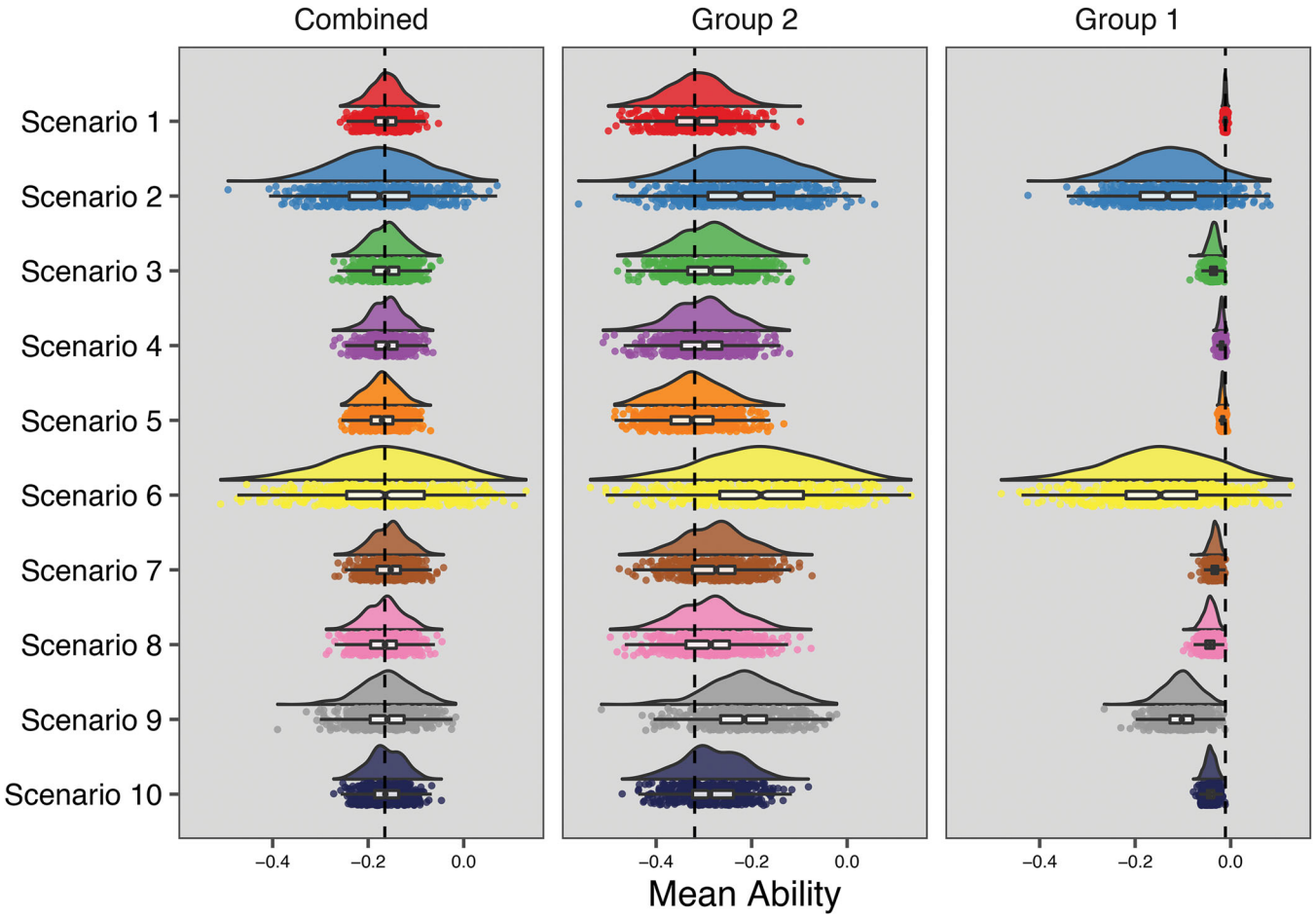

FIGURE 1 Empirical sampling distributions of sample factor score means derived from 500 simulations of each harmonization scenario. Each point represents one sample mean. Boxplots show median, approximate 95% confidence intervals for the median (as notches in the boxplot), interquartile range (hinges), and 1.5 times the interquartile range (whiskers). Dashed vertical lines represent the population means ($\mu_{\theta CG} = -0.165$, $\mu_{\theta C2} = -0.319$, $\mu_{\theta C1} = -0.011$). See Table 1 for more details about each scenario.

individual factor scores and the data-generating latent trait values were within an acceptable range. The poor performance of Scenario 2 is noteworthy, as it represents the actual HRS and MHAS data availability, with three dichotomous orientation items available as linking items.

Aside from Scenarios 2, 6, and 9, the remaining scenarios were more successful at approaching the harmonization success of Scenario 1. Whereas these other scenarios had moderate success at making unbiased and efficient estimates of group means, they were less effective for accurately estimating individual ability levels, as can be seen in the higher than desirable RMSE values. On the other hand, the high correlations show that the individual factor scores were strongly linearly related to the underlying trait. Overall, none of Scenarios 2–10 were able to match the outcomes produced by Scenario 1. However, six of these nine made reasonably unbiased and efficient sample mean estimates, and reliable, if not highly accurate, individual trait estimates.

We converted the linking items' sample-estimated discrimination and threshold parameters to item Information, and, subsequently, to standard errors (Figure 2). In IRT, Information is a statistic that is mathematically and conceptually related to the concept of reliability, and inversely related to measurement error.[24] Thus, the standard errors shown in Figure 2 can be interpreted as the expected magnitude of error when estimating the underlying trait (θ). Standard errors of < 0.30 are desirable, as that threshold is often used as a precision-based stopping rule in computerized adaptive testing.[25] When comparing the standard errors in Figure 2 to the data in Tables 2–4, the more successful harmonization scenarios were associated with more precisely estimated (and analogously, more reliable) factor scores.

4 | DISCUSSION

There are many benefits to combining cognitive data from multiple studies with similar, but not identical, batteries of tests. "Big data" are needed for statistical reasons to answer many important questions that address nature versus nurture questions. Collecting these data in a single study is unrealistic, likely flawed (one size does not fit all when conducting research across cultures and languages, etc.), and perhaps impossible. Rather, finding ways to leverage cognitive data in longitudinal studies of aging that already exist by harmonizing (or co-calibrating)

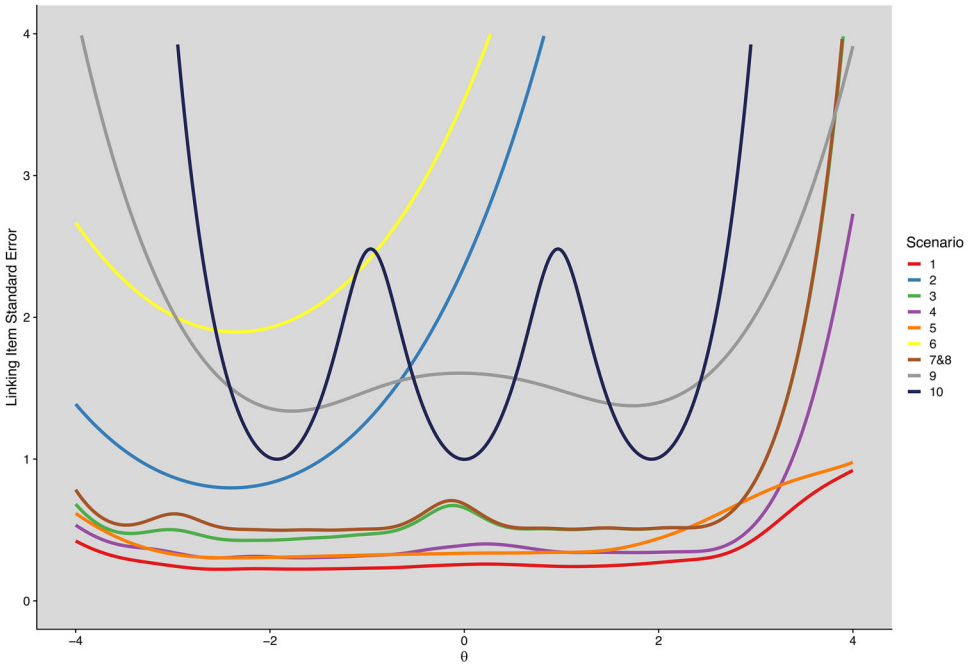

FIGURE 2 Standard errors for estimating cognitive ability derived from linking items' difficulty and discrimination parameters when estimated in the combined sample. Standard errors differ as a function of Scenario (colored lines) and the level of the underlying trait (θ; x-axis). See Table 1 for more details about each scenario. Scenarios 7 and 8 use the same linking items and therefore can be summarized together.

can be of great benefit.[3] What we have not seen in the literature thus far is a systematic method for considering the psychometric strengths of various linking items, both their quantity and quality, to determine a minimum set of requirements versus an optimal scenario. These are necessary pieces of information both for (1) considering whether it is appropriate to combine/harmonize data from certain studies as well as (2) for proactive design of studies so that they might be harmonized in the future. We postulated that simulation based on realistic parameters, in this case, the HRS and MHAS sister longitudinal studies of health determinants in aging, can address these important issues and questions. The results of this simulation demonstrated that successful cross-national equating of measures of cognition is quite difficult when few linking items are available and provide guidance for characteristics of linking items that are required for successful harmonization. Notably, ensuring there are linking items that cover most of the latent ability continuum is important; however, a more important feature is having highly discriminating linking items, that is, items that measure the intended construct well.

We examined many plausible scenarios reflecting different hypothetical and actual representations of HRS and MHAS linking item availability. When all items are linking items, (Scenarios 1, 4, and 5), relatively unbiased cross-group comparison of means is possible. In the scenario reflecting the naturally existing linking items that were fielded in the HRS and MHAS (Scenario 2), group mean comparisons were badly biased, especially in Group 1. Including additional low-quality linking items (Scenario 2 vs. Scenario 6) did not resolve the bias, nor did ensuring that the linking items spanned the full ability range (Scenario 2 vs. Scenario 9). Successful harmonization under the current pattern of data availability (three dichotomous linking items) can be possible when the threshold parameters spanned the full ability range and when these items' discrimination parameters are very high (Scenario 9 vs. Scenario 10). Reasonably successful harmonization also appears possible when there is at least one single high-quality polytomous linking item (Scenarios 7 and 8).

The literature contains numerous examples of research focused on deriving harmonized or composite scores from different cognitive tests administered to the same sample.[26–28] However, harmonizing cognitive data across samples – especially when the samples differ on important dimensions such as language and culture – is an equally important, yet less common research goal.[29] Comparative, including cross-national, studies of cognitive functioning are extremely challenging to plan.[30] Cognition is a construct that is defined within the broad environmental context of the individual.[31,32] The attempt to bring together cognitive data collected using similar methods in different settings can use data harmonization to align measurements. Our results suggest that co-calibration, using an IRT approach, requires linking items with good discrimination and threshold parameters that span the range of cognitive ability in the population. In other words, the linking items, when combined, should possess high Information and, consequently, low standard errors (Figure 2). Linking items effectively set the common metric of the harmonized test scores, and our results showed that the three orientation items shared by HRS and MHAS all were at very low difficulty levels and that the factor scores that resulted from using these linking items would be positively biased for MHAS and negatively biased for HRS. This means that true mean differences between these groups would be obscured. In real-world settings, systematic differences and biases in measurement are unknown, unlike this simulation study where true ability was known. Without more

numerous or higher-discrimination linking items spanning the ability level, unbiased comparisons between US and Mexican samples using the cognitive tests administered in these two studies is not possible.

Despite these impediments to making cross-group comparisons, the correlation between individuals' harmonized factor scores and individuals' latent ability was high; this pattern was evident even in Scenarios 2 and 6, which showed the most bias. This finding suggests that rank-ordering individuals within groups on the basis of their ability level can be done, even if, on average, those individual ability estimates contain higher than desirable measurement error. Such an approach would only require estimating an IRT model for each sample, not harmonizing data across samples, as long as no cross-sample comparisons were made or interpreted as reflecting true differences in the latent trait. This approach would support parallel analyses in the two samples to determine if the effects of risk and protective factors on cognitive outcomes are present in the two groups. Of course, conclusions about how risk and protective factors might exert similar or different influences on cognition across groups depends on the assumption that the cognitive tests are measuring the same latent trait in both groups. An assumption underlying the IRT approach to equating is that the linking items have the same psychometric characteristics in the two samples. Exploring issues related to measurement invariance and/or differential item functioning was beyond the scope of the current study, but future research should evaluate how deviations from this assumption affect cross-study harmonization accuracy.

A further limitation of this study is that the available items provided very coarse measures of cognition. The original HRS included cognitive measures that could be easily administered over the telephone in about 5 min. The MHAS was designed following the example of the HRS. More recent developments in the HRS and MHAS studies have included a more extensive battery of cognitive performance tests. The Harmonized Cognitive Assessment Protocol[33,34] might provide a more appropriate context for making cross-national comparisons if the fielded versions of these tests include plausible linking items that have high discrimination and span the range of underlying ability.

In conclusion, our results provide guidance for designing cognitive assessment instruments and choosing which instruments to include in research studies. Even just a few linking items could be adequate for harmonization, but only if they are very high in discrimination and able to capture a wide range of abilities. Further, the number and quality of the items in the two studies is important to ensure reliability is matched, such that differences across the two studies are not simply due differences in error variance of the cognitive outcome measure. In the absence of these conditions, we must find alternative methods to link scale scores or alternative data sources to conduct cross-national comparisons.[5] Our results also demonstrate the utility of simulation studies as a framework for assessing the validity of claims made from cross-national comparisons where the construct being compared is harmonized using modern measurement techniques. These results suggest that statistical simulation can be useful for planning cross-sample co-calibration efforts to ensure that the available linking items allow for unbiased harmonization of the desired outcomes.

ACKNOWLEDGMENTS

This work was funded by grants from the National Institute on Aging (R13 AG030995: Mungas, PI; R01 AG027161: Johnson, PI) and the National Institute of Environmental Health Sciences (T32 ES015459: Ilango, PI). The funding sources had no role in the study design, execution, or preparation of this article.

CONFLICT OF INTEREST STATEMENT

The authors have no conflicts of interest to disclose. Author disclosures are available in the supporting information.

REFERENCES

1. Juster FT, Suzman R. An overview of the Health and Retirement Study. *J Hum Resour*. 1995;30:S7-S56. doi: 10.2307/146277
2. Wong R, Michaels-Obregon A, Palloni A. Cohort profile: the Mexican Health and Aging Study (MHAS). *Int J Epidemiol*. 2017;46(2):e2. doi: 10.1093/ije/dyu263
3. Kobayashi LC, Gross AL, Gibbons LE, et al. You say tomato, I say radish: can brief cognitive assessments in the U.S. Health and Retirement Study be harmonized with its international partner studies? *J Gerontol B Psychol Sci Soc Sci*. 2021;76(9):1767-1776. doi: 10.1093/geronb/gbaa205
4. Chan KS, Gross AL, Pezzin LE, Brandt J, Kasper JD. Harmonizing measures of cognitive performance across international surveys of aging using item response theory. *J Aging Health*. 2015;27(8):1392-1414. doi: 10.1177/0898264315583054
5. Gross AL, Kueider-Paisley AM, Sullivan C, Schretlen D, International Neuropsychological Normative Database Initiative. Comparison of approaches for equating different versions of the mini-mental state examination administered in 22 studies. *Am J Epidemiol*. 2019;188(12):2202-2212. doi: 10.1093/aje/kwz228
6. Mellenbergh GJ. Generalized linear item response theory. *Psychol Bull*. 1994;115(2):300-307. doi: 10.1037/0033-2909.115.2.300
7. Dorans NJ. Linking scores from multiple health outcome instruments. *Qual Life Res*. 2007;16(Suppl 1):85-94. doi: 10.1007/s11136-006-9155-3
8. Griffith L, van den Heuvel E, Fortier I, et al. *Harmonization of Cognitive Measures in Individual Participant Data and Aggregate Data Meta-Analysis*. Agency for Healthcare Research and Quality (US); 2013.
9. Jones RN, Fonda SJ. Use of an IRT-based latent variable model to link different forms of the CES-D from the Health and Retirement Study. *Soc Psychiatry Psychiatr Epidemiol*. 2004;39(10):828-835. doi: 10.1007/s00127-004-0815-8
10. Gross AL, Mungas DM, Crane PK, et al. Effects of education and race on cognitive decline: an integrative study of generalizability versus study-specific results. *Psychol Aging*. 2015;30(4):863-880. doi: 10.1037/pag0000032
11. Fisher GG, Ryan LH. Overview of the Health and Retirement Study and introduction to the special issue. *Work Aging Retire*. 2018;4(1):1-9. doi: 10.1093/workar/wax032
12. Crimmins EM, Kim JK, Langa KM, Weir DR. Assessment of cognition using surveys and neuropsychological assessment: the Health and Retirement Study and the aging, demographics, and memory study. *J Gerontol B Psychol Sci Soc Sci*. 2011;66 Suppl 1(Suppl 1):i162-i171. doi: 10.1093/geronb/gbr048
13. Sonnega A, Faul JD, Ofstedal MB, Langa KM, Phillips JW, Weir DR. Cohort profile: the Health and Retirement Study (HRS). *Int J Epidemiol*. 2014;43(2):576-585. doi: 10.1093/ije/dyu067
14. Mejía-Arango S, Wong R, Michaels-Obregón A. Normative and standardized data for cognitive measures in the Mexican Health and Aging Study. *Salud Publica Mex*. 2015;57 Suppl 1(01):S90-S96. doi: 10.21149/spm.v57s1.7594

15. Mejía-Arango S, Gutierrez LM. Prevalence and incidence rates of dementia and cognitive impairment no dementia in the Mexican population: data from the Mexican Health and Aging Study. *J Aging Health*. 2011;23(7):1050-1074. doi: 10.1177/0898264311421199
16. Wong R, Michaels-Obregón A, Palloni A, et al. Progression of aging in Mexico: the Mexican Health and Aging Study (MHAS) 2012. *Salud Publica Mex*. 2015;57 Suppl 1(01):S79-S89. doi: 10.21149/spm.v57s1.7593
17. Wong R, Michaels-Obregon A, Palloni A. Cohort Profile: the Mexican Health and Aging Study (MHAS). *Int J Epidemiol*. 2017;46(2):e2. doi: 10.1093/ije/dyu263
18. Briceño EM, Gross AL, Giordani BJ, et al. Pre-statistical considerations for harmonization of cognitive instruments: harmonization of ARIC, CARDIA, CHS, FHS, MESA, and NOMAS. *J Alzheimer's Dis*. 2021;83(4):1803-1813. doi: 10.3233/JAD-210459
19. Brandt J, Spencer M, Folstein M. The telephone interview for cognitive status. *Neuropsychiatry Neuropsychol Behav Neurol*. 1998;1:111-117.
20. Glosser G, Wolfe N, Albert ML, et al. Cross-cultural cognitive examination: validation of a dementia screening instrument for neuroepidemiological research. *J Am Geriatr Soc*. 1993;41(9):931-939. doi: 10.1111/j.1532-5415.1993.tb06758.x
21. R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing; 2020. https://www.R-project.org/
22. Chalmers RP. mirt: a multidimensional item response theory package for the R environment. *J. Stat. Softw*. 2012;48(6):1-29. doi: 10.18637/jss.v048.i06
23. Samejima F. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*. 1969;34:100-114.
24. Embretson SE, Reise SP. *Item Response Theory for Psychologists*. Erlbaum; 2000.
25. Stafford RE, Runyon CR, Casabianca JM, Dodd BG. Comparing computer adaptive testing stopping rules under the generalized partial-credit model. *Behav. Res. Methods*. 2019;51(3):1305-1320. doi: 10.3758/s13428-018-1068-x
26. Balsis S, Benge JF, Lowe DA, Geraci L, Doody RS. How do scores on the ADAS-Cog, MMSE, and CDR-SOB correspond? *Clin Neuropsychol*. 2015;29(7):1002-1009. doi: 10.1080/13854046.2015.1119312
27. Crane PK, Narasimhalu K, Gibbons LE, et al. Item response theory facilitated cocalibrating cognitive tests and reduced bias in estimated rates of decline. *J Clin Epidemiol*. 2008;61(10):1018-1027.e9. doi: 10.1016/j.jclinepi.2007.11.011
28. Gross AL, Jones RN, Fong TG, Tommet D, Inouye SK. Calibration and validation of an innovative approach for estimating general cognitive performance. *Neuroepidemiology*. 2014;42(3):144-153. doi: 10.1159/000357647
29. Mukherjee S, Choi SE, Lee ML, et al. Cognitive domain harmonization and cocalibration in studies of older adults. *Neuropsychology*. 2022. doi: 10.1037/neu0000835
30. Menon RN, Varghese F, Paplikar A, et al. Validation of Indian council of medical research neurocognitive tool box in diagnosis of mild cognitive impairment in India: lessons from a harmonization process in a linguistically diverse society. *Dement Geriatr Cogn Disord*. 2020;49(4):355-364. doi: 10.1159/000512393
31. Franzen S, van den Berg E, Goudsmit M, et al. A systematic review of neuropsychological tests for the assessment of dementia in non-western, low-educated or illiterate populations. *J Int Neuropsychol Soc*. 2020;26(3):331-351. doi: 10.1017/S1355617719000894
32. Krch D, Lequerica A, Arango-Lasprilla JC, Rogers HL, DeLuca J, Chiaravalloti ND. The multidimensional influence of acculturation on digit symbol-coding and wisconsin card sorting test in hispanics. *Clin Neuropsychol*. 2015;29(5):624-638. doi: 10.1080/13854046.2015.1063696
33. Gross AL, Tommet D, D'Aquila M, et al; BASIL Study Group. Harmonization of delirium severity instruments: a comparison of the DRS-R-98, MDAS, and CAM-S using item response theory. *BMC Med Res Method*. 2018;18(1):92. doi: 10.1186/s12874-018-0552-4
34. Langa KM, Ryan LH, McCammon RJ, et al. The Health and Retirement Study harmonized cognitive assessment protocol project: study design and methods. *Neuroepidemiology*. 2020;54(1):64-74. doi: 10.1159/000503004

SUPPORTING INFORMATION

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

How to cite this article: Gavett BE, Ilango SD, Koscik R, et al. Harmonization of cognitive screening tools for dementia across diverse samples: A simulation study. *Alzheimer's Dement*. 2023;15:e12438. https://doi.org/10.1002/dad2.12438