THE LANCET
Healthy Longevity

Supplementary appendix

This appendix formed part of the original submission and has been peer reviewed. We post it as supplied by the authors.

Supplement to: Gross AL, Li C, Briceño EM, et al. Harmonisation of later-life cognitive function across national contexts: results from the Harmonized Cognitive Assessment Protocols. *Lancet Healthy Longev* 2023; **4:** e573–83.

Supplementary material

Contents

Supplementary Methods: Prestatistical harmonization

The original HCAP test battery[1] in the US included tests of memory (CERAD word list-learning[2], story recall from the Logical Memory Test and the East Boston Memory Test, three object recall), executive functioning (Letter Cancellation, Symbol Digit Modality Test,[3,4] HRS Number Series, Trail Making, Raven's Standard Progressive Matrices[5]), language (semantic fluency for animals, object naming)[6], orientation to time and place, and visuospatial function (CERAD Constructional Praxis)[7].

For linguistic, cultural, literacy, numeracy, and other reasons, each participating HCAP study adapted the neuropsychological battery for their study population. Adaptations to the HCAP battery included translations for test items and instructions, modifying administration and scoring procedures, changing test stimuli, and selecting alternative tests as necessary. Prestatistical review of the cognitive test items is necessary to determine that items are measuring cognitive constructs equivalently across linguistically, educationally, and culturally diverse populations. Determination of equivalent measures are necessary so that differences in the measures can be attributed to real differences between individuals and groups, instead of measurement-related differences. We seek to minimize between-study differences attributable to differences in test execution that might affect test score interpretations differently by study. Seemingly benign distinctions in test scoring, coding, or administration may have sizable effects on neuropsychological test interpretations, mean scores, and correlations with risk factors and outcomes.

Previous studies have described prestatistical harmonization procedures focusing on case examples between HCAPs in the US, Mexico, and India.[8-10] The present study leveraged and expanded on these previous efforts. To facilitate the decision-making required in prestatistical harmonization, we convened a multidisciplinary panel of professionals with cultural/linguistic expertise that included neuropsychologists, epidemiologists, psychometricians, and local study team members with cultural/linguistic expertise in the target populations with experience in administration of HCAPs (ALG, MAR, EMB, LCK). Most had working knowledge of cross-cultural neuropsychology. The team included two cultural neuropsychologists with expertise in Hispanic/Latinx populations (MAR, EMB). Additionally, the team included co-investigators on the HRS-HCAP (LCK), Mex-Cog (MAR), and LASI-DAD (ALG) studies, a former participant interviewer and database manager from the CHARLS-HCAP study (see acknowledgements), and a cognitive epidemiologist who assisted with training LASI-DAD field interviewers in India (ALG). From the HAALSI-HCAP, our team included an epidemiologist with fieldwork experience in South Africa's Agincourt sub-district and who participated in quality control efforts for the HAALSI study (LCK). The ELSA-HCAP's operations director joined meetings when we reviewed ELSA-HCAP (see acknowledgements), and other relevant experts were consulted while we conducted our prestatistical harmonization efforts for this study (see acknowledgements).

This group used available materials including codebooks, interviewer training manuals, and personal communication with study investigators and coordinators to document communalities and differences in test item content and administration across HCAPs. We determined which differences were substantial based on (1) reviews of procedural details involving test administration and scoring, (2) guidelines for coding of responses, and (3) when available, reviews of test item translations to examine construct equivalence of the item from a neuropsychological perspective. The latter entailed evaluation of cultural and linguistic equivalence. Cultural equivalence was ascertained by accounting for cultural familiarity of items. Linguistic equivalence consisted of studying translations of both test items themselves (e.g., are words from a word list learning task of similar imagery value, lexical frequency in the language, and syllabic count?) as well as instructions (e.g., are translated instructions comparably clear or complex in ways intended for the test?).

To facilitate detection of differential item functioning (DIF) among cross-HCAP linking items, described in the Methods, using the HRS-HCAP as the reference and based on available information, we rated each item from other HCAPs as a *confident linking item* that has no known issues with item equivalence, a *tentative linking item* that has possible issues with item equivalence based on reviews of differences described above, or a non-linking item based on available information.[8,9] As described in the Methods, we conducted DIF testing first among confident linking items, followed by testing among tentative linking items (where items confirmed as confident linking items were treated as anchors). One would hypothesize minimal evidence of DIF in items rated as confident linking items. But, importantly, DIF is still tested among confident linking items. Thus, any misclassification of linking items as confident when they should be tentative, or vice versa, will be rectified if such misclassification is not widespread. For this reason, we erred towards assigning items as tentative linking items, but balanced that decision against the number of available linking items in each domain.

The following sections summarize, by domain, rationales for assignment of cognitive test items as confident, tentative, or non-linking items. Supplementary Table 1 summarizes item classification.

Prestatistical harmonization of orientation items

For the most part, cognitive test items assessing orientation to time and place were considered confident items, given their ease of translation at face value. There were exceptions: season of year was considered a tentative linking item in ELSA-HCAP, CHARLS-HCAP, and HAALSI-HCAP due to concerns regarding recognizability of seasons in and across these countries, some lack of clarity in documentation around how correct answers were coded (e.g., is Christmas accepted as a season, or are only weather patterns considered as correct responses?), and the local importance of tracking time in seasons (Supplementary Table 1).

Orientation to year is an item our team did anticipate concerns with because local study representatives did not flag them as major points of contention given the ease with which it can be translated, administered, and scored. Formal DIF testing suggested, however, that after controlling for underlying orientation ability as well as age and gender, performance on this item was much poorer in LASI-DAD and HAALSI-HCAP than in HRS-HCAP. Upon further scrutiny, we recognized that while this type of question may be translated, administered, and coded the same way across HCAPs, the degree to which it is a measure of cognitive orientation can vary. There are potential rationales for why asking older adults certain questions like orientation to year in contexts like India or South Africa are more difficult than in other contexts. Not all cultures rely on a Gregorian calendar system. Even if they do, there are rural areas in many countries in which calendar year may not be of large importance in the daily lives of many older adults.

Prestatistical harmonization of memory items

The primary memory tests in the original HCAP battery were CERAD word list learning and two story recall tests: Logical Memory and East Boston Memory Test. Other tests included delayed constructional praxis and three-word recall. During prestatistical harmonization, we determined the CERAD word recall test was administered comparably between HRS-HCAP, ELSA-HCAP, and CHARLS-HCAP but in a different way from LASI-DAD, Mex-Cog, and HAALSI-HCAP. Participants in the former three countries were presented with the words both verbally and visually, but in the latter three countries, participants were presented with the words only verbally. Moreover, while all studies presented the words verbally, there was variation in the order in which words were presented (i.e., alternating per trial vs fixed). In HRS-HCAP, the list of words is read in a different order in each of the three trials, while in Mex-Cog, LASI-DAD, and others, the intertrial order of the word list is fixed. Because of these distinctions, we considered the CERAD word recall test a linking item between HRS-HCAP, ELSA-HCAP, and CHARLS-HCAP, as well as between LASI-DAD, Mex-Cog, and HAALSI-HCAP, but as a non-linking item between these two sets. These were confident linking items except for CHARLS-HCAP: although administration procedures were similar to HRS-HCAP and ELSA-HCAP, there were questions around the familiarity of certain words in the list. For example, *butter* and *beach* are uncommon Mandarin terms for many older adults in China.

Regarding story recall tests, we considered these as linking items across all HCAPs, albeit usually as tentative due to questions around coding, scoring, and translations. The link between HRS-HCAP and LASI-DAD's story recall items was considered confident because despite having 13 translations into different languages of the HCAP battery in India, the LASI-DAD battery and materials were rigorously back-translated and pilot tested. A previous study investigated measurement invariance of the HCAP battery by language in India, and found little evidence of language differences.[11] Scoring across HCAPs was conducted in a way that allowed "exact" and "gist" coding of responses. In exact or precise scoring, respondents are expected to recollect exact details of each story element. In gist scoring, points are awarded for recalling a story element's main idea.[8] in this study, we relied on gist or approximate scoring as this approach provided greater variability in observed scores in HAALSI-HCAP, LASI-DAD, CHARLS-HCAP, and Mex-Cog. In all studies for which instruments were translated into different languages, there was uncertainty about comparability due to various idiosyncrasies around translation. Some story elements might have varying difficulty in different languages. Place names that were adapted for the local population may have varying familiarity to local respondents, as compared to the original English version. Further discussion of this is available in Briceño et al.[8]

The link between story recall in HRS-HCAP and ELSA-HCAP was tentative because coding instructions in the latter study appeared to follow stricter criteria than the Wechsler Memory Scale-Fourth Edition (WMS-IV) criteria[3] used in HRS-HCAP, Mex-Cog, and others: interviewers were instructed to "only code if respondent mentions underlined words/phrases." For example, recalling a minor variation of the name of the individual in the story, which is a story element, would be incorrect according to interviewer instructions in ELSA-HCAP.

For story recall in HAALSI-HCAP, as in all HCAPs, character names and places in the stories were changed to make it more relatable to the South African population. For the Logical Memory test, the maximum number of points was 24 in HAALSI-HCAP while 25 points were possible in all other HCAPs because of one

missing story element. Instead of "Anna Thompson of South Boston," where "south" is considered a story element, HAALSI-HCAP's adaptation was, "Anna Khosa of Johannesburg."

Prestatistical harmonization of executive functioning items

Tests considered in the executive functioning domain, which includes tasks spanning problem-solving, set-shifting, attentional control, and processing speed, were very similar between HRS-HCAP and ELSA-HCAP. Several of these tests, such as Number Series and Letter Cancellation, have been commonly administered in studies in the US and England for decades. No language translation was necessary, and scrutiny of test administration, coding, and scoring revealed no concerns. There was extensive discussion around whether ELSA-HCAP's Letter Cancellation task was done on A4-sized paper (210 x 297 mm, common in the UK) instead of letter-sized paper (215.9 x 279.4 mm, common in the US) and whether this meant more rows of letters were provided or font sizes differed. We were relieved to learn ELSA-HCAP used letter-sized pages for that test.

Due to numeracy and literacy differences, most tests in the executive functioning domain were infeasible in other HCAPs. Raven's progressive matrices could be administered in LASI-DAD and HAALSI-HCAP, while serial 7s was administered everywhere outside of HRS-HCAP. Otherwise, this domain proved a hotbed of innovation across HCAPs as each study used different instruments. As evidenced by common tests in this domain among Mex-Cog, LASI-DAD, and HAALSI-HCAP, investigative teams in these studies shared recommended practices. Some of these tests included Go-No-Go, Similarities and Differences, backwards day naming (to replace serial 7s and spelling a word backwards), and a Symbol Cancellation Test (to replace the Symbol Digit Modalities test administered in HRS-HCAP and ELSA-HCAP). In general, investigators shared administration and coding instructions for these instruments, for which translations of stimuli and instructions were straightforward compared to items from the memory domain.

Prestatistical harmonization of language items

Language tests across HCAP studies included semantic fluency for animals, a variety of naming tasks (parts of the body, common objects, confrontational naming, etc.), and following commands (point to 2 things in the vicinity; read and follow a written command).

We considered animal fluency a confident linking item across all HCAPs. There were no strong indications of differences in administration, scoring, or coding of this item. Regarding use in different language groups and cultures, previous studies have found no evidence of sizable measurement differences in animal fluency between English and Spanish[12] and English and Arabic.[13] Despite our final confidence rating, we did discuss this item at length; in no existing HCAP are imaginary animals (e.g., unicorns, mermaids) acceptable responses. There was extensive discussion about what this test is intended to represent. Generally, tests of semantic fluency assess the ability to generate words from a specific semantic category, such as animals, vegetables, or fruits, within a certain time limit (in HCAPs, 60 seconds). Semantic fluency tests measure one's language abilities to access previously acquired semantic knowledge and generative fluency related to the cognitive constructs of semantic memory and executive functions.[14] Although animal types and names are not necessarily an overlearned feature in all cultures and countries starting from childhood,[15,16] animals are thought to be a universal category familiar to most individuals across different cultures and languages.

Object naming tests are popular measures of receptive and expressive language abilities. Assessing an individual's ability to retrieve or to produce object names depends on the stimulus used and the salience of that stimulus with one's lived experiences in a cultural setting. The original HCAP battery included questions about receptive and expressive language abilities by requiring individuals to name an item given a specific prompt.[1] For example, a cactus (*What do you call the kind of prickly plant that grows in the desert?*) or scissors (*What do people usually use to cut paper?*) Cognitive test items assessing receptive language ability for which the stimuli were changed were considered non-linking items for purposes of cross-national harmonization. For example, since cacti are not native to many parts of the world, this item has been substituted with varying degrees of success in other countries; in India, the LASI-DAD study asked participants "*What is a brown nut that contains milk?*" (coconut). The cactus item was also administered in CHARLS-HCAP, and while we initially considered it a tentative linking item, evaluation of DIF revealed significant differences in that item between HRS-HCAP and CHARLS-HCAP. We subsequently recognized that 61% of the CHARLS-HCAP sample responded "Don't know" to this item, which is more than 1.5 times the number of respondents who correctly or incorrectly answered it. This finding suggests cacti are not well-recognized plants in China, and thus this is a non-linking item between HRS-HCAP and CHARLS-HCAP.

To assess confrontational naming, participants were shown common objects and asked to name them. In most HCAPs, participants were shown the physical objects, whereas in CHARLS-HCAP participants were shown a 2-dimensional color picture of the objects. Thus, we considered these as tentative linking items between CHARLS-HCAP and other studies.

For object naming, in addition to considerations of the actual stimulus provided (e.g., cacti vs coconuts), scoring rules matter. For example, *what do people usually use to cut paper*? The most correct answer is scissors, but pen knives for opening letters are common as well. Because of differences in scoring rules, we considered this a tentative linking item in Mex-Cog vs other studies but a confident linking item across all other HCAPs. As mentioned earlier in this section, we were not greatly concerned with misclassification of items as confident when they should be tentative, as long as we were certain such misclassification was limited.

A common test of language ability is to ask a participant to read a command and do what it says (e.g., *Close your Eyes*). As the goal of the test is not to disadvantage illiterate people, an adaptation in several HCAP studies (LASI-DAD, CHARLS-HCAP) was made to, "Do what I [interviewer] am doing." The interviewer then demonstrates the task. Such an adaptation would be considered a non-linking item because the original task requires a respondent to successfully read a command, interpret words into actions, and then do the written command. If a person does not read the stimulus, a crucial step of translating words into action is lost. Following evaluation of DIF, we found the variable in CHARLS-HCAP did not make a distinction between whether people read the instruction or imitated the interviewer; thus, thus item had to be considered unique to CHARLS-HCAP and different from the instruction to read (available in most HCAPs) as well as the imitated-only instruction among illiterate participants in LASI-DAD.

References for Supplementary Materials on Prestatistical Harmonization

1. 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: 64–74.
2. Strauss ME, Fritsch T. Factor structure of the CERAD neuropsychological battery, *J Int Neuropsychol Soc* 2004; 10: 559-65.
3. Wechsler D. 2009. *Wechsler Memory Scale - WMS-IV Technical and Interpretive Manual - Fourth Edition* (Pearson: San Antonio, TX).
4. Lowery N, Ragland JD, Gur RC, Gur RE, Moberg PJ. Normative data for the symbol cancellation test in young healthy adults. *Appl Neuropsychol* 2004; 11: 218-21.
5. Raven J. The Ravens progressive matrices: change and stability over culture and time, *Cogn Psychol* 2000; 41: 1-48.
6. Henry JD, Crawford JR, Phillips LH. Verbal fluency performance in dementia of the Alzheimers type: a meta-analysis, *Neuropsychologia* 2004; 42: 1212-22.
7. Yuspeh RL, Vanderploeg RD, Kershaw DAJ. CERAD Praxis memory and recognition in relation to other measures of memory, *Clinical Neuropsychologist* 1998; 12: 468-74.
8. Briceño EM, Arce Rentería M, Gross AL, et al. A cultural neuropsychological approach to harmonization of cognitive data across culturally and linguistically diverse older adult populations, *Neuropsychology* 2023; 37: 247-57.
9. Arce Rentería M, Briceño EM, Chen D, et al. Memory and language cognitive data harmonization between the United States and Mexico. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* 2023, in press.
10. Vonk JMJ, Gross AL, Zammit AR, et al. Cross-national harmonization of cognitive measures across HRS HCAP (USA) and LASI-DAD (India). *PLoS One* 2022;17(2).
11. Gross AL, Khobragade PY, Meijer E, Saxton JA. Measurement and Structure of Cognition in the Longitudinal Aging Study in India–Diagnostic Assessment of Dementia. *J Am Geriatr Soc* 2020;68(S3).
12. Siedlecki KL, Manly JJ, Brickman AM, Schupf N, Tang MX, Stern Y. Do neuropsychological tests have the same meaning in Spanish speakers as they do in English speakers? *Neuropsychology* 2010;24(3):402-411.
13. Zahodne LB, Brauer S, Tarraf W, Morris EP, Antonucci TC, Ajrouch KJ. Measurement and structural invariance of a neuropsychological battery among Middle Eastern/North African, Black, and White Older Adults. *Neuropsychology* 2023; in press.
14. Lezak MD. 2004. *Neuropsychological assessment* (Oxford University Press: Oxford; New York).
15. Scarry S. 1999. *Richard Scarry's Best Word Book Ever* (Random House Childrens Books).
16. Carle E. 2019. *Eric Carle's Book of Many Things* (World of Eric Carle).

Supplementary Table 1. Assignments of cognitive test items based on prestatistical and statistical procedures: Results from HCAP (N=21,144)

Variable	HRS-HCAP	ELSA-HCAP	LASI-DAD	Mex-Cog	CHARLS-HCAP	HAALSI-HCAP
Orientation						
Day of month	Administered	Confident	Confident	Confident	Confident	Confident
Month	Administered	Confident	Confident	Confident	Confident	Confident
Year	Administered	Confident		Confident	Confident	
Year (DIF adjusted)			DIF-corrected			DIF-corrected
Day of the week	Administered	Confident	Confident	Confident	Confident	Confident
What time is it				Novel		Novel
Where are we				Unique		
What country are we in		Novel		Novel		Novel
What state are we in	Administered		Confident	Confident	Confident	Confident
What county are we in	Administered	Tentative			Confident	
What city are we in	Administered	Confident	Confident		Confident	Confident
Season of year	Administered	Tentative	Confident		Tentative	Tentative
Floor of building	Administered		Tentative		Confident	Tentative
Address (street name and/or building number)	Administered	Tentative	Tentative		Confident	
Name of hospital or district/municipality			Novel			Novel
Memory						
CERAD immediate sum of 3 trials	Administered	Confident			Tentative	
CERAD immediate sum of 3 trials			Novel	Novel		Novel
CERAD word list delay	Administered	Confident			Tentative	
CERAD word list delay			Novel	Novel		Novel
CERAD recognition	Administered	Confident			Tentative	
CERAD recognition			Novel	Novel		Novel
Three word immediate registration	Administered	Confident	Confident	Tentative	Confident	Confident
Three word delayed recall	Administered	Confident	Confident	Tentative	Confident	Confident
Logical Memory immediate	Administered	Tentative	Confident	Tentative		Tentative
Logical Memory delay	Administered	Tentative	Confident	Tentative		Tentative
Logical memory recognition	Administered	Tentative	Confident			Tentative
Brave man immediate (East Boston Memory Test)	Administered	Tentative	Confident	Tentative		
Brave man delay (East Boston Memory Test)	Administered	Tentative	Confident	Tentative		
CERAD constructional praxis delay	Administered	Confident	Tentative	Confident		Tentative
Executive functioning						
Problem solving			Unique			
Ravens progressive matrices	Administered	Confident	Tentative			Tentative
HRS Number series	Administered	Confident				
Number series					Unique	
Trails A time (letters and numbers)	Administered	Confident				
Trails B time (letters and numbers)	Administered	Confident				
Similarities			Novel	Novel		Novel
Token Test			Novel			Novel
Digit Span Forward (single item)			Unique			
Digit Span Backward (single item)			Unique			
Digit Span Forward (multiple items)						Unique
Digit Span Backward (multiple items)						Unique
Go-No-Go			Novel	Novel		Novel
Motor Programming						Unique

MMSE Spelling backwards	Administered					
Backward counting, 100-0	Administered	Confident				
Backward counting, 20-0				Unique		
Symbol Digit Modalities Test *	Administered	Confident				
Symbols and Digits test **				Unique		
Symbol Cancellation Test			Novel	Novel		Novel
Letter cancellation	Administered	Confident				
Serial 3s				Unique		
Serial 7s		Novel	Novel	Novel	Novel	Novel
Backward Day naming			Novel			Novel
Forward day naming						Unique
CDR calculation-cent						Unique
Language						
Animal fluency	Administered	Confident	Confident	Confident	Confident	Confident
Name a described cactus	Administered	Confident				
Name a described cactus (DIF adjusted)					Unique	
Name a described coconut			Unique			
What are scissors used for?	Administered	Confident	Confident	Tentative	Confident	Confident
Object naming (watch)	Administered		Confident	Tentative	Tentative	Confident
Object naming (pencil)	Administered		Confident	Confident	Tentative	Confident
Name the elbow	Administered	Confident	Confident	Confident	Confident	Confident
Write a sentence (or write one's name)	Administered	Tentative	Confident	Tentative	Confident	Tentative
Say a sentence						
Read and follow command (Close your eyes)	Administered	Confident	Confident	Confident		Confident
Read and follow command (DIF adjusted)					Unique	
Follow example (close your eyes)					Unique	
Repetition of a phrase	Administered	Confident	Tentative	Tentative	Confident	Confident
What does one do with a hammer	Administered	Confident	Tentative	Tentative	Confident	Tentative
Define Bridge				Unique		
Point to 2 things in the vicinity	Administered	Confident	Confident	Confident	Confident	Confident
Where is the local market?	Administered	Tentative	Tentative	Tentative	Tentative	Tentative
Follow 3-stage instruction	Administered	Confident	Confident	Confident		Confident
Name president or Prime Minister	Administered	Confident	Confident		Confident	Confident
Name deputy president						Unique
Phonemic Fluency						Unique
Boston Naming Test, uncued						Unique

Legend. Confident and tentative linking items are as described in the Methods. Unique items are those administered in a single study. Novel items are administered in more than one HCAP but not in HRS-HCAP; DIF testing was not conducted for these items but most were judged to be confident linking items in reference to Mex-Cog. Unique and novel items are non-linking items.

Supplementary Methods: HCAP-specific factor analyses

Methods

To illustrate empirically that similar organizations of cognitive test items fit well across countries prior to imposing assumptions about cross-national linking items, we estimated confirmatory factor analysis (CFA) models for cognitive domains of general and domain-specific cognitive function separately for each HCAP study. The CFA models estimated two relevant parameters for each cognitive test item: factor loadings, and item thresholds (for categorical items) or intercepts (for continuous items). Factor loadings characterize how strongly correlated a cognitive test item is with the other items in the model. In general, loadings between 0·3 and 0·9 indicate an item is meaningfully related to other items without overwhelming other items in the model.[1,2] Item thresholds characterize the location along the factor at which the cognitive test item provides maximal information regarding underlying cognitive function. We ascertained model fit for CFAs using three standard fit statistics: the Root Mean Square Error of Approximation (RMSEA), Comparative Fit Index (CFI), and Standardized Root Mean Residual (SRMR).[3] When possible, we improved model fit through the use of bifactor models to address additional correlations between similar items (e.g., immediate and delayed recall).[1,2] From these fit statistics, we summarized model fit using the following guidelines that have been described in prior research.[4] Model fit was considered perfect if CFI = 1 and RMSEA = 0 and SRMR = 0, good if CFI ≥ 0·95 and RMSEA ≤ 0·05 and SRMR ≤ 0·05, adequate if CFI ≥ 0·90 and RMSEA ≤ 0·08 and SRMR ≤ 0·08, and poor if either CFI < 0·9 or RMSEA > 0·08 or SRMR > 0·08. We chose this combination because each fit statistic has advantages and disadvantages. Together, these three statistics considered in conjunction minimize the risk of choosing a bad model. Although low SRMR implies low model residuals, it does not incorporate model complexity and may be partial to overly complex models or models with larger sample sizes. The RMSEA provides an index of model discrepancy per degree of freedom (which accounts for model complexity), but tends to improve with larger sample size. The CFI compares an estimated model with a hypothetical null baseline model.

Results

Supplementary Table 10 displays model fit statistics for measurement models of each of the five cognitive domains, by each of the seven study groups (six HCAP studies with LASI-DAD stratified by literacy). Of these 35 measurement models, 31% (11 models) were of perfect or good fit, 57% (20 models) were of adequate fit, and the remaining 11% (4 models) were of poor fit. The single "perfect" model, for executive functioning in CHARLS-HCAP, included only 2 indicators, and thus was a saturated model. Two of the four poorly fitting models were in the general cognitive function domain. Ultimately, we proceeded with these factor structures because most model fits, including all for HRS-HCAP, were good or adequate.

References for Supplementary material on HCAP-specific factor analyses

1. Mukherjee S, Choi S-E, Lee ML, *et al.* Cognitive domain harmonization and cocalibration in studies of older adults. *Neuropsychology* 2022; **37**(4):409-423.
2. Gibbons RD, Bock RD, Hedeker D, *et al.* Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement* 2007; **31**: 4–19.
3. Hu LT, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling* 1999;6(1).
4. Gross AL, Khobragade PY, Meijer E, Saxton JA. Measurement and Structure of Cognition in the Longitudinal Aging Study in India–Diagnostic Assessment of Dementia. *J Am Geriatr Soc* 2020;68: S11-9.

Supplementary Table 2. Model fit statistics of CFAs for each cognitive domain in each study: Results from HCAP studies (N=21,144)

Cognitive domain	Study	Number of items	RMSEA	CFI	SRMR	Bifactor structure	Summary of fit
General cognition	HRS-HCAP	45	0.035	0.925	0.078	Yes	Adequate
General cognition	ELSA-HCAP	41	0.027	0.968	0.089	Yes	Poor
General cognition	LASI-DAD - literate	48	0.033	0.904	0.066	Yes	Adequate
General cognition	LASI-DAD - illiterate	48	0.036	0.904	0.063	Yes	Adequate
General cognition	Mex-Cog	40	0.040	0.932	0.072	Yes	Adequate
General cognition	CHARLS-HCAP	31	0.032	0.949	0.051	No	Adequate
General cognition	HAALSI-HCAP	51	0.043	0.913	0.122	Yes	Poor
Memory	HRS-HCAP	11	0.045	0.980	0.023	Yes	Good
Memory	ELSA-HCAP	11	0.060	0.971	0.038	Yes	Adequate
Memory	LASI-DAD - literate	11	0.046	0.978	0.027	Yes	Good
Memory	LASI-DAD - illiterate	11	0.049	0.965	0.031	Yes	Good
Memory	Mex-Cog	10	0.048	0.985	0.033	Yes	Good
Memory	CHARLS-HCAP	5	0.047	0.984	0.020	No	Good
Memory	HAALSI-HCAP	9	0.026	0.995	0.018	Yes	Good
Orientation	HRS-HCAP	10	0.028	0.971	0.064	No	Adequate
Orientation	ELSA-HCAP	9	0.010	0.999	0.052	No	Adequate
Orientation	LASI-DAD - literate	10	0.053	0.924	0.077	Yes	Adequate
Orientation	LASI-DAD - illiterate	10	0.049	0.945	0.064	Yes	Adequate
Orientation	Mex-Cog	8	0.062	0.924	0.066	No	Adequate
Orientation	CHARLS-HCAP	10	0.043	0.968	0.051	No	Adequate
Orientation	HAALSI-HCAP	10	0.032	0.989	0.069	No	Adequate
Language	HRS-HCAP	14	0.020	0.971	0.071	No	Adequate
Language	ELSA-HCAP	12	0.007	0.997	0.070	Yes	Adequate
Language	LASI-DAD - literate	14	0.034	0.897	0.070	Yes	Poor
Language	LASI-DAD - illiterate	14	0.032	0.949	0.050	Yes	Adequate
Language	Mex-Cog	13	0.016	0.986	0.073	Yes	Adequate
Language	CHARLS-HCAP	13	0.029	0.960	0.046	No	Good
Language	HAALSI-HCAP	16	0.039	0.971	0.127	Yes	Poor
Executive functioning	HRS-HCAP	8	0.076	0.973	0.020	Yes	Adequate
Executive functioning	ELSA-HCAP	8	0.080	0.966	0.020	Yes	Adequate
Executive functioning	LASI-DAD - literate	10	0.029	0.989	0.024	No	Good
Executive functioning	LASI-DAD - illiterate	10	0.038	0.975	0.034	No	Good
Executive functioning	Mex-Cog	7	0.043	0.995	0.018	Yes	Good
Executive functioning	CHARLS-HCAP	2	0.000	1.000	0.000	No	Perfect
Executive functioning	HAALSI-HCAP	12	0.076	0.922	0.059	Yes	Adequate

Legend. CFI: confirmatory fit index; RMSEA: root mean square error of approximation; SRMR: standardized root mean squared residual.

Using the combination of the RMSEA, CFI, and SRMR, we summarized model fit as perfect, good, adequate, or poor. Model fit was considered perfect if CFI = 1 and RMSEA = 0 and SRMR = 0, good if CFI ≥ 0·95 and RMSEA ≤ 0·05 and SRMR ≤ 0·05, adequate if CFI ≥ 0·90 and RMSEA ≤ 0·08 and SRMR ≤ 0·08, and poor if either CFI < 0·9 or RMSEA > 0·08 or SRMR > 0·08. We chose this combination because each fit statistic has advantages and disadvantages. Together, these three statistics considered in conjunction minimize the risk of choosing a bad model. Although low SRMR implies low model residuals, it does not incorporate model complexity and may be partial to

overly complex models or models with larger sample sizes. The RMSEA provides an index of model discrepancy per degree of freedom (which accounts for model complexity), but tends to improve with larger sample size. The CFI compares an estimated model with a hypothetical null baseline model.

Supplementary Methods: Does ordering of studies matter in the statistical co-calibration procedure that uses item banking?

Based on knowledge of factor analysis and statistical co-calibration, one should expect the ordering in which studies are added should not greatly affect the distributions of harmonized factor scores or cross-study comparisons. This is because, if the prestatistical harmonization was rigorous in its identification of linking items, then in the absence of strong DIF in many cognitive test items, the relative ranking of people (and studies) should be relatively unaffected.

To address the question of whether ordering matters, in a limited sensitivity analysis we re-arranged the order of studies in our item banking procedure 3 times. The first scenario was the original, with an ordering of studies of HRS-HCAP, ELSA-HCAP, LASI-DAD (literate then illiterate), Mex-Cog, CHARLS-HCAP, and HAALSI-HCAP. The ordering of studies in the second scenario was LASI-DAD (illiterate then literate), CHARLS-HCAP, Mex-Cog, HAALSI-HCAP, ELSA-HCAP, and HRS-HCAP. The ordering of studies in the third scenario was CHARLS-HCAP, HAALSI-HCAP, LASI-DAD (illiterate then literate), Mex-Cog, HRS-HCAP, and ELSA-HCAP. We did not estimate scores for every possible ordering of studies, but based on methodological theory, that is probably not necessary.

Below is a table showing overall and study-specific correlations among the ordering scenarios for general cognitive function and each domain-specific score. General cognitive scores from all ordering scenarios are highly correlated with each other (r's>0.96).

Domain	Correlation	Overall	HRS-HCAP	ELSA-HCAP	LASI-DAD (literate)	LASI-DAD (illiterate)	Mex-Cog	CHARLS-HCAP	HAALSI-HCAP
General cognitive performance									
	Corr(scenario1,scenario2)	0.963	0.985	0.988	0.993	0.995	0.992	0.963	0.995
	Corr(scenario1,scenario3)	0.984	0.989	0.990	0.990	0.990	0.995	0.975	0.994
	Corr(scenario2,scenario3)	0.970	0.990	0.990	0.988	0.992	0.991	0.987	0.989
Orientation									
	Corr(scenario1,scenario2)	0.923	0.945	0.959	0.944	0.947	0.883	0.916	0.951
	Corr(scenario1,scenario3)	0.966	0.978	0.985	0.971	0.969	0.960	0.956	0.970
	Corr(scenario2,scenario3)	0.929	0.937	0.958	0.933	0.925	0.886	0.943	0.937
Memory									
	Corr(scenario1,scenario2)	0.943	0.988	0.990	0.992	0.993	0.992	0.990	0.994
	Corr(scenario1,scenario3)	0.925	0.984	0.984	0.991	0.986	0.994	0.975	0.991
	Corr(scenario2,scenario3)	0.960	0.973	0.972	0.991	0.984	0.997	0.988	0.990
Executive functioning									
	Corr(scenario1,scenario2)	0.856	0.995	0.982	0.985	0.988	0.970	0.944	0.988
	Corr(scenario1,scenario3)	0.824	1.000	0.964	0.867	0.888	0.955	0.896	0.969
	Corr(scenario2,scenario3)	0.800	0.995	0.989	0.863	0.902	0.944	0.942	0.973
Language									
	Corr(scenario1,scenario2)	0.937	0.938	0.959	0.904	0.918	0.936	0.942	0.963
	Corr(scenario1,scenario3)	0.947	0.967	0.986	0.930	0.925	0.956	0.918	0.979
	Corr(scenario2,scenario3)	0.958	0.965	0.984	0.899	0.979	0.946	0.967	0.978

With respect to correlations among domain-specific factors, most study-specific correlations among the scores from different ordering scenarios were r>0.95 or higher; correlations were uniformly highest (r>0.97) for the memory which had the richest set of cognitive test indicators across studies, defined by polytomous items and linking items between studies. Language was similarly highly correlated across the ordering scenarios. Correlations among ordering scenarios for orientation were as low as r=0.88 in Mex-Cog between scenario 2 and others, but otherwise the average correlation among study-specific correlations for orientation was r=0.95. Correlations among ordering

scenarios for executive function were as low as r=0.86 in LASI-DAD, but otherwise the average correlation among study-specific correlations for executive functioning was r=0.95. In the Table below, we highlighted in yellow the lowest correlations observed by domain.

The following table shows study-specific means and standard deviations of the general cognitive factor; we do not show specific domains because they all follow the same general patterns as for the general cognitive factor. Highlighted in yellow in each column is the study that appeared first in the item banking procedure.

Study number	HCAP study	Means by scenario			SD by scenario		
		Scenario1	Scenario2	Scenario3	Scenario1	Scenario2	Scenario3
1	HRS-HCAP	-0.003	2.141	1.059	1.004	1.244	1.028
2	ELSA-HCAP	0.003	2.266	1.125	1.073	1.351	1.098
4	LASI-DAD (literate)	-0.770	1.659	0.361	0.779	1.131	0.919
6	LASI-DAD (illiterate)	-1.853	-0.009	-0.965	0.615	0.950	0.745
3	Mex-Cog	-0.840	1.386	0.177	0.993	1.433	1.090
9	CHARLS-HCAP	-1.188	0.410	-0.084	0.983	1.091	0.934
5	HAALSI-HCAP	-1.283	0.883	-0.163	0.771	1.167	0.995

We can glean 4 inferences from this table.

1. First, by design, the mean of the estimated factor score is close to 0 and the SD is close to 1 in the first study in the given ordering scenario (e.g., CHARLS-HCAP was the first study in ordering scenario 3, thus it's mean of -0.084 is closer to 0 than any other study for that ordering scenario). This result is by design because the latent trait mean (SD) is set to 0 (1) in the first study; this is what helps us interpret deviation differences in scores.
    a. Because of differences between a latent variable in latent variable space and an observed factor score estimated in real data, we do not expect factor scores themselves to have a mean of exactly 0 and a standard deviation of 1.0. Usually, the standard deviation of an estimated factor score is less than 1.0 because of the lack of extreme outliers beyond -4 or +4 SD units in real data.
2. A second inference we can make is that, because the scores in different ordering scenarios are scaled to the first HCAP, a point shift means something different across different scenarios. This happens because different HCAPs have differing variances of cognitive function. Looking at the column of standard deviations, the SD for general cognition in ELSA-HCAP is greater than in HAALSI-HCAP, regardless of whether 1.073>0.771 (scenario 1), 1.351>1.167 (scenario 2), or 1.098>0.995 (scenario 3).
3. The third inference, notwithstanding the above caveat of inference 2, is that between ordering scenarios 1 and 2, there is a fairly constant mean upshift in scores of between 2.1-2.4 points. Two exceptions that have a relatively attenuated upshift are: LASI-DAD illiterates (upshift of just 1.84 points) and CHARLS-HCAP (upshift of just 1.60 points). Between ordering scenarios 1 and 3, there is a constant mean upshift in scores of between 1.0-1.1 points; the exception was among LASI-DAD illiterates (upshift 0.89).
4. A fourth inference we make is that, because of the lesser upshift in mean scores of certain studies between ordering scenarios, the relative ranking of country means across ordering scenarios varies slightly. Using study numbers in the above table:
    Under ordering scenario 1, the ordering of means is 6<5=9<3<4<1=2.
    Under ordering scenario 2, the ordering of means is 6<9<5<3<4<2>1.
    Under ordering scenario 3, the ordering of means is 6<5=9<3<4<1=2.

Ultimately, based on evidence from patterns in correlations among the different ordering scenarios, coupled with comparisons of means across the different scenarios, we conclude that the order of studies in the item bank should not affect the validity or reliability of resulting scores. The interpretation of a unit change does depend on the reference group, to the extent that different groups are different in variability of the underlying latent trait.

A caveat to the empirical conclusion regarding ordering of studies is that ordering of studies may be constrained based on availability of linking items. For example, in the following hypothetical scenario, studies A and C have no linking items and thus cannot be co-calibrated without study B. Study B cannot be added last. Acceptable ordering could be ABC, BAC, BCA, CBA. Unacceptable orders for the item banking procedure would be ACB, CAB.

Indicator	Hypothetical study		
	A	B	C
Item 1	Present		
Item 2	Present		
Item 3	Present	Present	
Item 4		Present	Present
Item 5		Present	Present
Item 6			Present

Supplementary Table 3. Information about HCAP studies

Characteristic	United States	England	India	Mexico	China	South Africa
Parent cohort study	Health and Retirement study (HRS)	English Longitudinal Study on Ageing (ELSA)	Longitudinal Aging Study in India (LASI)	Mexican Health and Aging Study (MHAS)	China Health and Retirement Longitudinal Study (CHARLS)	Health and Aging in Africa: A Longitudinal Study of an INDEPTH Community in South Africa (HAALSI)
HCAP sub-study	Harmonized Cognitive Assessment Protocol Project of Health and Retirement study (HRS-HCAP)	Harmonised Cognitive Assessment Protocol Sub-study of the English Longitudinal Study of Ageing (ELSA-HCAP)	Harmonised Diagnostic Assessment of Dementia for the Longitudinal Aging Study in India (LASI-DAD)	Mexican Cognitive Aging Ancillary Study (Mex-Cog)	Harmonized Cognitive Assessment Protocol for the China Health and Retirement Longitudinal Study (CHARLS-HCAP)	Cognition and dementia in the Health and Aging in Africa Longitudinal Study of an INDEPTH community in South Africa (HAALSI-HCAP)
Dates of study recruitment	June 2016 – October 2017	January 2018 – April 2018	October 2017 – June 2018, October 2018 – May 2019	October 2015 – December 2015	July 12, 2017 – August 31, 2017	September 9, 2019 – January 13, 2020
Eligibility criteria for HCAPs	1. Aged 65 years and over at the time of HRS-HCAP survey 2. Completed core interview	1. Aged 65 years and over at the time of ELSA-HCAP survey 2. Completed an ELSA interview in wave 7 (2014-15) or wave 8 (2016-17)	1. Aged 60 years and over at the time of LASI-DAD survey	1. Aged 55 and over in the MHAS 2015 2. Completed a direct or proxy interview for health reasons in the MHAS 2015	1. Aged 60 and over at the time of CHARLS-HCAP	1. Aged 50 and over
Number of target cases to conduct HCAP interviews	4,425	1,778	3,891	3,250	Not reported	Not reported
Number of completed HCAP interviews	3,496	1,273	4,096	2,042	9,755	628
Language used during interviews						
English	3174	1273	10			
Spanish	170			2042		
Mandarin					9755	
Hindi			1,393			
Kannada			245			
Malayalam			349			
Gujarati			288			
Tamil			301			
Punjabi			159			

Urdu	152	
Bengali	309	
Assamese	199	
Odiya	252	
Marathi	250	
Telugu	189	
xiTsonga		631

Supplementary Table 4. Factor loadings of cognitive test items and item overlap across studies: Results from HCAP studies (N=21,144)

Variable	HRS-HCAP	ELSA-HCAP	LASI-DAD (literate)	LASI-DAD (illiterate)	Mex-Cog	CHARLS-HCAP	HAALSI-HCAP
Orientation							
Day of month	0·68 (0·43)	0·68 (0·43)	0·68 (0·43)	0·68 (0·43)	0·68 (0·43)	0·68 (0·43)	0·68 (0·43)
Month	0·88 (0·73)	0·88 (0·73)	0·88 (0·73)	0·88 (0·73)	0·88 (0·73)	0·88 (0·73)	0·88 (0·73)
Year	0·90 (0·76)	0·90 (0·76)			0·90 (0·76)	0·90 (0·76)	
Year (DIF adjusted)			0·70 (0·82)	0·70 (0·82)			0·70 (0·82)
Day of the week	0·76 (0·63)	0·76 (0·63)	0·76 (0·63)	0·76 (0·63)	0·76 (0·63)	0·76 (0·63)	0·76 (0·63)
What time is it					0·17 (0·32)		0·17 (0·32)
Where are we					0·13 (0·35)		
What country are we in		0·62 (0·72)			0·62 (0·72)		0·62 (0·72)
What state are we in	0·64 (0·61)		0·64 (0·61)	0·64 (0·61)	0·64 (0·61)	0·64 (0·61)	0·64 (0·61)
What county are we in	0·68 (0·55)	0·68 (0·55)				0·68 (0·55)	
What city are we in	0·83 (0·72)	0·83 (0·72)	0·83 (0·72)	0·83 (0·72)		0·83 (0·72)	0·83 (0·72)
Season of year	0·54 (0·42)	0·54 (0·42)	0·54 (0·42)	0·54 (0·42)		0·54 (0·42)	0·54 (0·42)
Floor of building	0·66 (0·56)		0·66 (0·56)	0·66 (0·56)		0·66 (0·56)	0·66 (0·56)
Address (street name and/or building number)	0·76 (0·62)	0·76 (0·62)	0·76 (0·62)	0·76 (0·62)		0·76 (0·62)	
Name of hospital or district/municipality			0·63 (0·69)	0·63 (0·69)			0·63 (0·69)
Memory							
CERAD immediate sum of 3 trials	0·87 (0·82)	0·87 (0·82)				0·87 (0·82)	
CERAD immediate sum of 3 trials			0·89 (0·69)	0·89 (0·69)	0·89 (0·69)		0·89 (0·69)
CERAD word list delay	0·88 (0·85)	0·88 (0·85)				0·88 (0·85)	
CERAD word list delay			0·88 (0·65)	0·88 (0·65)	0·88 (0·65)		0·88 (0·65)
CERAD recognition	0·75 (0·58)	0·75 (0·58)				0·75 (0·58)	
CERAD recognition			0·79 (0·65)	0·79 (0·65)	0·79 (0·65)		0·79 (0·65)
Three word immediate registration	0·51 (0·49)	0·51 (0·49)	0·51 (0·49)	0·51 (0·49)	0·51 (0·49)	0·51 (0·49)	0·51 (0·49)
Three word delayed recall	0·76 (0·65)	0·76 (0·65)	0·76 (0·65)	0·76 (0·65)	0·76 (0·65)	0·76 (0·65)	0·76 (0·65)
Logical Memory immediate	0·71 (0·65)	0·71 (0·65)	0·71 (0·65)	0·71 (0·65)	0·71 (0·65)		0·71 (0·65)
Logical Memory delay	0·74 (0·67)	0·74 (0·67)	0·74 (0·67)	0·74 (0·67)	0·74 (0·67)		0·74 (0·67)
Logical memory recognition	0·62 (0·52)	0·62 (0·52)	0·62 (0·52)	0·62 (0·52)			0·62 (0·52)
Brave man immediate (East Boston Memory Test)	0·42 (0·39)	0·42 (0·39)	0·42 (0·39)	0·42 (0·39)	0·42 (0·39)		
Brave man delay (East Boston Memory Test)	0·53 (0·43)	0·53 (0·43)	0·53 (0·43)	0·53 (0·43)	0·53 (0·43)		
CERAD constructional praxis delay	0·70 (0·67)	0·70 (0·67)	0·70 (0·67)	0·70 (0·67)	0·70 (0·67)		0·70 (0·67)
Executive functioning							
Problem solving			0·75 (0·71)	0·75 (0·71)			

Ravens progressive matrices	0·74 (0·68)	0·74 (0·68)	0·74 (0·68)	0·74 (0·68)			0·74 (0·68)
HRS Number series	0·64 (0·57)	0·64 (0·57)					
Number series						0·56 (0·44)	
Trails A time (letters and numbers)	0·79 (0·76)	0·79 (0·76)					
Trails B time (letters and numbers)	0·76 (0·67)	0·76 (0·67)					
Similarities			0·58 (0·57)	0·58 (0·57)	0·58 (0·57)		0·58 (0·57)
Token Test			0·77 (0·75)	0·77 (0·75)			0·77 (0·75)
Digit Span Forward (single item)			0·68 (0·60)	0·68 (0·60)			
Digit Span Backward (single item)			0·73 (0·64)	0·73 (0·64)			
Digit Span Forward (multiple items)							0·48 (0·38)
Digit Span Backward (multiple items)							0·43 (0·28)
Go-No-Go			0·71 (0·67)	0·71 (0·67)	0·71 (0·67)		0·71 (0·67)
Motor Programming							0·76 (0·74)
MMSE Spelling backwards	0·62 (0·65)						
Backward counting, 100-0	0·69 (0·63)	0·69 (0·63)					
Backward counting, 20-0					0·54 (0·79)		
Symbol Digit Modalities Test *	0·88 (0·77)	0·88 (0·77)					
Symbols and Digits test **					0·58 (0·77)		
Symbol Cancellation Test			0·67 (0·68)	0·67 (0·68)	0·67 (0·68)		0·67 (0·68)
Letter cancellation	0·59 (0·57)	0·59 (0·57)					
Serial 3s					0·36 (0·54)		
Serial 7s		0·56 (0·53)	0·56 (0·53)	0·56 (0·53)	0·56 (0·53)	0·56 (0·53)	0·56 (0·53)
Backward Day naming			0·68 (0·68)	0·68 (0·68)			0·68 (0·68)
Forward day naming							0·74 (0·68)
CDR calculation-cent							0·62 (0·69)
Language							
Animal fluency	0·74 (0·68)	0·74 (0·68)	0·74 (0·68)	0·74 (0·68)	0·74 (0·68)	0·74 (0·68)	0·74 (0·68)
Name a described cactus	0·81 (0·70)	0·81 (0·70)					
Name a described cactus (DIF adjusted)						0·79 (0·60)	
Name a described coconut			0·56 (0·45)	0·56 (0·45)			
What are scissors used for?	0·71 (0·54)	0·71 (0·54)	0·71 (0·54)	0·71 (0·54)	0·71 (0·54)	0·71 (0·54)	0·71 (0·54)
Object naming (watch)	0·69 (0·59)		0·69 (0·59)	0·69 (0·59)	0·69 (0·59)	0·69 (0·59)	0·69 (0·59)
Object naming (pencil)	0·56 (0·57)		0·56 (0·57)	0·56 (0·57)	0·56 (0·57)	0·56 (0·57)	0·56 (0·57)
Name the elbow	0·88 (0·68)	0·88 (0·68)	0·88 (0·68)	0·88 (0·68)	0·88 (0·68)	0·88 (0·68)	0·88 (0·68)
Write a sentence (or write one's name)	0·62 (0·52)	0·62 (0·52)	0·62 (0·52)		0·62 (0·52)	0·62 (0·52)	0·62 (0·52)
Say a sentence				0·81 (0·67)			

Read and follow command (Close your eyes)	0·61 (0·46)	0·61 (0·46)	0·61 (0·46)		0·61 (0·46)		0·61 (0·46)
Read and follow command (DIF adjusted)						0·73 (0·58)	
Follow example (close your eyes)				0·86 (0·71)			
Repetition of a phrase	0·46 (0·37)	0·46 (0·37)	0·46 (0·37)	0·46 (0·37)	0·46 (0·37)	0·46 (0·37)	0·46 (0·37)
What does one do with a hammer	0·43 (0·24)	0·43 (0·24)	0·43 (0·24)	0·43 (0·24)	0·43 (0·24)	0·43 (0·24)	0·43 (0·24)
Define Bridge					0·62 (0·54)		
Point to 2 things in the vicinity	0·85 (0·65)	0·85 (0·65)	0·85 (0·65)	0·85 (0·65)	0·85 (0·65)	0·85 (0·65)	0·85 (0·65)
Where is the local market?	0·58 (0·48)	0·58 (0·48)	0·58 (0·48)	0·58 (0·48)	0·58 (0·48)	0·58 (0·48)	0·58 (0·48)
Follow 3-stage instruction	0·39 (0·30)	0·39 (0·30)	0·39 (0·30)	0·39 (0·30)	0·39 (0·30)		0·39 (0·30)
Name president or Prime Minister	0·85 (0·82)	0·85 (0·82)	0·85 (0·82)	0·85 (0·82)		0·85 (0·82)	0·85 (0·82)
Name deputy president							0·87 (0·75)
Phonemic Fluency							0·59 (0·63)
Boston Naming Test, uncued							0·68 (0·62)

Legend. This table shows factor loadings for domain-specific factor analyses, and in parentheses the factor loadings for the model for general cognitive function. Loadings are standardized to have a range from -1 to 1, thus can be interpretable as correlations between items and the underlying factor. The presence of factor loadings for a given test item in each column reflects decisions about the comparability of items made at the prestatistical harmonization as well as after testing for differential item functioning. Refer to the Methods and Results for details.

* (110 items, correctly assign numbers to symbols provided on the test's page based on a key)

** (56 items, correctly assign symbols provided on the test's page based on a key)

Supplementary Table 5. Crosswalk between scaled educational attainment equivalents based on ISCED 2011 and education categories in each study

ISCED attainment level	HRS-HCAP	ELSA-HCAP	LASI-DAD	Mex-Cog	CHARLS-HCAP	HAALSI-HCAP
0: No or Early Childhood education	No degree	No formal education	(1) 0 year of schooling (2) Less than primary school (Standard 1-4)	0-5 years of school	(1) No formal education (illiterate) (2) Did not finish primary school	(1) No formal education; (2) Preschool
1: Primary education	Grades 1-6	Not available from ELSA data	Primary school completed (Standard 5-7)	6 years of school	(1) Sishu/home school (2) Elementary school	Grades 1-6 (Sub-A through Std 4)
2: Lower secondary education	Grades 7-12	No qualification; Foreign/undetermined qualification, if age at which continuous full-time education finished is ≤ 16 years	Middle school completed (Standard 8- 9)	7-9 years of school	Middle school	Grades 7-9 (Std 4-7)
3: Upper secondary education	GED/high school diploma	(1) NVQ1/CSE or other grade equivalent (2) NVQ2/GCE O level equivalent (3) Foreign/undetermined qualification, if age at which continuous full-time education finished is 17 and 18 years	(1) Secondary school/Matriculation completed (2) Higher secondary/intermediate/senior secondary completed	10-12 years of school	High School	Grades 10-12 (Std 8-10)
4-8: Post-secondary education	(1) Some college; (2) Masters degree; (3) Professional degree; (4) Grades > 12 if degree unknown	(1) NVQ3/GCE A level equivalent; (2) Higher education below degree; (3) NVQ4/NVQ5/degree or equivalent; (4) Foreign/undetermined qualification if age at which full-time education finished is 19 years or older	(1) Diploma or certificate; (2) Graduate degree (B.A., B.Sc., B. Com.); (3) Post-graduate degree or (M.A., M.Sc., M. Com., M.Phil, Ph.D., Post-Doc); (4) Professional course/degree	13+ years of school	(1) Vocational school (2) Two/three year college/associate degree (3) Four year college/Bachelors' degree (4) Master's degree (5) Doctoral degree	(1) Partial or complete tertiary education; (2) partial or complete University education

Supplementary Table 6. Estimated thresholds and intercepts of cognitive test items: Results from HCAP studies (N=21,144)

Cognitive test item	Is the item categorical or continuous	Intercept (continuous items)	Threshold (categorical items)								
			1	2	3	4	5	6	7	8	9
Orientation											
Day of month	Categorical		-1.29 (-1.01)								
Month	Categorical		-4.28 (-2.87)								
Year	Categorical		-3.85 (-2.81)								
Year (DIF adjusted)	Categorical		-1.49 (-2.44)								
Day of the week	Categorical		-2.51 (-2.26)								
What time is it	Categorical		-0.68 (-0.85)								
Where are we	Categorical		-1.19 (-1.47)								
What country are we in	Categorical		-3.36 (-3.75)								
What state are we in	Categorical		-2.92 (-3.02)								
What county are we in	Categorical		-2.11 (-1.95)								
What city are we in	Categorical		-3.76 (-3.30)								
Season of year	Categorical		-1.33 (-1.27)								
Floor of building	Categorical		-2.65 (-2.51)								
Address (street name and/or building number)	Categorical		-2.54 (-2.25)								
Name of hospital or district/municipality	Categorical		-2.04 (-2.51)								
Memory											
CERAD immediate sum of 3 trials	Continuous	17.40 (17.40)									
CERAD immediate sum of 3 trials	Continuous	17.29 (16.64)									
CERAD word list delay	Continuous	5.12 (5.12)									
CERAD word list delay	Continuous	5.64 (5.25)									
CERAD recognition	Continuous	18.49 (18.49)									
CERAD recognition	Continuous	19.17 (19.02)									
Three word immediate registration	Categorical		-2.37 (-2.47)	-1.51 (-1.57)							
Three word delayed recall	Categorical		-2.67 (-2.55)	-1.87 (-1.78)	-0.64 (-0.61)						
Logical Memory immediate	Continuous	9.83 (9.83)									
Logical Memory delay	Continuous	7.39 (7.39)									
Logical memory recognition	Continuous	10.28 (10.28)									

Brave man immediate (East Boston Memory Test)	Categorical		-2.61 (-2.66)	-1.36 (-1.38)	-0.16 (-0.17)	0.90 (0.92)	2.03 (2.07)	3.28 (3.34)				
Brave man delay (East Boston Memory Test)	Categorical		-0.96 (-0.92)	-0.09 (-0.09)	0.82 (0.79)	1.85 (1.78)	3.23 (3.11)	4.41 (4.24)				
CERAD constructional praxis delay	Continuous	5.81 (5.81)										
Executive functioning												
Problem solving	Categorical		-3.35 (-3.40)	-2.03 (-2.11)	-0.73 (-0.84)							
Ravens progressive matrices	Continuous	12.34 (12.34)										
HRS Number series	Continuous	-0.06 (-0.06)										
Number series	Continuous	7.04 (9.12)										
Trails A time (letters and numbers)	Continuous	0.06 (0.06)										
Trails B time (letters and numbers)	Continuous	0.01 (0.01)										
Similarities	Categorical		-2.06 (-2.20)	-0.98 (-1.11)	-0.05 (-0.17)							
Token Test	Categorical		-3.99 (-4.16)	-3.46 (-3.63)	-3.06 (-3.23)	-2.27 (-2.45)	-1.53 (-1.71)	-0.80 (-0.97)	0.01 (-0.16)			
Digit Span Forward (single item)	Categorical		-0.46 (-0.51)									
Digit Span Backward (single item)	Categorical		-0.81 (-0.81)									
Digit Span Forward (multiple items)	Categorical		-1.51 (-1.34)	-0.30 (-0.17)	0.82 (0.92)							
Digit Span Backward (multiple items)	Categorical		-0.69 (-0.48)	0.36 (0.52)	1.25 (1.38)							
Go-No-Go	Categorical		-2.69 (-2.77)	-2.49 (-2.57)	-2.36 (-2.45)	-2.17 (-2.26)	-1.89 (-1.99)	-1.71 (-1.80)	-1.48 (-1.58)	-1.19 (-1.30)	-0.79 (-0.90)	
Motor Programming	Categorical		-4.01 (-3.93)	-3.81 (-3.75)	-3.48 (-3.42)	-3.08 (-3.03)	-2.39 (-2.35)	-1.65 (-1.64)				
MMSE Spelling backwards	Categorical		-3.28 (-2.07)	-2.90 (-1.82)	-2.36 (-1.49)	-1.54 (-0.97)	-1.24 (-0.78)					
Backward counting, 100-0	Continuous	29.27 (29.27)										
Backward counting, 20-0	Categorical		-2.04 (-2.84)	-1.93 (-2.73)	-1.71 (-2.51)	-0.80 (-1.58)						
Symbol Digit Modalities Test *	Continuous	31.81 (31.81)										
Symbols and Digits test **	Continuous	25.48 (31.50)										
Symbol Cancellation Test	Continuous	16.39 (17.68)										
Letter cancellation	Continuous	9.69 (9.69)										
Serial 3s	Categorical		-2.28 (-2.53)	-1.64 (-1.92)	-1.28 (-1.58)	-0.80 (-1.12)	0.13 (-0.23)					

Serial 7s	Categorical		-1.49 (-1.41)	-0.44 (-0.37)	-0.22 (-0.15)	-0.02 (0.05)	0.15 (0.23)
Backward Day naming	Categorical		-2.46 (-2.70)	-2.34 (-2.57)	-2.19 (-2.42)	-2.08 (-2.31)	-1.70 (-1.92)
Forward day naming	Categorical		-2.81 (-2.60)				
CDR calculation-cent	Categorical		-0.13 (-0.36)				
Language							
Animal fluency	Continuous	15.97 (15.97)					
Name a described cactus	Categorical		-2.26 (-2.06)				
Name a described cactus (DIF adjusted)	Categorical		-0.55 (0.21)				
Name a described coconut	Categorical		-1.25 (-0.96)				
What are scissors used for?	Categorical		-3.01 (-2.58)				
Object naming (watch)	Categorical		-4.20 (-3.33)				
Object naming (pencil)	Categorical		-3.31 (-3.09)				
Name the elbow	Categorical		-4.50 (-3.21)				
Write a sentence (or write one's name)	Categorical		-1.84 (-1.80)				
Say a sentence	Categorical		-3.12 (-2.96)				
Read and follow command (Close your eyes)	Categorical		-2.36 (-2.21)				
Read and follow command (DIF adjusted)	Categorical		-2.48 (-1.76)				
Follow example (close your eyes)	Categorical		-3.56 (-3.21)				
Repetition of a phrase	Categorical		-0.59 (-0.58)				
What does one do with a hammer	Categorical		-1.57 (-1.49)				
Define Bridge	Categorical		-1.39 (-1.56)				
Point to 2 things in the vicinity	Categorical		-4.05 (-3.13)				
Where is the local market?	Categorical		-1.05 (-1.03)				
Follow 3-stage instruction	Categorical		-2.80 (-2.76)	-2.03 (-2.00)	-0.67 (-0.66)		
Name president or Prime Minister	Categorical		-2.85 (-2.99)				
Name deputy president	Categorical		-0.98 (-0.75)				
Phonemic Fluency	Continuous	3.86 (4.37)					
Boston Naming Test, uncued	Continuous	15.21 (15.35)					

Legend. Intercepts and thresholds were estimated from factor analysis models using a maximum likelihood estimator with a probit link. For each item, parameters from domain-specific factor analyses are shown, and in parentheses is the corresponding parameter for the model for general cognitive function.

Supplementary Table 7. Results of differential item functioning among confident and tentative linking items for the language domain: Results from HCAP studies (N=21,144)

Study	Stage of DIF testing	Cognitive test item	Association with cohort (REF: HRS-HCAP)a	95% CI lower bound	95% CI upper bound	Interpretation (b)
HAALSI-HCAP						
	DIF among confident linking items					
		Name the elbow	1.65	1.31	2.08	DIF
		Animal fluency	N/A			No DIF
		What are scissors used for?	N/A			No DIF
		Point to 2 things in the vicinity	N/A			No DIF
		Name president or Prime Minister	N/A			No DIF
	DIF among tentative linking items, treating confident items as anchors					
		What does one do with a hammer	2.64	2.14	3.27	DIF
		Where is the local market?	2.06	1.81	2.34	DIF
ELSA-HCAP						
	DIF among confident linking items					
		Name a described cactus	0.61	0.55	0.67	DIF
		What does one do with a hammer	2.29	1.86	2.82	DIF
		Follow 3-stage instruction	1.83	1.66	2.03	DIF
		Name president or Prime Minister	0.43	0.39	0.47	DIF
		Animal fluency	N/A			No DIF
		Point to 2 things in the vicinity	N/A			No DIF
		What are scissors used for?	N/A			No DIF
		Name the elbow	N/A			No DIF
		Read and follow command (Close your eyes)	N/A			No DIF
		Repetition of a phrase	N/A			No DIF
	DIF among tentative linking items, treating confident items as anchors					
		Where is the local market?	2.31	2.03	2.62	DIF

	Write a sentence (or write one's name)	N/A			No DIF
LASI-DAD					
	DIF among confident linking items				
	What are scissors used for?	N/A			No DIF
	Object naming (pencil)	N/A			No DIF
	Write a sentence (or write one's name)	1.40	1.24	1.58	Negligible
	Read and follow command (Close your eyes)	0.14	0.13	0.16	DIF
	Follow 3-stage instruction	1.65	1.51	1.80	DIF
	Animal fluency	N/A			No DIF
	Object naming (watch)	N/A			No DIF
	Name the elbow	1.13	0.96	1.32	Negligible
	Point to 2 things in the vicinity	N/A			No DIF
	Name president or Prime Minister	N/A			No DIF
	DIF among tentative linking items, treating confident items as anchors				
	Repetition of a phrase	4.85	4.32	5.44	DIF
	What does one do with a hammer	N/A			No DIF
	Where is the local market?	3.79	3.36	4.27	DIF
Mex-Cog					
	DIF among confident linking items				
	Animal fluency	N/A			No DIF
	Object naming (pencil)	N/A			No DIF
	Name the elbow	1.03	0.87	1.21	Negligible
	Read and follow command (Close your eyes)	N/A			No DIF
	Point to 2 things in the vicinity	N/A			No DIF
	Follow 3-stage instruction	N/A			No DIF
	DIF among tentative linking items, treating confident items as anchors				
	What are scissors used for?	1.71	1.42	2.06	DIF
	Repetition of a phrase	3.21	2.93	3.51	DIF
	What does one do with a hammer	2.08	1.83	2.37	DIF
	Where is the local market?	0.48	0.45	0.52	DIF
	Object naming (watch)	N/A			No DIF

	Write a sentence (or write one's name)	N/A			No DIF
CHARLS-HCAP					
DIF among confident linking items					
	Name the elbow	0.44	0.39	0.50	DIF
	Write a sentence (or write one's name)	0.53	0.48	0.58	DIF
	What does one do with a hammer	2.32	2.12	2.54	DIF
	Point to 2 things in the vicinity	0.43	0.37	0.49	DIF
	Name president or Prime Minister	1.40	1.29	1.52	Negligible
	What are scissors used for?	N/A			No DIF
	Repetition of a phrase	N/A			No DIF
	Animal fluency	N/A			No DIF
DIF among tentative linking items, treating confident items as anchors					
	Object naming (watch)	0.59	0.49	0.70	DIF
	Object naming (pencil)	N/A			No DIF
	Read and follow command (Close your eyes)	1.34	1.27	1.41	Negligible
	Where is the local market?	1.47	1.37	1.57	Negligible
	Name a described cactus	3.65	3.3	4.04	DIF
	Follow 3-stage instruction	N/A			No DIF

Legend. Odds Ratios (OR) illustrate the difference (on an odds scale) in outcome between and a given study and HRS-HCAP, adjusting for the latent ability. An OR greater than 1 implies better performance than expected on the item in the focal study, compared to HRS-HCAP, whereas an OR less than 1 indicates better performance on the item than expected in HRS-HCAP, compared to the focal study.

N/A =Not Applicable. This is shown when an item is present in a given study, but no DIF was detected; in these cases the "Interpretation" column notes "No DIF" was detected.

a Reference group is HRS-HCAP.

b Interpretation of the magnitude of DIF as negligible (between 0.66 and 1.5) or nonnegligible (DIF).

Supplementary Table 8. Results of differential item functioning among confident and tentative linking items for the memory domain: Results from HCAP studies (N=21,144)

Study	Stage of DIF testing	Cognitive test item	Association with cohort (REF: HRS-HCAP)	95% CI lower bound	95% CI upper bound	Interpretation
HAALSI-HCAP						
	DIF among confident linking items					
		Three word delayed recall	N/A			No DIF
		Three word immediate registration	N/A			No DIF
	DIF among tentative linking items, with no anchors					
		Logical Memory immediate	N/A			No DIF
		Logical Memory delay	3.49	3.01	3.97	DIF
		CERAD constructional praxis delay	-1.76	-2.10	-1.43	DIF
		Logical memory recognition	-0.17	-0.46	0.12	DIF
ELSA-HCAP						
	DIF among confident linking items					
		Three word delayed recall	0.51	0.48	0.54	Negligible
		Three word immediate registration	1.43	1.25	1.63	Negligible
		CERAD immediate sum of 3 trials	N/A			No DIF
		CERAD word list delay	N/A			No DIF
		CERAD constructional praxis delay	1.47	1.27	1.66	Negligible
		CERAD recognition	N/A			No DIF
	DIF among tentative linking items, treating confident items as anchors					
		Logical Memory immediate	-2.05	-2.27	-1.82	DIF
		Logical memory recognition	0.45	0.30	0.60	DIF
		Brave man delay (East Boston Memory Test)	1.68	1.58	1.78	DIF
		Brave man immediate (East Boston Memory Test)	2.29	2.17	2.42	DIF
		Logical Memory delay	N/A			No DIF
LASI-DAD						

	DIF among confident linking items				
	Logical Memory delay	-0.34	-0.56	-0.12	DIF
	Three word delayed recall	0.86	0.81	0.92	Negligible
	Brave man delay (East Boston Memory Test)	1.83	1.73	1.93	DIF
	Brave man immediate (East Boston Memory Test)	1.58	1.49	1.66	DIF
	Three word immediate registration	N/A			No DIF
	Logical Memory immediate	N/A			No DIF
	Logical memory recognition	N/A			No DIF
	DIF among tentative linking items, treating confident items as anchors				
	CERAD constructional praxis delay	-0.89	-1.07	-0.72	DIF
Mex-Cog					
	DIF among confident linking items				
	CERAD constructional praxis delay	N/A			No DIF
	DIF among tentative linking items, treating confident items as anchors				
	Logical Memory immediate	-2.72	-2.94	-2.49	DIF
	Logical Memory delay	-1.73	-1.98	-1.47	DIF
	Three word delayed recall	0.80	0.75	0.86	Negligible
	Brave man delay (East Boston Memory Test)	1.69	1.60	1.80	DIF
	Brave man immediate (East Boston Memory Test)	N/A			No DIF
	Three word immediate registration	N/A			No DIF
CHARLS-HCAP					
	DIF among confident linking items				
	Three word delayed recall	N/A			No DIF
	Three word immediate registration	N/A			No DIF
	DIF among tentative linking items, treating confident items as anchors				
	CERAD immediate sum of 3 trials	-0.14	-0.35	0.07	Negligible
	CERAD word list delay	0.92	0.82	1.01	Negligible
	CERAD recognition	0.87	0.73	1.01	Negligible

22

Legend. Odds Ratios (OR) illustrate the difference (on an odds scale) in outcome between and a given study and HRS-HCAP, adjusting for the latent ability. An OR greater than 1 implies better performance than expected on the item in the focal study, compared to HRS-HCAP, whereas an OR less than 1 indicates better performance on the item than expected in HRS-HCAP, compared to the focal study.
N/A =Not Applicable. This is shown when an item is present in a given study, but no DIF was detected; in these cases the "Interpretation" column notes "No DIF" was detected.
a Reference group is HRS-HCAP.
b Interpretation of the magnitude of DIF as negligible (between 0.66 and 1.5) or nonnegligible (DIF).

Supplementary Table 9. Results of differential item functioning among confident and tentative linking items for the orientation domain: Results from HCAP studies (N=21,144)

Study	Stage of DIF testing	Cognitive test item	Association with cohort (REF: HRS-HCAP)	95% CI lower bound	95% CI upper bound	Interpretation
HAALSI-HCAP						
	DIF among confident linking items					
		Day of month	N/A			No DIF
		Month	0.57	0.50	0.65	DIF
		Year	2.11	1.90	2.35	DIF
		What state are we in	0.16	0.13	0.19	DIF
		Day of the week	N/A			No DIF
		What city are we in	N/A			No DIF
	DIF among tentative linking items, treating confident items as anchors					
		Floor of building	N/A			No DIF
		Season of year	N/A			No DIF
ELSA-HCAP						
	DIF among confident linking items					
		Day of month	N/A			No DIF
		Month	N/A			No DIF
		Year	N/A			No DIF
		Day of the week	N/A			No DIF
		What city are we in	N/A			No DIF
	DIF among tentative linking items, treating confident items as anchors					
		Season of year	1.66	1.48	1.86	DIF
		Address	1.38	1.20	1.59	Negligible
		What county are we in	N/A			No DIF
LASI-DAD						
	DIF among confident linking items					
		Day of month	1.35	1.23	1.48	Negligible

	Year	1.35	1.26	1.46	Negligible
	Day of the week	0.81	0.73	0.90	Negligible
	What state are we in	0.28	0.24	0.32	DIF
	Season of year	1.47	1.33	1.63	Negligible
	Month	N/A			No DIF
	What city are we in	N/A			No DIF
DIF among tentative linking items, treating confident items as anchors					
	Floor of building	N/A			No DIF
	Address	N/A			No DIF
Mex-Cog					
DIF among confident linking items					
	Day of month	N/A			No DIF
	Month	N/A			No DIF
	Year	N/A			No DIF
	Day of the week	N/A			No DIF
	What state are we in	N/A			No DIF
DIF among tentative linking items, treating confident items as anchors					
	None				
CHARLS-HCAP					
DIF among confident linking items					
	Year	0.71	0.66	0.77	Negligible
	Day of the week	0.44	0.41	0.48	DIF
	What state are we in	0.66	0.57	0.76	Negligible
	What county are we in	1.50	1.38	1.62	Negligible
	Address	1.79	1.63	1.96	DIF
	Day of month	N/A			No DIF
	Month	N/A			No DIF
	What city are we in	N/A			No DIF
	Floor of building	N/A			No DIF
DIF among tentative linking items, treating confident items as anchors					
	Season of year	N/A			No DIF

33

Legend. Odds Ratios (OR) illustrate the difference (on an odds scale) in outcome between and a given study and HRS-HCAP, adjusting for the latent ability. An OR greater than 1 implies better performance than expected on the item in the focal study, compared to HRS-HCAP, whereas an OR less than 1 indicates better performance on the item than expected in HRS-HCAP, compared to the focal study.
N/A =Not Applicable. This is shown when an item is present in a given study, but no DIF was detected; in these cases the "Interpretation" column notes "No DIF" was detected.
a Reference group is HRS-HCAP.
b Interpretation of the magnitude of DIF as negligible (between 0.66 and 1.5) or nonnegligible (DIF).

Supplementary Table 10. Validation of the domain-specific cognitive function factors: Results from HCAP studies (N=21,144)

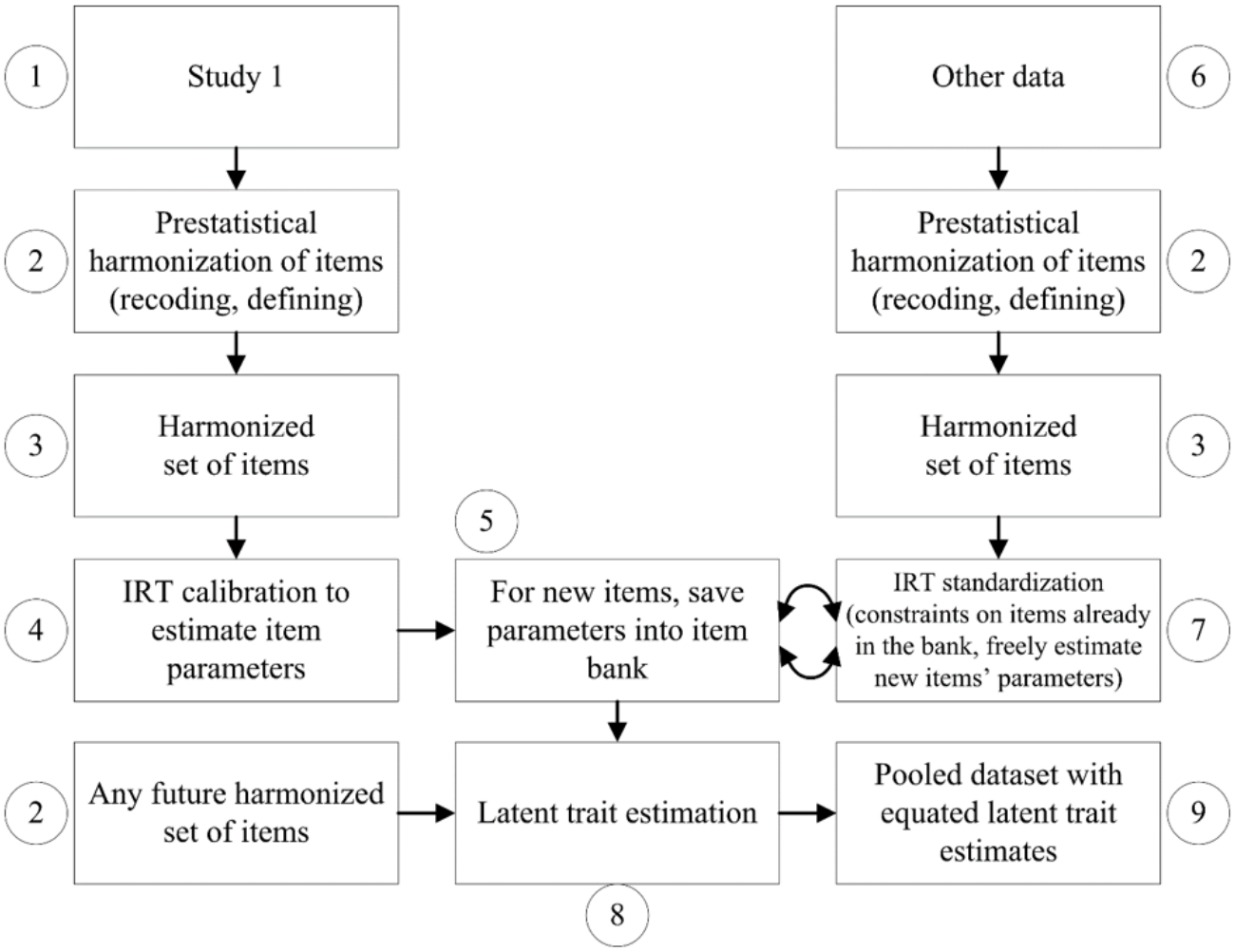
Cognitive domain	Covariate	Overall sample	HRS-HCAP	ELSA-HCAP	LASI-DAD	Mex-Cog	CHARLS-HCAP	HAALSI-HCAP
		Beta coefficient (SE)	Beta coefficient (SE)	Beta coefficient (SE)	Beta coefficient (SE)	Beta coefficient (SE)	Beta coefficient (SE)	Beta coefficient (SE)
Orientation								
	Female gender	-0.15* (-0.18, -0.13)	0.03 (-0.01, 0.07)	-0.01 (-0.09, 0.07)	-0.36* (-0.41, -0.31)	-0.37* (-0.45, -0.30)	-0.16* (-0.20, -0.13)	-0.33* (-0.46, -0.20)
	Age group							
	50-59 years	0.01 (-0.06, 0.08)	N/A	N/A	N/A	-0.01 (-0.12, 0.09)	N/A	0.28* (0.09, 0.46)
	60-69 years	REF	REF	REF	REF	REF	REF	REF
	70-79 years	-0.09* (-0.11, -0.06)	-0.05 (-0.11, 0.00)	-0.11* (-0.21, -0.01)	-0.19* (-0.25, -0.14)	-0.21* (-0.30, -0.12)	-0.10* (-0.14, -0.06)	-0.34* (-0.51, -0.16)
	80-89 years	-0.26* (-0.30, -0.23)	-0.25* (-0.31, -0.19)	-0.40* (-0.51, -0.29)	-0.42* (-0.51, -0.34)	-0.59* (-0.72, -0.46)	-0.24* (-0.31, -0.17)	-0.62* (-0.82, -0.43)
	90+ years	-0.64* (-0.72, -0.56)	-0.59* (-0.69, -0.49)	-0.87* (-1.08, -0.66)	-0.80* (-0.97, -0.62)	-1.19* (-1.47, -0.90)	-0.35* (-0.68, -0.02)	-0.79* (-1.13, -0.44)
	Education							
	No or Early Childhood Education	-0.93* (-0.96, -0.89)	-0.42* (-0.69, -0.15)	-0.64 (-1.43, 0.16)	-0.96* (-1.06, -0.87)	-0.69* (-0.80, -0.58)	-0.76* (-0.81, -0.71)	-0.70* (-1.02, -0.38)
	Primary education	-0.30* (-0.34, -0.26)	-0.44* (-0.56, -0.32)	N/A	-0.21* (-0.32, -0.10)	-0.16* (-0.28, -0.03)	-0.20* (-0.25, -0.14)	-0.15 (-0.47, 0.17)
	Lower secondary	REF	REF	REF	REF	REF	REF	REF
	Upper secondary	0.28* (0.24, 0.32)	0.23* (0.16, 0.29)	0.07 (-0.03, 0.17)	0.24* (0.13, 0.36)	0.04 (-0.20, 0.27)	0.11* (0.04, 0.19)	0.10 (-0.32, 0.51)
	Any college	0.40* (0.35, 0.45)	0.29* (0.22, 0.36)	0.15* (0.06, 0.24)	0.38* (0.23, 0.52)	0.27* (0.11, 0.42)	0.29* (0.14, 0.44)	0.19 (-0.35, 0.74)
Memory								
	Female gender	0.14* (0.12, 0.17)	0.32* (0.26, 0.37)	0.17* (0.07, 0.26)	0.08* (0.03, 0.12)	0.19* (0.13, 0.24)	0.02 (-0.01, 0.06)	-0.00 (-0.10, 0.09)
	Age group							
	50-59 years	0.11* (0.05, 0.18)	N/A	N/A	N/A	0.17* (0.09, 0.24)	N/A	0.24* (0.10, 0.38)
	60-69 years	REF	REF	REF	REF	REF	REF	REF
	70-79 years	-0.18* (-0.21, -0.16)	-0.24* (-0.32, -0.17)	-0.44* (-0.57, -0.31)	-0.27* (-0.32, -0.22)	-0.33* (-0.40, -0.27)	-0.23* (-0.26, -0.19)	-0.28* (-0.40, -0.15)
	80-89 years	-0.49* (-0.53, -0.46)	-0.72* (-0.80, -0.64)	-0.98* (-1.12, -0.84)	-0.61* (-0.68, -0.53)	-0.76* (-0.85, -0.66)	-0.39* (-0.46, -0.32)	-0.52* (-0.67, -0.38)
	90+ years	-0.95* (-1.03, -0.87)	-1.29* (-1.42, -1.16)	-1.46* (-1.73, -1.19)	-0.93* (-1.07, -0.78)	-1.17* (-1.38, -0.97)	-0.53* (-0.85, -0.21)	-0.67* (-0.93, -0.41)
	Education							
	No or Early Childhood Education	-0.92* (-0.95, -0.88)	-0.26 (-0.61, 0.09)	0.03 (-0.98, 1.03)	-0.62* (-0.70, -0.53)	-0.61* (-0.69, -0.54)	-0.89* (-0.94, -0.84)	-0.49* (-0.73, -0.25)
	Primary education	-0.30* (-0.34, -0.26)	-0.41* (-0.57, -0.25)	N/A	-0.15* (-0.25, -0.05)	-0.19* (-0.28, -0.10)	-0.25* (-0.30, -0.19)	-0.23 (-0.47, 0.01)
	Lower secondary	REF	REF	REF	REF	REF	REF	REF
	Upper secondary	0.37* (0.33, 0.41)	0.43* (0.35, 0.52)	0.26* (0.13, 0.39)	0.29* (0.19, 0.39)	0.05 (-0.12, 0.22)	0.17* (0.09, 0.24)	0.10 (-0.20, 0.41)
	Any college	0.68* (0.63, 0.73)	0.75* (0.66, 0.85)	0.48* (0.36, 0.59)	0.43* (0.30, 0.55)	0.33* (0.22, 0.45)	0.42* (0.27, 0.57)	0.38 (-0.03, 0.78)
Executive functioning								
	Female gender	-0.16* (-0.18, -0.13)	0.04 (-0.02, 0.09)	-0.04 (-0.14, 0.06)	-0.20* (-0.24, -0.16)	-0.19* (-0.26, -0.12)	-0.12* (-0.15, -0.09)	-0.16* (-0.25, -0.07)
	Age group							
	50-59 years	0.14* (0.07, 0.21)	N/A	N/A	N/A	0.19* (0.10, 0.28)	N/A	0.17* (0.04, 0.29)
	60-69 years	REF	REF	REF	REF	REF	REF	REF
	70-79 years	-0.23* (-0.26, -0.21)	-0.27* (-0.34, -0.20)	-0.47* (-0.60, -0.35)	-0.17* (-0.21, -0.13)	-0.41* (-0.49, -0.33)	-0.04* (-0.07, -0.01)	-0.29* (-0.41, -0.18)
	80-89 years	-0.72* (-0.75, -0.68)	-0.77* (-0.85, -0.70)	-1.02* (-1.16, -0.89)	-0.37* (-0.43, -0.30)	-0.82* (-0.94, -0.70)	-0.21* (-0.26, -0.16)	-0.51* (-0.64, -0.37)
	90+ years	-1.27* (-1.35, -1.19)	-1.25* (-1.38, -1.12)	-1.76* (-2.02, -1.50)	-0.66* (-0.79, -0.53)	-0.97* (-1.22, -0.73)	-0.38* (-0.66, -0.10)	-0.67* (-0.90, -0.43)
	Education							

	No or Early Childhood Education	-0.82* (-0.85, -0.78)	-0.88* (-1.22, -0.54)	-1.10* (-2.07, -0.13)	-0.85* (-0.93, -0.78)	-1.39* (-1.49, -1.30)	-0.44* (-0.47, -0.40)	-0.54* (-0.75, -0.32)
	Primary education	-0.16* (-0.20, -0.12)	-0.67* (-0.82, -0.51)	N/A	-0.21* (-0.30, -0.12)	-0.59* (-0.70, -0.49)	-0.15* (-0.19, -0.12)	-0.14 (-0.36, 0.08)
	Lower secondary	REF	REF	REF	REF	REF	REF	REF
	Upper secondary	0.24* (0.21, 0.28)	0.78* (0.70, 0.86)	0.43* (0.30, 0.55)	0.33* (0.25, 0.42)	0.18 (-0.02, 0.39)	0.18* (0.13, 0.22)	0.33* (0.05, 0.61)
	Any college	0.66* (0.61, 0.71)	1.22* (1.13, 1.31)	0.59* (0.47, 0.70)	0.64* (0.52, 0.75)	0.64* (0.50, 0.78)	0.43* (0.34, 0.53)	0.55* (0.18, 0.92)
Language								
	Female gender	0.03* (0.01, 0.05)	0.02 (-0.03, 0.07)	-0.08 (-0.19, 0.02)	-0.13* (-0.17, -0.09)	0.01 (-0.06, 0.07)	-0.04* (-0.07, -0.02)	-0.32* (-0.41, -0.23)
	Age group							
	50-59 years	0.51* (0.44, 0.57)	N/A	N/A	N/A	0.01 (-0.07, 0.09)	N/A	0.04 (-0.09, 0.17)
	60-69 years	REF	REF	REF	REF	REF	REF	REF
	70-79 years	-0.02 (-0.04, 0.00)	-0.08* (-0.14, -0.01)	-0.44* (-0.57, -0.30)	-0.15* (-0.20, -0.10)	-0.18* (-0.26, -0.11)	-0.10* (-0.13, -0.07)	-0.21* (-0.33, -0.09)
	80-89 years	-0.24* (-0.27, -0.21)	-0.43* (-0.49, -0.36)	-0.84* (-0.99, -0.70)	-0.36* (-0.43, -0.29)	-0.55* (-0.65, -0.45)	-0.36* (-0.41, -0.31)	-0.45* (-0.59, -0.32)
	90+ years	-0.53* (-0.60, -0.45)	-0.81* (-0.92, -0.70)	-1.27* (-1.54, -0.99)	-0.76* (-0.90, -0.61)	-0.89* (-1.11, -0.66)	-0.47* (-0.73, -0.21)	-0.52* (-0.76, -0.28)
	Education							
	No or Early Childhood Education	-0.75* (-0.78, -0.72)	-0.24 (-0.55, 0.06)	-0.74 (-1.79, 0.30)	-0.49* (-0.57, -0.41)	-0.77* (-0.85, -0.68)	-0.58* (-0.62, -0.54)	-0.53* (-0.75, -0.31)
	Primary education	-0.36* (-0.39, -0.32)	-0.29* (-0.43, -0.15)	N/A	-0.17* (-0.27, -0.08)	-0.28* (-0.38, -0.18)	-0.22* (-0.26, -0.17)	-0.15 (-0.37, 0.07)
	Lower secondary	REF	REF	REF	REF	REF	REF	REF
	Upper secondary	0.39* (0.36, 0.43)	0.41* (0.34, 0.49)	0.25* (0.12, 0.38)	0.15* (0.06, 0.25)	-0.12 (-0.31, 0.07)	0.29* (0.23, 0.35)	0.36* (0.08, 0.64)
	Any college	0.84* (0.79, 0.88)	0.72* (0.64, 0.80)	0.49* (0.37, 0.61)	0.23* (0.10, 0.35)	0.37* (0.24, 0.50)	0.65* (0.53, 0.76)	0.64* (0.27, 1.01)

Legend. Beta coefficients represent overall and study-specific differences in cognitive functioning between a given exposure grouping and the reference category. For age, persons aged 60-69 comprised the reference group. For education, persons with a lower secondary education comprised the reference group. Models for each exposure are mutually adjusted for other exposures in this table. N/A = Not applicable (e.g., no observations in the group)

* $p<0.05$

Supplementary Figure 1: Flowchart of item banking procedure implemented to statistically harmonize cognition across HCAP studies



Legend. This procedure was implemented separately for cognitive domains of orientation, memory, executive functioning, language, and general cognitive performance. Starting with a reference study, HRS-HCAP (step 1), pre-statistical harmonization was conducted by recoding and redefining cognitive test items as needed. Minor transformations were performed as needed to handle missing data and outlying values (step 2), resulting in a harmonized set of cognitive test items to support a measurement model (step 3). Step 4 entails calibration via item response theory methods (equivalent to confirmatory factor analysis, CFA) to freely estimate item parameters including factor loadings, and thresholds (for categorical test items) or intercepts (for continuous test items). Resulting item parameters are saved into an item bank (step 5). Next, additional studies are serially brought in (step 6) to have their cognitive test items recoded as necessary in the same manner as in other studies (step 2 at right), resulting in a unique set of harmonized cognitive test items for a study (step 3 at right). In step 7, item response theeory standardization is implemented with a CFA model that places constraints on parameters for items in common between the other study and HRS-HCAP as well as previous studies already processed, and freely estimates parameters for items unique to the new study. Parameters for these new items are iteratively added to the item bank (step 5), such that eventually any future harmonized set of items can be used to estimate latent traits (step 8) and save out factor scores into a pooled dataset for all included studies (step 9).

Supplemental Table 11. Number of participants with scores with salient DIF by HCAP study and cognitive domain, prior to DIF adjustment: Results from HCAP studies (N=21,144)

Domain		HRS-HCAP (N=3347)	ELSA-HCAP (N=1273)	LASI-DAD (N=1777)	Mex-Cog (N=2042)	CHARLS-HCAP (N=9755)	HAALSI-HCAP (N=631)
			n (%)	n (%)	n (%)	n (%)	n (%)
Memory		Ref	2 (0.2%)	1 (<0.1%)	105 (5.1%)	216 (2.2%)	0
Orientation		Ref	22 (1.7%)	290 (16.3%)	0	23 (0.2%)	326 (51.7%)
Language/fluency		Ref	57 (4.5%)	23 (0.6%)	50 (2.4%)	6668 (68.4%)	6 (0.9%)
Executive function		Ref	0	0	NA	NA	0

Legend. Salience of DIF was calculated as the difference between DIF-adjusted and non-DIF-adjusted factor scores. The number of participants whose DIF-adjusted scores differed by more than 0·3 SDs from non-DIF-adjusted scores are shown here. DIF=differential item functioning. HCAP=Harmonized Cognitive Assessment Protocol. HRS=US Health and Retirement Study. ELSA=English Longitudinal Study on Ageing. LASI-DAD=Longitudinal Aging Study in India-Diagnostic Assessment of Dementia. Mex-Cog=Mexican Health and Aging Study Cognitive Aging Ancillary Study. CHARLS=China Health and Retirement Longitudinal Study. HAALSI=Health and Aging in Africa: A Longitudinal Study in South Africa. NA=not applicable (no overlap).