Journal of Clinical Epidemiology 188 (2025) 111972

ORIGINAL RESEARCH

International application of an optimized harmonization approach for longitudinal cognitive data in people with HIV

Lang Lang[a,1], Leah H. Rubin[b,c,d,e,**,1], Beau M. Ances[f], Aggrey Anok[g], Sarah Cooley[f], Raha M. Dastgheyb[b], Rebecca E. Easter[b], Donald R. Franklin Jr.[h], Robert K. Heaton[h], Scott L. Letendre[h], Gertrude Nakijozi[g], Thomas Marcotte[h], Robert Paul[i], Eran F. Shorer[b], Stephan Tomusange[g], David E. Vance[j], Yanxun Xu[a,k,*,1]

[a]*Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, MD, USA*
[b]*Departments of Neurology, Johns Hopkins University School of Medicine, Baltimore, MD, USA*
[c]*Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine, Baltimore, MD, USA*
[d]*Department of Epidemiology, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA*
[e]*Department of Molecular and Comparative Pathobiology, Johns Hopkins University School of Medicine, Baltimore, MD, USA*
[f]*Department of Neurology, Washington University in Saint Louis, St. Louis, Missouri, USA*
[g]*Rakai Health Sciences Program, Kalisizo, Uganda*
[h]*HIV Neurobehavioral Research Program, Departments of Medicine and Psychiatry, University of California San Diego, San Diego, CA, USA*
[i]*Missouri Institute of Mental Health, University of Missouri-St. Louis, St. Louis, Missouri, USA*
[j]*School of Nursing, University of Alabama at Birmingham, Birmingham, AL, USA*
[k]*Division of Biostatistics and Bioinformatics at The Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, Baltimore, MD, USA*

Accepted 5 September 2025; Published online 11 September 2025

Funding: The contents of this publication are solely the responsibility of the authors and do not represent the official views of the NNTC or National Institutes of Health (NIH). The contents of this publication are solely the responsibility of the authors and do not represent the official views of the National Institutes of Health (NIH). MWCCS (Principal Investigators): Atlanta CRS (Ighovwerha Ofotokun, Anandi Sheth, and Gina Wingood), U01-HL146241; Baltimore CRS (Todd Brown and Joseph Margolick), U01-HL146201; Bronx CRS (Kathryn Anastos, David Hanna, and Anjali Sharma), U01-HL146204; Brooklyn CRS (Deborah Gustafson and Tracey Wilson), U01-HL146202; Data Analysis and Coordination Center (Gypsyamber D'Souza, Stephen Gange and Elizabeth Topper), U01-HL146193; Chicago-Cook County CRS (Mardge Cohen, Audrey French, and Ryan Ross), U01-HL146245; Chicago-Northwestern CRS (Steven Wolinsky, Frank Palella, and Valentina Stosor), U01-HL146240; Northern California CRS (Bradley Aouizerat, Jennifer Price, and Phyllis Tien), U01-HL146242; Los Angeles CRS (Roger Detels and Matthew Mimiaga), U01-HL146333; Metropolitan Washington CRS (Seble Kassaye and Daniel Merenstein), U01-HL146205; Miami CRS (Maria Alcaide, Margaret Fischl, and Deborah Jones), U01-HL146203; Pittsburgh CRS (Jeremy Martinson and Charles Rinaldo), U01-HL146208; UAB-MS CRS (Mirjam-Colette Kempf, James B. Brock, Emily Levitan, and Deborah Konkle-Parker), U01-HL146192; UNC CRS (M. Bradley Drummond and Michelle Floris-Moore), U01-HL146194. The MWCCS is funded primarily by the National Heart, Lung, and Blood Institute (NHLBI), with additional co-funding from the Eunice Kennedy Shriver National Institute Of Child Health & Human Development (NICHD), National Institute On Aging (NIA), National Institute Of Dental & Craniofacial Research (NIDCR), National Institute Of Allergy And Infectious Diseases (NIAID), National Institute Of Neurological Disorders And Stroke (NINDS), National Institute Of Mental Health (NIMH), National Institute On Drug Abuse (NIDA), National Institute Of Nursing Research (NINR), National Cancer Institute (NCI), National Institute on Alcohol Abuse and Alcoholism (NIAAA), National Institute on Deafness and Other Communication Disorders (NIDCD), National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), National Institute on Minority Health and Health Disparities (NIMHD), and in coordination and alignment with the research priorities of the National Institutes of Health, Office of AIDS Research (OAR). MWCCS data collection is also supported by UL1-TR000004 (UCSF CTSA), UL1-TR003098 (JHU ICTR), UL1-TR001881 (UCLA CTSI), P30-AI-050409 (Atlanta CFAR), P30-AI-073961 (Miami CFAR), P30-AI-050410 (UNC CFAR), P30-AI-027767 (UAB CFAR), P30-MH-116867 (Miami CHARM), UL1-TR001409 (DC CTSA), KL2-TR001432 (DC CTSA), and TL1-TR001431 (DC CTSA). This publication was also made possible by the NNTC project which is a funded contract mechanism supported by NIMH, NIDA, NIA, and NINDS: Manhattan HIV Brain Bank (MHBB): U24MH100931; Texas NeuroAIDS Research Center (TNRC): U24MH100930; National Neurological AIDS Bank (NNAB): U24MH100929; California NeuroAIDS Tissue Network (CNTN): U24MH100928; Data Coordinating Center (DCC): U24MH100925. This work was in part supported by the Johns Hopkins Center for the Advancement of HIV Neurotherapeutics (P30MH075773; Rubin, Slusher), CHARTER (N01 MH22005; Grant, HHSN271201000036 C; Grant, HHSN271201000030 C; Grant, R01 MH107345; Heaton, Letendre, R24 MH129166; Letendre, Ellis), the HIV Neurobehavioral Research Center (P30 MH062512; Moore, Ellis), R01 MH078748 (Marcotte), R01 MH073433 (Heaton), National Science Foundation grants 1918854 (Xu), R01 MH128085 (Xu), R01 MH099733 (Wawer, Sacktor), R01 MH119947 (Rubin, Paul), R01 NR015738 (Ances), R01 NR012907 (Ances), R01 NR014449 (Ances), R01 MH118031 (Ances), and a funded contract mechanism supported by NIMH (75N95023C00013 to Ances, Burdo, Letendre, Paul, Rubin).

[1] Equally contributing.

* Corresponding author. Department of Applied Mathematics and Statistics, Johns Hopkins University Whiting School of Engineering, 3400 North Charles Street, Wyman N429, Baltimore, MD 21218, USA.

** Corresponding author. Department of Neurology, Johns Hopkins University School of Medicine, 600 N. Wolfe Street/ Carnegie 3-301, Baltimore, MD 21287-7613, USA.

E-mail addresses: yanxun.xu@jhu.edu (L.H. Rubin); lrubin@jhmi.edu (Y. Xu).

https://doi.org/10.1016/j.jclinepi.2025.111972

Abstract

Objectives: We previously developed a refined longitudinal data harmonization method to address the challenge of nonoverlapping cognitive tests across cohorts, successfully harmonizing data from 5 large-scale US HIV studies. Building on this harmonized data set, we now aim to apply this method to an additional US HIV study and cognitive data from HIV studies in China, India, and Uganda. This effort will result in a more comprehensive data set with a larger, internationally diverse sample that includes both people with HIV and people without HIV.

Study Design and Setting: The new cohorts to be harmonized included cognitive tests that did not fully overlap across studies, a challenge for traditional harmonization methods. We applied our refined approach, designed for scenarios without direct test linkage. In the Uganda cohort, where a key method assumption was violated, we implemented targeted adjustments.

Results: The harmonized cognitive domain scores were consistent across cohorts and strongly correlated with raw or log-transformed cognitive test data (eg, timed outcomes). These scores preserved key patterns of variation observed in the raw data for key demographics—such as age, education, and race—and maintained age-related longitudinal trajectories of cognitive performance derived from all participants' visits.

Conclusion: The resulting harmonized data set includes 18,270 participants across multiple countries, significantly enhancing its diversity and utility. It lays the groundwork for developing normative data and conducting more robust analyses to address critical neuroHIV research questions. This study also demonstrates the adaptability of the refined harmonization method in integrating new data and accommodating methodological challenges.

Keywords: Cognition; Factor model; Harmonization; HIV; Psychometrics

Plain Language Summary

People with HIV (PWH) often face a variety of cognitive challenges, but these issues can look different for each person. As different studies use different tests to measure cognitive abilities, it is difficult to combine the results from multiple studies and draw clear conclusions. In our previous work, we developed a refined method to harmonize data from 5 large US-based HIV neuro studies. Such method could handle the scenarios where nonoverlapping cognitive tests are used in certain domains across different studies. We now aim to include additional cohorts from the United States, China, India, and Uganda. Because these new cohorts also use nonoverlapping cognitive tests in certain domains, we applied our developed approach to harmonize the new data into our previously harmonized data. Our refined method created ''harmonized scores'' for cognitive abilities that closely matched the original test results. These scores captured differences related to age, education, and other factors while preserving how each person's cognitive abilities changed over time. By using this method to combine new data with existing data, we were able to create a more comprehensive and diverse data set. This will aid researchers to better understand the wide range of cognitive changes in PWH, leading to stronger, more inclusive studies on the impact of HIV on cognition.

1. Introduction

People with HIV (PWH) often experience cognitive impairment due to a complex interplay of factors, including psychiatric and substance use comorbidities, coinfections, cardiometabolic disease, antiretroviral and other drugs, and social determinants of health. While specific subgroups of PWH—such as those who are virally suppressed or initiating ART—have been studied [1–6], no large-scale longitudinal cohort has yet captured the full range of cognitive outcomes across geographic regions and contributing factors.

To better understand cognitive complications in HIV, large-scale data integration across diverse groups is essential. However, cross-cultural differences in cognitive processes, such as attention and perception [7,8], may skew test performance, particularly when Western-developed assessments are applied in non-Western settings. Thus, methods that account for cultural variation and support harmonization are critically needed in global neuroHIV research.

To address the limited overlap in cognitive test batteries across 5 major US HIV studies (Women's Interagency HIV Study [WIHS], the Multicenter AIDS Cohort Study [MACS], the CNS HIV Antiretroviral Therapy Effects Research [CHARTER], the National NeuroAIDS Tissue Consortium [NNTC], and the HIV Neurobehavioral Research Program [HNRP]), we developed a refined

What is new?

Key findings

- The harmonized data set provides a key foundation for future research aiming to understand cognitive complications in people with HIV across diverse global populations.

What this adds to what is known?

- Applied a refined method from earlier work to include additional international cohorts in the harmonized data set.
- Refined the harmonization procedure to include participants with partially missing data.
- Showed how to adjust the method when key assumptions are violated in the data.

What is the implication and what should change now?

- Harmonized scores accurately capture variations based on age and education, preserving longitudinal cognitive trends.

harmonization approach [9]. Unlike traditional methods that rely on directly linking items for each domain to be harmonized [10], our approach uses a second-order factor model [11] to leverage structural relationships between cognitive domains. This enables harmonization even when tests do not directly overlap by modeling shared cognitive structure. It yields harmonized factor scores from raw data, enabling valid comparisons across cohorts. These scores demonstrated construct validity via strong correlations with original test scores, alignment with known demographic influences, and preservation of individual cognitive changes over time.

Building on this prior work, the present study expands our harmonization approach to one additional US cohort (Washington University in St. Louis) and 3 international cohorts: the Rakai Neurology Cohort Study (RNCS) in Uganda along with neuroHIV studies in China and India. Applying this method to global longitudinal datasets may support a scalable, culturally inclusive platform for neuroHIV research. Furthermore, we refined the harmonization procedure to incorporate partially complete data, in contrast to the complete-case approach used in our previous work.

2. Methods

2.1. Study participants

We harmonized cognitive data from 4 cohorts with longitudinal assessments (ranging from every 3 months to 2 years). Participants included in the analysis completed at least one cognitive test and provided sociodemographic data (age, biological sex, race/ethnicity, education).

1. Washington University Saint Louis (WUSTL) [3]. Data from January 2015 to April 2022 included 225 PWH, primarily men (79%).
2. RNCS [4,12,13]. Data from July 2013 to April 2024 included 913 participants (52.5% men) with 398 PWH and 515 people without HIV (PWoH).
3. Neurobehavioral Effects of HIV and Host Genetics in China Study. Data from January 2006 to December 2010 included 1013 participants (63.9% men), with 407 PWH and 606 PWoH.
4. NeuroAIDS in India Study. Data from October 2007 to May 2013 included 529 participants (58% men), with 252 PWH and 277 PWoH.

We also incorporated data from our previous study, which included WIHS, MACS, CHARTER, NNTC, and HNRP [9]. All cohorts are multisite US studies, except HNRP, which is based in San Diego, CA. Additional details are provided in the original publication and Supplemental Materials. While we acknowledge that registering secondary analyses is best practice for transparency and reproducibility, this specific study was not registered because it uses well-established data sets for which we lack permission to share.

2.2. Cognitive assessments

Table 1 outlines the cognitive tests and their corresponding outcomes. To handle non-Gaussian distributions, we log-transformed timed outcomes and reverse-coded these timed scores so that higher values represented better performance, and unified cutoff times across all cohorts.

2.3. Statistical analyses

Following the harmonization procedure detailed in our prior work [9], we began by extracting reference line data from each cohort. Subsequently, within each cohort, Exploratory Factor Analysis (EFA) was conducted on the cognitive tests to identify underlying cognitive domains, guided by predefined criteria (see Supplemental Section *EFA criteria*). The EFA results informed the specification of cohort-specific second-order linear factor models, which were then evaluated using Confirmatory Factor Analysis (CFA) on the reference line data. Model fit was assessed using the Comparative Fit Index (CFI), Root Mean Square Error of Approximation (RMSEA), and Standardized Root Mean Square Residual (SRMR), adhering to established criteria (good fit: CFI >0.95, RMSEA <0.05, SRMR <0.05; acceptable fit: RMSEA <0.08, SRMR <0.08) [14–16] before proceeding with harmonization.

We then performed fixed-parameter calibration [17]. Following the established procedure from our previous

Table 1. Cognitive tests administered in at least one of the harmonization countries.

Test (outcomes available)	US	China	India	Uganda
Hopkins verbal learning test-revised (total learning across trials 1-3, delayed free recall)	•	•	•	
Rey auditory verbal learning test (total learning across trials 1-5, immediate and delayed free recall)	•			
WHO - UCLA Auditory Verbal Learning Test (total learning across trials 1-6, recognition, immediate, and delayed free recall)				•
Stroop Color-Word Test-Comalli version (color naming, word reading, & color-word interference trials-time to complete [seconds])[a]	•			
Stroop Color-Word Test-Golden version (color naming, word reading, & color-word interference trials-stimuli read in 45 seconds)[a]	•	•	•	
D-KEFS Color-Word Interference Test (color naming, word reading, & inhibition trials-time to complete [seconds])[a]	•			
Trail Making Test: Part A and B (time to complete [seconds])[a]	•	Part A	Part A	
Color Trail 1 and 2 (time to complete [seconds])[a]		•	•	•
Grooved Pegboard Test (dominant and nondominant hand-time to complete [seconds])[a]	•	•	•	•
Symbol Digit Modalities Test (total correct)	•			•
Paced Auditory Serial Addition Task - 50-item (total correct)	•	•	•	
Controlled Oral Word Association Test (total correct)	•			
Animal Fluency (total correct)	•	•	•	•
Action Fluency (total correct)	•	•	•	

Dots indicate the test was administered within the cohort.
[a] Outcomes are log transformed and reverse scored so that higher equates to better performance for data harmonization.

work, we first identified reference visit data within each cohort. For each participant with at least one cognitive test, we selected their earliest visit with the most tests completed; these collectively defined the cohort-specific reference visit. Fixed-parameter calibration sequentially was then conducted sequentially across cohorts. For WIHS, MACS, CHARTER, NNTC, and HNRP, we used the same order as in our previous work, and for newly added cohorts, we proceeded in the order of WUSTL, China, India, and Uganda. In fixed-parameter calibration, item parameters (factor loadings, intercepts, and residual variances) estimated in earlier cohorts were held fixed while calibrating later cohorts, thereby placing all cohorts on a common measurement scale. For WIHS, the general factor mean was fixed at 0 for model identification; for all subsequent cohorts, first-order domain means were constrained to 0, and the general factor mean was freely estimated. The second-order factor model and harmonization assumptions are critical for linking nonoverlapping domains (see Supplemental Section *How second-order factor models link non-overlapping domains and assumptions*). The rationale for the ordering of new cohorts and the potential impact of harmonization order are provided in the Supplemental Section *Harmonization order* along with supporting sensitivity analyses. We perform Differential Item Functioning (DIF) analyses [18,19] when significant model assumption violations occur to assess whether these violations result from cross-cultural differences in the anchor items.

After harmonizing the reference line data, we proceed to harmonize the longitudinal data. This step assumes longitudinal measurement invariance—that the factor structure, item loadings, and intercepts of the cognitive battery remain stable over time. This assumption is supported by consistent administration procedures, item content, and scoring rules across time points and cohorts. For all cohorts except Uganda, the only exception involved the use of psychometrically equivalent alternate forms of verbal learning tests (eg, HVLT-R), which were alternated to mitigate practice effects. As noted in prior research [20], alternate forms can vary in difficulty, potentially jeopardizing longitudinal measurement invariance and the validity of derived scores. Given limited metadata on form versions, we evaluated equivalence and applied equipercentile equating [20] as a preharmonization step to reduce discrepancies in WIHS and MACS (see Supplemental Section *Alternating Forms*). We then applied the derived parameters to all longitudinal

Table 2. Cognitive tests administered in at least one of the harmonization countries.

Test (outcomes available)	US	China	India	Uganda
Hopkins verbal learning test-revised (total learning across trials 1-3, delayed free recall)	•	•	•	
Rey auditory verbal learning test (total learning across trials 1-5, immediate and delayed free recall)	•			
WHO - UCLA Auditory Verbal Learning Test (total learning across trials 1-6, recognition, immediate, and delayed free recall)				•
Stroop Color-Word Test-Comalli version (color naming, word reading, & color-word interference trials-time to complete [seconds])[a]	•			
Stroop Color-Word Test-Golden version (color naming, word reading, & color-word interference trials-stimuli read in 45 seconds)[a]	•	•	•	
D-KEFS Color-Word Interference Test (color naming, word reading, & inhibition trials-time to complete [seconds])[a]	•			
Trail Making Test: Part A and B (time to complete [seconds])[a]	•	Part A	Part A	
Color Trail 1 and 2 (time to complete [seconds])[a]		•	•	•
Grooved Pegboard Test (dominant and nondominant hand-time to complete [seconds])[a]	•	•	•	•
Symbol Digit Modalities Test (total correct)	•			•
Paced Auditory Serial Addition Task - 50-item (total correct)	•	•	•	
Controlled Oral Word Association Test (total correct)	•			
Animal Fluency (total correct)	•	•	•	•
Action Fluency (total correct)	•	•	•	

The entries under each cohort represent the mean and SD of each domain factor in that cohort. Analysis of variance was used to compare cohorts on the domain factors. Post-hoc ordering is obtained through pairwise two-sample *t*-test between cohorts and Bonferroni correction.

a indicates United States, b indicates China, c indicates India, and d indicates Uganda. Dots indicate the test was administered within the cohort.

[a] Outcomes are log transformed and reverse scored so that higher equates to better performance for data harmonization.

data, yielding harmonized factor scores for every cognitive domain at each participant visit. These scores provide valid cross-cohort comparisons even when raw test scores differ. Unlike our previous work, we retained all visits with at least one cognitive test score, as participants with lower function—often the very group of greatest interest—😕 commonly skip portions of assessments, producing partial but informative data.

See Supplemental Figure 17 for a flow chart summarizing the harmonization process. EFA was performed using the psych package [21] in R [22]. More complex factor models were estimated using Mplus version 8.3 [23] with Full Information Maximum Likelihood (FIML) to handle partially missing data. Mplus code for the reference line and longitudinal harmonization procedures is available at https://github.com/BHPDataSci/DataHarmonization. Reporting follows the STROBE guidelines (see Supplemental Section: "STROBE Statement" [24]).

3. Results

3.1. Cohort demographics

Table 2 summarizes the demographic characteristics of participants' reference visits by country. Using US Census Bureau divisions, the US sample was geographically distributed as follows: 18% from the Northeast, 13% Midwest, 21% South, and 48% West. The combined data set included 18,270 participants (11,911 PWH; 6359 PWoH), aged 18 to 99 years, with 34% women, 35% Black, 14.3% Hispanic, and 70.5% having $\geq$12 years of education. Participant characteristics varied across cohorts in age, education, race/ethnicity, proportion of PWH, and number of completed visits.

3.2. First-order factor structure obtained through EFA

Figure 1 illustrates the first-order factor structure for each cohort identified through EFA (Supplemental Tables 1-4).

The structure aligns closely with established cognitive domains and demonstrates overall good performance in the first-order CFA (CFI $\geq$ 0.95, RMSEA ~ 0.06, except for China ~0.08, SRMR $\leq$0.05, except for WUSTL = 0.051, see Supplemental Table 5 for fit statistics). Test abbreviations are listed in the table accompanying Figure 2 and are used throughout the remainder of this paper. The 5 factors were defined as follows: *Factor 1-Declarative Memory* (verbal learning and memory test outcomes), *Factor 2-Verbal Processing Speed* (Stroop), *Factor 3-Attention/Processing Speed* (Trail Making Test, Color Test, Paced Auditory Serial Addition Test), *Factor 4-Motor Function* (Grooved Pegboard Test), and *Factor 5-Language* (Animal, Letter, and Action Fluency). Based on the EFA results and predefined criteria (Supplemental Section *EFA criteria*), RAVLT Trial VI was excluded from further harmonization due to a factor loading below the cutoff (0.379 $<$ 0.4). In addition, PASAT was excluded for the India cohort due to an inconsistent factor structure: its highest loading was associated with *Factor 2-Verbal Processing Speed*, whereas in other cohorts, the highest loading was associated with *Factor 3-Attention/Processing Speed*.

3.3. Diagnostic statistics showed good fit performance of the second-order factor model to the reference line data

The model fit results for the included cohorts indicate that our models demonstrated acceptable fit performance, with CFI values ~0.95, RMSEA ~$<$ 0.08, and SRMR $<$ 0.07. Details are provided in Supplemental Figures 1-9 and Supplemental Table 6.

3.4. Adjustment in RNCS reference line sample

During harmonization, we encountered model fitting issues when applying parameter estimates from prior cohorts to the RNCS reference line sample. Further analysis revealed substantial discrepancies between domains—⊜ specifically between Motor Function and Attention/Processing Speed domains—in the RNCS data. Specifically, we compared representative Motor Function tests (GPEGDOM, GPEGNDOM) and an Attention/Processing Speed test (SYDM) in the RNCS reference line sample with WIHS and MACS (see Supplemental Table 7 (a) and Supplemental Figure 10 (a)). The RNCS means for GPEGDOM and GPEGNDOM were −1.89 and −1.98, closely aligning with WIHS values of −1.93 and −1.98. However, the SYDM mean in RNCS was 17.21—more than 2 SDs lower than the WIHS mean of 42.82. A similar pattern emerged when comparing COLOR1 and COLOR2 (Attention/Processing Speed) with GPEGDOM and GPEGNDOM in the China and Uganda cohorts (see Supplemental Table 7 (b) and Supplemental Figure 10 (b)). These discrepancies indicated a violation of the mean-structure assumption, primarily due to DIF—that is, instances in which individuals

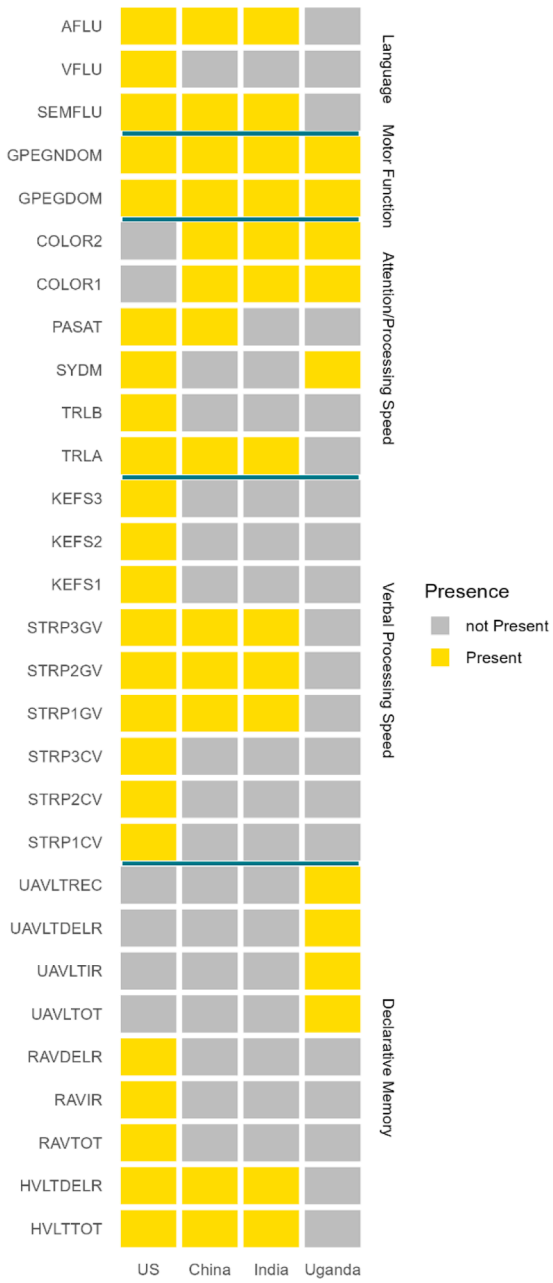

Figure 1. Cognitive tests included in the data harmonization process. Overview of test administration across countries and corresponding exploratory factor analysis (EFA) results. Yellow cells indicate that a given test was administered in the specified country, whereas gray cells denote its absence. Green divider lines group tests by cognitive domain, visually separating each domain and its associated measures. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

from different groups, despite equivalent underlying ability systematically differ in observed scores [18]. DIF is well documented in cross-cultural settings [19], often arising from differences in language, education, or culturally specific content that alter item difficulty or interpretation. To account for this, we conducted DIF analysis for 5 Uganda-based tests (GPEGDOM, GPEGNDOM, SYDM,

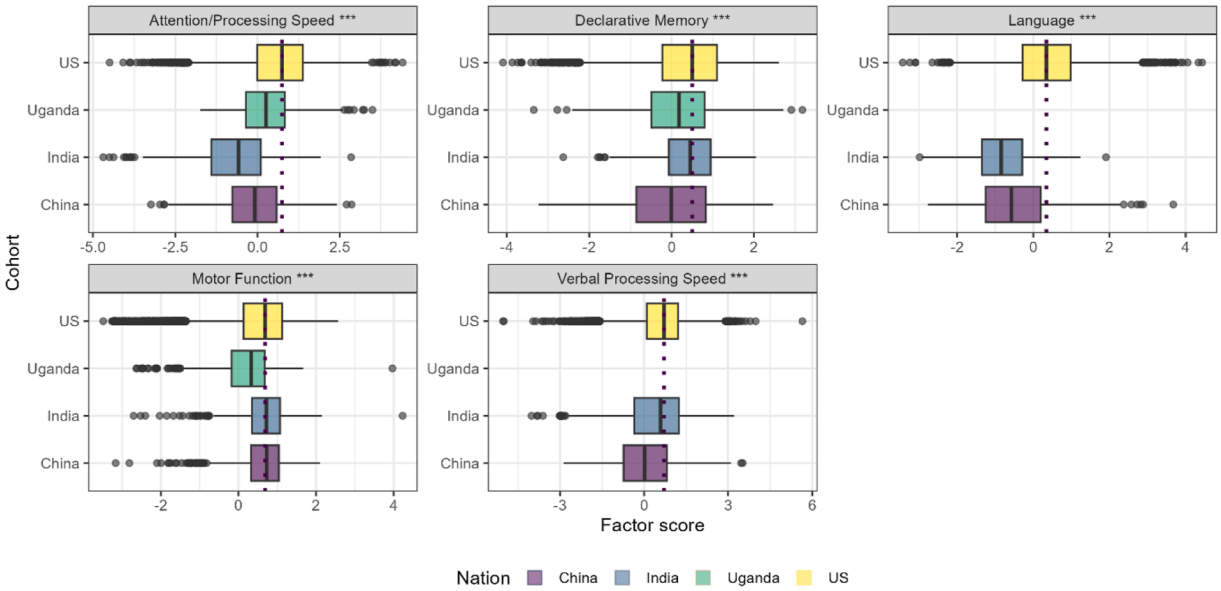

Figure 2. Harmonized factor scores for the 4 countries. ***$P < .01$ for ANOVA test.

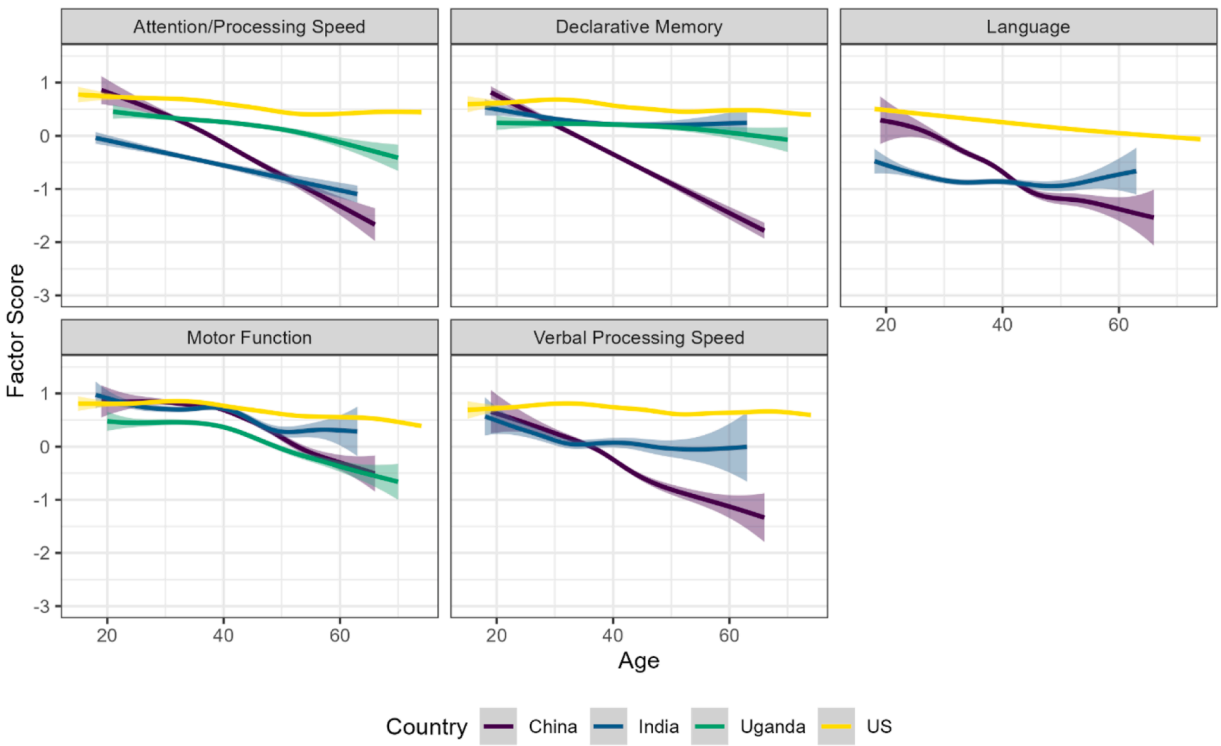

Figure 3. GAM-smoothed lines for longitudinal harmonized factor scores of participants from the 4 countries against age. The colored band is the associated 95% confidence band. The plot displays trends in harmonized scores across age based on data from all individuals with varying follow-up durations.

Table 3. Summary for harmonized domain scores across cohorts at the reference line visit

Characteristics	US (N = 15,815)	China (N = 1013)	India, (N = 529)	Uganda, (N = 913)	P-value	Post-hoc
Declarative memory	0.311 (1.044)	−0.173 (1.166)	0.302 (0.75)	0.102 (0.957)	<0.01	a,c>d>b
Verbal processing speed	0.496 (1.001)	−0.021 (1.176)	0.221 (1.355)	*n*.a	<0.01	a>c>b
Attention/processing speed	0.562 (1.143)	−0.299 (1.08)	−0.992 (1.323)	0.197 (0.9)	<0.01	a>d>b>c
Motor function	0.494 (0.925)	0.562 (0.695)	0.58 (0.722)	0.097 (0.762)	<0.01	c,b>a>d
Language	0.322 (0.998)	−0.593 (1.065)	−0.952 (0.789)	*n*.a	<0.01	a>b>c

The entries under each cohort represent the mean and SD of each domain factor in that cohort. Analysis of variance was used to compare cohorts on the domain factors. Post-hoc ordering is obtained through pairwise two-sample *t*-test between cohorts and Bonferroni correction.

a indicates United States, b indicates China, c indicates India, and d indicates Uganda.

Dots indicate the test was administered within the cohort.

COLOR1, and COLOR2) using Multi-Group CFA (MG-CFA) [25−27], applying ΔCFI <0.01 as the threshold for invariance [28]. Details of the method and results are presented in Supplemental Section *DIF analysis*.

DIF results indicated that GPEGDOM and GPEGNDOM demonstrated full measurement invariance (factor loadings, intercepts, residual variances) relative to the WIHS, the reference cohort in which these parameters were originally estimated. In contrast, SYDM, COLOR1, and COLOR2 exhibited only partial invariance: factor loadings were invariant when compared with their source cohorts. Therefore, when harmonizing the Uganda cohort, we fix all parameters for GPEGDOM and GPEGNDOM based on estimates from preceding cohorts. For SYDM, COLOR1, and COLOR2, we fixed factor loadings but allowed intercepts and residual variances to be freely estimated.

3.5. Raw cognitive test scores associated with harmonized factor scores

We validated our harmonization procedure by assessing the correlations between the original test scores (with timed tests log-transformed and reverse scored) and the resulting harmonized scores within each cohort. The analysis revealed strong alignment, as all correlations were statistically significant ($P < .001$) and typically surpassed a value of 0.6 (details in Supplemental Table 8 and Supplemental Figure 11).

3.6. Harmonized scores maintain expected differences across demographic variables

We also investigated the relationships between key sociodemographic variables (education, age, sex, race/ethnicity, HIV status) and both the original raw scores and the derived harmonized scores. We found that the nature of these associations was preserved through harmonization, with both types of scores exhibiting similar trends and significant differences across demographic groups (Supplemental Tables 9-18; Supplemental Figures 12-15).

3.7. Harmonized scores preserve the overall age-related longitudinal trajectory

We evaluated age-related cognitive trajectories using GAM-smoothed curves of harmonized scores and, separately, raw test scores, using all available longitudinal observations and grouping curves by country. As shown in Figure 3, the harmonized scores reproduce the expected age-related decline across 5 cognitive domains in all 4 countries, closely mirroring patterns observed in the raw data (Supplemental Figures 16−20). These results support the validity of our harmonization approach.

3.8. Distributions of harmonized cognition factor scores by cohorts

We visualized the harmonized reference line scores by cognitive domain across countries using box plots in Figure 2. Detailed summary statistics, presented in Table 3, indicate that participants from the United States and India demonstrated superior performance across most of cognitive factors. Item parameters from the harmonization process are provided in Supplemental Table 19.

4. Discussion

We applied our previously developed harmonization method to integrate cognitive data from an additional US cohort and 3 international longitudinal HIV studies, which used different tests in some domains. In this work, we also refined the procedure to incorporate participants with partially missing cognitive data, thereby increasing sample inclusiveness and reducing potential selection bias. The resulting harmonized scores aligned with demographic patterns, showed strong correlations with raw/log-transformed test scores, and preserved expected demographic differentiation and longitudinal trajectories. These findings confirm the method's effectiveness and flexibility, even when harmonization assumptions are not fully met.

In evaluating validity, it is important to recognize the limitations of the evidence presented. While strong correlations between raw scores and harmonized factor scores (see Section 3.5) are expected in a factor analytic framework that summarizes shared variance, they provide only modest support for construct validity. Demographic comparisons provide somewhat stronger evidence, as harmonized scores showed expected associations with age, sex, and education. Nonetheless, more robust validation approaches are needed. Future efforts should prioritize assessments of predictive validity (eg, associations with longitudinal clinical outcomes), convergent validity (eg, correlations with neuroimaging/biomarker data), and external validity in diverse settings. These steps will be essential for establishing the broader utility of harmonized cognitive scores in global health research.

In our cross-cultural harmonization, violations of the mean-structure assumption appear to stem primarily from DIF. DIF may arise from cultural or linguistic differences, differences in test administration, or participant characteristics such as educational attainment. Notably, in the RNCS cohort, only 7.1% of participants completed high school—the lowest proportion among all cohorts—☹ suggesting that observed DIF in this group may reflect education-related effects [29,30] (additional discussion on the possible cause of the observed discrepancy is in Supplemental Section *More discussion on discrepancies observed in Uganda cohort*). Guided by MG-CFA diagnostics, we addressed DIF by freeing the intercepts and residual variances for the 3 Attention/Processing Speed (ATT/PS) tests in the Uganda cohort. As a result, the Motor factor—anchored by the Grooved Pegboard measures—☹ now serves as the sole linkage between Uganda and the common scale. This limitation should be considered in interpreting findings involving the Uganda cohort. While the current DIF analyses specifically addressed anchor items whose misfit could directly affect the mean-structure assumptions, we acknowledge that subtle DIF may persist in other items. Future work should concentrate on developing a more comprehensive framework to evaluate and address DIF.

To provide readers with more insights on the various analytic options, we applied the Linear Linking for Related Traits (LLRT) method [31], which is another method designed for harmonization when some domains lack overlapping items. A comparison of LLRT-derived scores for WIHS and MACS with those from our approach is provided in the Supplemental Section *Comparison with LLRT*.

This study advances our harmonization framework in 4 key ways. First, it demonstrates the methods applicability to international cohorts with nonoverlapping test batteries (eg, Declarative Memory in Uganda). Second, we introduce diagnostic-adjustment strategy to address model violations and harmonization failures, providing a practical guide for real-word implementation. Third, we refined the procedure to incorporate participants with partially complete cognitive data, increasing effective sample size and reducing selection bias. Fourth, by successfully harmonizing data across culturally diverse cohorts, we created the largest integrated cognitive data set in PWH to date, substantially expanding opportunities for cross-cultural and longitudinal research that were previously limited by methodological heterogeneity.

CRediT authorship contribution statement

Lang Lang: Writing − review & editing, Writing − original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Leah H. Rubin:** Writing − review & editing, Writing − original draft, Supervision, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization. **Beau M. Ances:** Writing − review & editing, Resources, Funding acquisition, Data curation. **Aggrey Anok:** Writing − review & editing, Data curation. **Sarah Cooley:** Writing − review & editing, Data curation. **Raha M. Dastgheyb:** Writing − review & editing, Visualization, Data curation. **Rebecca E. Easter:** Writing − review & editing, Writing − original draft. **Donald R. Franklin:** Writing − review & editing, Data curation. **Robert K. Heaton:** Writing − review & editing, Funding acquisition, Data curation. **Scott L. Letendre:** Writing − review & editing, Funding acquisition. **Gertrude Nakijozi:** Writing − original draft, Resources. **Thomas Marcotte:** Writing − review & editing, Funding acquisition. **Robert Paul:** Writing − review & editing, Funding acquisition. **Eran F. Shorer:** Writing − review & editing, Writing − original draft. **Stephan Tomusange:** Writing − review & editing, Data curation. **David E. Vance:** Writing − review & editing. **Yanxun Xu:** Writing − review & editing, Writing − original draft, Supervision, Methodology, Investigation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.

Acknowledgments

The authors gratefully acknowledge the contributions of Dr. Manisha Ghate from the National AIDS Research Institute, MIDC, Bhosari, Pune, India, and Dr. Chuan Shi from the Institute of Mental Health, National Clinical Research Center for Mental Disorders, Peking Univ. Sixth Hospital, Key Laboratory of Mental Health, Ministry of Health, Peking University, Beijing, China as well as all of the study participants for all studies (MWCCS, CHARTER, NNTC, HNRP, WUSTL, RNCS, Neurobehavioral Effects of HIV

and Host Genetics in China study, NeuroAIDS in India study) and dedication of the staff at the sites.

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jclinepi.2025.111972.

Data availability

Data will be made available on request.

References

[1] Dastgheyb RM, Buchholz AS, Fitzgerald KC, Xu Y, Williams DW, Springer G, et al. Patterns and predictors of cognitive function among virally suppressed women with HIV. Front Neurol 2021;12:604984.
[2] Dastgheyb RM, Sacktor N, Franklin D, Letendre S, Marcotte T, Heaton R, et al. Cognitive trajectory phenotypes in human immunodeficiency virus-infected patients. J Acquired Immune Deficiency Syndromes 2019;82:61−70. 1999.
[3] Paul RH, Cho K, Belden A, Carrico AW, Martin E, Bolzenius J, et al. Cognitive phenotypes of HIV defined using a novel data-driven approach. J Neuroimmune Pharmacol 2022;17:515−25.
[4] Rubin LH, Saylor D, Nakigozi G, Nakasujja N, Robertson K, Kisakye A, et al. Heterogeneity in neurocognitive change trajectories among people with HIV starting antiretroviral therapy in Rakai, Uganda. J Neurovirol 2019;25:800−13.
[5] Rubin LH, Sundermann EE, Dastgheyb R, Buchholz AS, Pasipanodya E, Heaton RK, et al. Sex differences in the patterns and predictors of cognitive function in HIV. Front Neurol 2020;11:551921.
[6] Sundermann EE, Dastgheyb R, Moore DJ, Buchholz AS, Bondi MW, Ellis RJ, et al. Identifying and distinguishing cognitive profiles among virally suppressed people with HIV. Neuropsychology 2024;38:169−83.
[7] Chua HF, Boland JE, Nisbett RE. Cultural variation in eye movements during scene perception. Proc Natl Acad Sci USA 2005;102:12629−33.
[8] Goh JOS, Leshikar ED, Sutton BP, Tan JC, Sim SKY, Hebrank AC, et al. Culture differences in neural processing of faces and houses in the ventral visual cortex. Social Cogn Affective Neurosci 2010;5:227−35.
[9] Lang L, Rubin LH, Dastgheyb RM, Vance DE, Letendre SL, Franklin DR, et al. Development of a refined harmonization approach for longitudinal cognitive data in people with HIV. J Clin Epidemiol 2025;178:111620.
[10] Kolen MJ, Brennan RL. Test equating, scaling, and linking: Methods and practices. 3rd ed. New York, NY: Springer Science + Business Media; 2014.
[11] Brunner M, Nagy G, Wilhelm O. A tutorial on hierarchically structured constructs. J Personal 2012;80:796−846.
[12] Sacktor N, Saylor D, Nakigozi G, Nakasujja N, Robertson K, Grabowski MK, et al. Effect of HIV subtype and antiretroviral therapy on HIV-associated neurocognitive disorder stage in Rakai, Uganda. J Acquir Immune Defic Syndr 2019;81:216−23.
[13] Vecchio A, Robertson K, Saylor D, Nakigozi G, Nakasujja N, Kisakye A, et al. Neurocognitive effects of antiretroviral initiation among people living with HIV in rural Uganda. J Acquired Immune Deficiency Syndromes 2020;84:534−42.
[14] Hooper D, Coughlan J, Mullen M. Structural equation modeling: guidelines for determining model fit. Electron J Business Res Methods 2007;6:53−60.
[15] Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Struct Equation Model A Multidisciplinary J 1999;6:1−55.
[16] Schermelleh-Engel K, Moosbrugger H, Müller H. Evaluating the fit of structural equation models: tests of significance and descriptive goodness-of-fit measures. Methods Psychol Res 2003;8:23−74.
[17] von Davier M, von Davier AA. A unified approach to irt scale linking and scale transformations. ETS Res Rep Ser 2004;2004:i-21.
[18] Meredith W. Measurement invariance, factor analysis and factorial invariance. Psychometrika 1993;58:525−43.
[19] van de Vijver F, Tanzer NK. Bias and equivalence in cross-cultural assessment: an overview. Eur Rev Appl Psychol 2004;54:119−35.
[20] Gross AL, Inouye SK, Rebok GW, Brandt J, Crane PK, Parisi JM, et al. Parallel but not equivalent: challenges and solutions for repeated assessment of cognition over time. J Clin Exp Neuropsychol 2012;34:758−72.
[21] Revelle W. psych: Procedures for Psychological, Psychometric, and Personality Research. Evanston, IL: Evanston; 2024.
[22] R Core Team. R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing; 2023.
[23] Muthén LK, Muthén BO. Mplus: Statistical Analysis with Latent Variables: User's Guide. Los Angeles, CA: Muthén & Muthén; 2017.
[24] Cuschieri S. The STROBE guidelines. Saudi J Anaesth 2019;13:S31−4.
[25] Vandenberg RJ, Lance CE. A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. Organizational Res Methods 2000;3:4−70.
[26] Hirschfeld G, Von Brachel R. Improving Multiple-Group confirmatory factor analysis in R−A tutorial in measurement invariance with continuous and ordinal indicators. Pract Assess Res Eval 2014;19:7.
[27] Brown TA. Confirmatory factor analysis for applied research. New York, NY: Guilford publications; 2015.
[28] Cheung GW, Rensvold RB. Evaluating goodness-of-fit indexes for testing measurement invariance. Struct Equation Model 2002;9:233−55.
[29] Holland PW, Wainer H. Differential item functioning. New York, NY: Routledge; 2012.
[30] Scott NW, Fayers PM, Aaronson NK, Bottomley A, de Graeff A, Groenvold M, et al. Differential item functioning (DIF) analyses of health-related quality of life instruments using logistic regression. Health Qual Life Outcomes 2010;8:1−9.
[31] Nichols EL, Cadar D, Lee J, Jones RN, Gross AL. Linear linking for related traits (LLRT): a novel method for the harmonization of cognitive domains with no or few common items. Methods 2022;204:179−88.