Appendix A

Table of Contents

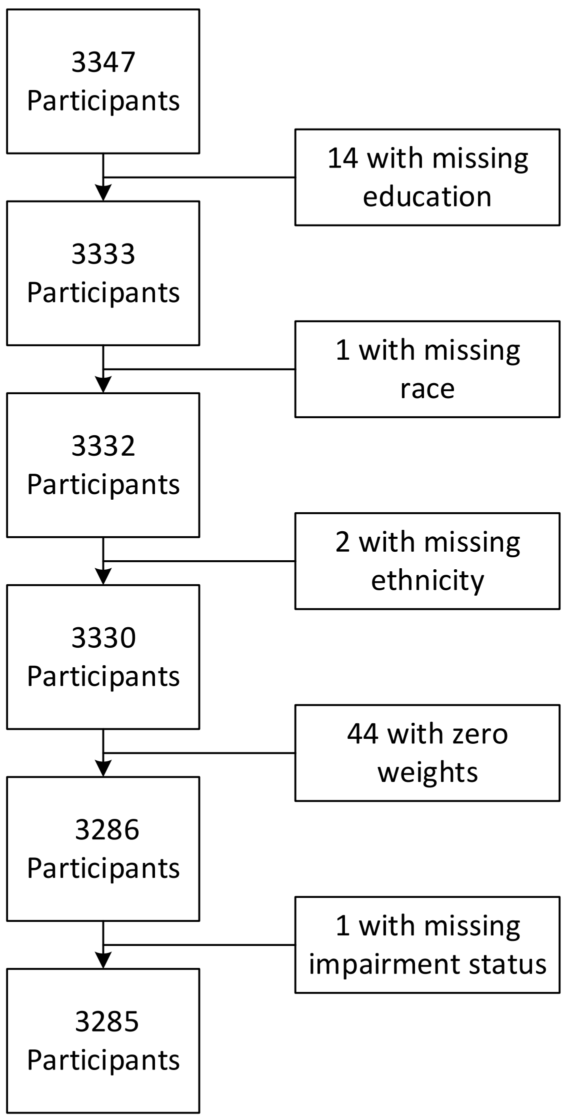
Sample selection and survey weights

The parent Health and Retirement Survey (HRS) International Partner Studies (IPS) were nationally representative, with the exception of the South African study, which was representative of the rural Agincourt sub-district of the Mpumalanga Province in South Africa [1]. The HCAP studies in the US and Mexico randomly sampled eligible participants from the parent HRS IPS surveys; all other locations oversampled those with low levels of cognition for inclusion in the HCAP samples to ensure that enough individuals with cognitive impairment and dementia would be selected. Survey weights were used in analyses to ensure that results are generalizable to broader populations, and therefore can better inform future survey design. In the United States and Mexico, study participants were randomly sampled from the prior HRS IPS studies. Therefore, in this analysis we used the survey weights provided in these broader HRS IPS studies. In the United States, additional data was provided on individuals who were selected to be in the HCAP sample but declined participation. We accounted for potential selection bias by calculating stabilized inverse probability of selection weights and multiplying these weights by the provided survey weights. We used a logistic regression models with predictor variables for gender and 5-year age group to predict selection. The England, India, and South Africa HCAP samples over-selected individuals with low cognition, and provided survey weights to allow for the generalization of results to the broader samples. For England, and India, survey results enable generalization to nationally-representative samples. Survey weights for South Africa enable generalization to the population of the Agincourt sub-district of the Mpumalanga Province in South Africa.

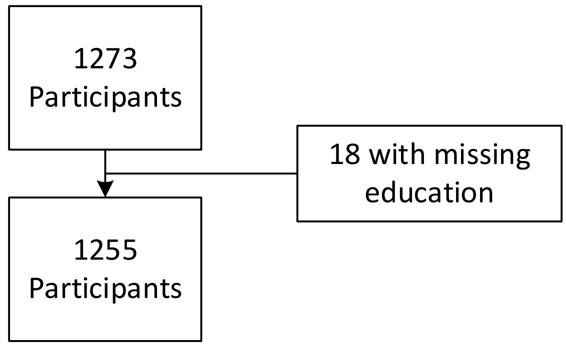
Sample exclusions

For each of the HCAP samples, we included individuals with valid responses on the variables required to ascertain cognitive impairment status. For each of the samples, we required non-missing values for age, sex, and education. Additionally, we required non-missing values for race and ethnicity in the United States, rurality in Mexico and India, and illiteracy status in India. In the United States and Mexico, a few participants had estimated survey weights of 0 for the HRS IPS survey wave from which HCAP participants were selected. As we used these survey weights to ensure results would be applicable to the general population, these individuals were excluded from our analyses. Finally, we excluded all individuals with missingness on cognitive impairment status due to high missingness in cognitive items and low reliability of scores across all cognitive domains. The flow charts for sample exclusions for each one of the HCAP samples are shown below.

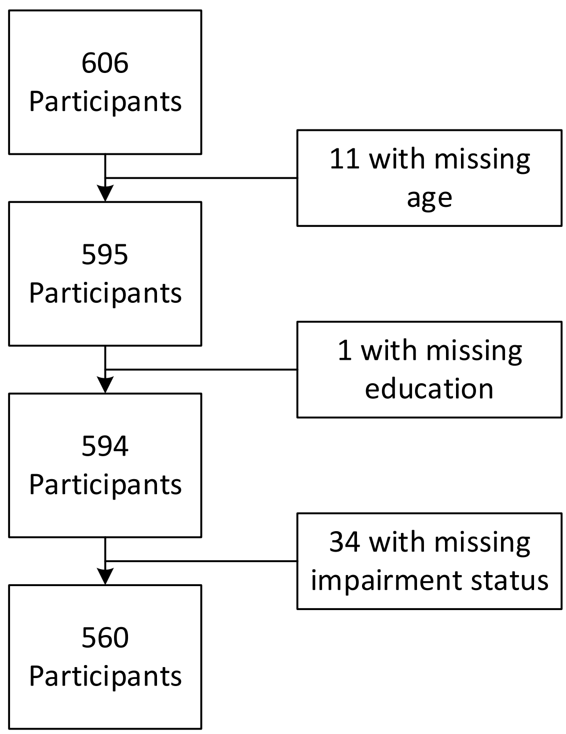
United States



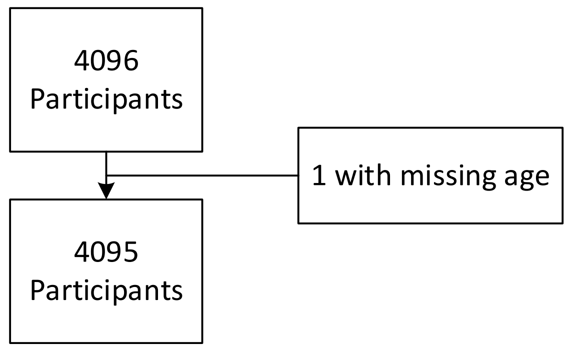
England



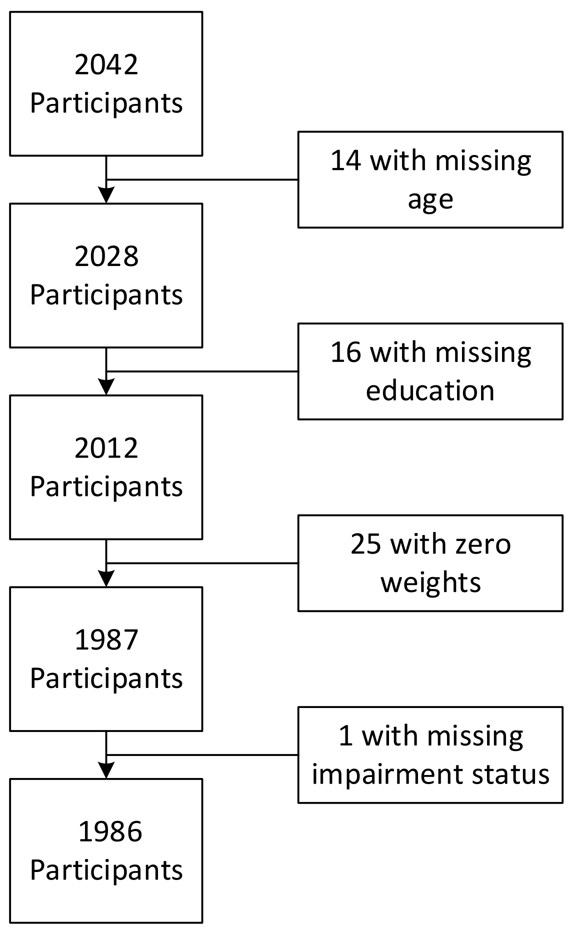
South Africa



India



Mexico



Definitions of sociodemographic variables

Race and ethnicity (US): We categorized race into white, black, and other, due to small numbers of other racial categories, and categorized ethnicity into Hispanic and non-Hispanic.

Rurality (Mexico and India): In the Mexico HCAP study, rurality was categorized into four levels: (1) population greater than 100,000; (2) population 15,000-99,999; (3) population 2,500-14,999; and (4) population <2,500. We categorized individuals as living in rural areas if they lived in an area with a population of 14,999 people or fewer. The categorization of urban versus rural in India was based on recorded information in census data.

Literacy (India and South Africa): In India and South Africa we also considered literacy as an important demographic variable.

Education (all countries): We dichotomized educational attainment based on the distribution of educational attainment in each study. In the United States and England, individuals were grouped into those with secondary education or lower versus those with post-secondary education. In South Africa, India and Mexico, we grouped individuals into those with primary school education and lower, versus those with secondary education or higher.

Depressive symptomology (all countries): We considered all items administered from the Center for Epidemiologic Studies – Depression scale (CESD). Table S1 shows the items administered across the HCAP studies. We classified individuals with the highest 10% of self-reported symptoms in each setting as having high depressive symptomology. All HCAP studies except Mexico included items from the CESD. For Mexico, we considered responses to the CESD from the prior HRS IPS wave.

Table S1. Items from the Center for Epidemiologic Studies – Depression scale (CESD) administered in each of the United States, England, South Africa, India, and Mexico Harmonized Cognitive Assessment Protocol (HCAP) samples

CESD Item	United States	England	South Africa	India	Mexico
Depressed	X	X	X	X	X
Effortful	X	X	X	X	X
Restless Sleep	X	X	X		X
Felt Happy	X	X	X	X	X
Felt Lonely	X	X	X	X	X
Enjoyed Life	X	X	X	X	X
Felt Sad	X	X	X		X
Felt Tired				X	X
Could Not Get Going	X	X	X		
Lots Of Energy					X
Fearful	X	X		X	
Hopeful	X	X		X	
Appetite	X	X			

Trouble Concentrating	X
Bothered By Things	X

The estimation of cognitive domains using confirmatory factor analysis

All models were scaled such that the scores on the latent cognitive domain estimated would have a mean of 0 and standard deviation of 1 within each specific HCAP study. To ascertain fit of estimated models to the data, we initially fit CFA models using a Weighted Least Squares estimator and evaluated model fit using the Root Mean Square Error of Approximation (RMSEA), the Comparative Fit Index (CFI) and the Standardized Root Mean Residual (SRMR) [2]. Model fit was considered excellent if the RMSEA was <0.05, CFI >0.95, TLI >0.95, and SRMR <0.05. Model fit was considered poor if the RMSEA >0.1, CFI <0.9, TLI <0.9, and SRMR >0.1. Where model fit was poor, we added additional methods factors to the model structure to explain higher correlations between similar items or items from the same cognitive test (i.e. immediate and delayed recall) [3]. These methods factors were selected by examining evidence of model misfit in conjunction with a priori knowledge of the content of items.

The following methods factors were used:

<u>United States</u>

Memory:

- Methods factor 1: Logical memory immediate, logical memory delayed, logical memory recognition
- Methods factor 2: Brave man immediate, brave man delayed
- Methods factor 3: CERAD word list immediate, CERAD word list delayed, CERAD word list recognition

Language:

- Methods factor 1: Naming a watch, naming a pencil

<u>England</u>

Memory:

- Methods factor 1: Logical memory immediate, logical memory delayed
- Methods factor 2: Brave man immediate, brave man delayed

Executive functioning:

- Methods factor 1: Trail-making test part A, trail-making test part B

Language:

- Methods factor 1: Naming a cactus, naming scissors

<u>South Africa</u>

Memory:

- Methods factor 1: CERAD word list immediate, CERAD word list delayed
- Methods factor 2: Logical memory immediate, logical memory delayed

India

Memory:

- Methods factor 1: Logical memory immediate, logical memory delayed, logical memory recognition
- Methods factor 2: Brave man immediate, brave man delayed
- Methods factor 3: CERAD word list immediate, CERAD word list delayed, CERAD word list recognition

Language:

- Methods factor 1: Naming scissors, naming a coconut

Mexico

Memory:

- Methods factor 1: Brave man immediate, brave man delayed
- Methods factor 2: MMSE 3-word immediate, MMSE 3-word delayed
- Methods factor 3: Logical memory immediate, logical memory delayed

Executive functioning:

- Methods factor 1: Serial 7's, serial 3's

Language:

- Methods factor 1: Naming a pencil, naming a shoe

To estimate factor scores used in the classification of cognitive impairment using the actuarial neuropsychological approach, a second set of CFA models was fit using the Maximum Likelihood (MLR) estimator to generate estimates of cognitive impairment for all study participants. CFA models estimated with an MLR estimator make an assumption that data are missing at random (MAR) and allow for the estimation of cognitive ability even with large amounts of missing data by relying on information from each of the other non-missing cognitive items included in the model. However, the reliability of the estimation of cognition in individuals with large amounts of missing data may be poor [4]. To prevent the scores estimated with low reliability from having an outsized influence on the estimation of cognitive impairment, we estimated reliability using the formula: $Reliability = 1 - standard\ error^2$. Reliability as measured using this formula is akin to a measure of internal consistency and reflects the precision of estimated cognitive scores [5]. We set all scores with a reliability of under 0.6 and greater than 50% missingness on cognitive items to be missing [6].

Final CFA models by study and cognitive domain generally showed adequate to good fit (Table S2).

Table S2. Fit statistics for confirmatory factor analysis models of cognitive domains in the United States, England, South Africa, India, and Mexico Harmonized Cognitive Assessment Protocol (HCAP) samples

Sample	Domain	Parameters	CFI	TLI	RMSEA	SRMR
United States	Orientation	20	0.971	0.963	0.028 (0.023-0.034)	0.064
England	Orientation	20	0.997	0.997	0.015 (0.000-0.027)	0.064
South Africa	Orientation	8	0.989	0.967	0.042 (0.000-0.100)	0.089
India	Orientation	20	0.954	0.941	0.039 (0.034-0.043)	0.089
Mexico	Orientation	16	0.924	0.894	0.062 (0.053-0.070)	0.066
United States	Memory	53	0.986	0.977	0.040 (0.035-0.046)	0.019
England	Memory	46	0.954	0.938	0.075 (0.067-0.082)	0.044
South Africa	Memory	24	0.978	0.959	0.084 (0.063-0.106)	0.036
India	Memory	53	0.982	0.971	0.044 (0.040-0.049)	0.023
Mexico	Memory	46	0.986	0.979	0.045 (0.038-0.052)	0.028
United States	Executive Functioning	27	0.927	0.898	0.110 (0.104-0.117)	0.036
England	Executive Functioning	28	0.886	0.832	0.140 (0.130-0.151)	0.050
South Africa	Executive Functioning	23	0.914	0.857	0.086 (0.063-0.110)	0.062
India	Executive Functioning	41	0.989	0.986	0.033 (0.028-0.037)	0.023
Mexico	Executive Functioning	31	0.971	0.953	0.086 (0.076-0.097)	0.034
United States	Language	32	0.973	0.967	0.020 (0.016-0.024)	0.067
England	Language	28	0.997	0.997	0.007 (0.000-0.019)	0.070
South Africa	Language	17	0.975	0.965	0.030 (0.000-0.050)	0.120
India	Language	32	0.922	0.906	0.032 (0.029-0.035)	0.060
Mexico	Language	30	0.975	0.970	0.021 (0.016-0.027)	0.076

* CFI = Comparative Fit Index, TLI = Tucker-Lewis Index, RMSEA = Root Mean Square Error of Approximation, SRMR = Standardized Root Mean Residual

There was some evidence of poor fit in models for executive functioning in the United States and England, likely due to the lower inter-correlations between the more diverse set of items measuring executive functioning. The reliability, or precision of the measurement, of the cognitive scores was generally fair to good, although lower reliabilities were observed for the orientation domain, as this domain was composed solely of binary cognitive items which provide less information than continuous items, particularly in the less impaired range of cognitive performance (Figure 2). Outside of the orientation domain, only language scores in England (41%) and India (47%) had greater than 25% of scores with reliabilities under 0.6. Many CFA models had scores with high reliabilities, and models of executive functioning in the United States, England, India, and Mexico, memory in the United States, England, Mexico, and South Africa all had scores with mean reliabilities of over 0.8.

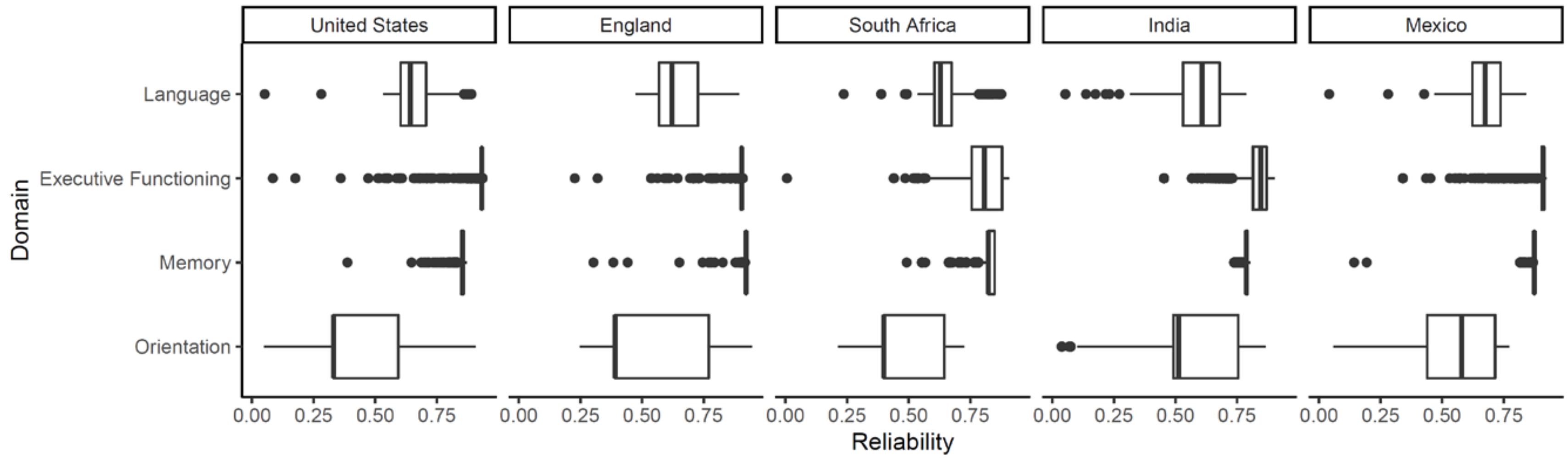


Figure S1. Distributions of the reliability of estimated cognitive scores by cognitive domain and study sample. Cognitive scores were estimated using confirmatory factor analysis

Definition of the cognitively robust group for the neuropsychological norms classification approach

We used questions from the CSID and the 10/66 assessment of functional limitations to select individuals who had no reports of: (1) changes in daily activities, (2) general decline, (3) difficulty remembering, (4) changes in the ability to handle money, (5) forgetting friends or family's names, (6) using the wrong words, (7) forgetting where they are, when they last saw the informant, or what happened yesterday, (8) getting lost in the community or at home, (9) changes in ability to think or reason things through. We also excluded individuals with the highest 10% of depressive symptom burden in each sample, and individuals who either used a proxy informant or had a self-reported stroke or heart attack in the prior HRS IPS wave. Additionally, we excluded individuals with informant-reported stroke, Alzheimer's disease or memory problems in all samples except Mexico, as these data were not available in the Mexican HCAP survey. Due to low endorsement of limitations in the South African sample as compared to other HCAP samples, we further excluded individuals with fair or poor self-reported health in the South African sample. While individuals with missing data on individual items used to create the normative sample were considered to be "not impaired" on these individual items, we additionally excluded individuals with greater than 50% missing data on the items used for the selection of the normative group from the normative group.

Estimation of residual scores for the neuropsychological norms classification approach

We estimated multiple linear regression models within the normative sample of each HCAP study to quantify the relationship between cognitive functioning on each domain, adjusting for basic demographic factors. For all studies we included age, gender, and educational attainment (dichotomized). We included this crude marker of education to ensure that we accounted for differences in educational attainment that would confer advantages or disadvantages on the cognitive testing used to measure cognitive functioning (bias), but that we did not control for variation in education that is expected to be associated with true variability in cognitive functioning [28]. We additionally included race and ethnicity in regression model for the US HCAP study, rurality in the regression model for the India and Mexico studies, and literacy status in the regression models for the India and South Africa studies.

Using the coefficients from these models we calculated expected cognitive performance on each domain for each respondent in the full sample by predicting scores based on participants' demographic characteristics. We then used these predictions to calculate residual scores, scaled by the standard error of the regression equation, using the formula:

$$Residual\ Score = \frac{True\ Cognitive\ Score - Predicted\ Cognitive\ Score}{SE\ of\ the\ Regression\ Equation}.$$

This ensures that the variability in the residual scores (which is used to determine cutoffs of cognitive impairment) is proportional to the variability that remained unexplained by the regression equation, and is therefore not attributable to demographic factors.

Details on latent class analysis models

Latent class analysis is a data-driven approach to dividing survey respondents into unimpaired and impaired groups, to derive these classifications and to evaluate the associations between impairment status and responses to individual items on cognition and function. We fit latent class models for each sample separately and a priori specified two classes based on our assumption of the existence of an unimpaired and an impaired class. We used the conditional item probabilities to label each class as impaired or unimpaired. We examined either the item odds ratios (the odds of either endorsing the item or getting the item correct, comparing the impaired class to the unimpaired class) for binary items or the item differences (the difference in the mean item score comparing the impaired class to the unimpaired class) for continuous items to assess the relationship between each item and impairment status. Where estimated item odds ratios were either zero or infinite due to the estimation of boundary values in the maximum likelihood procedure, we capped these odds ratios at 0.01 (for an estimate of 0) or 100 (for an infinite estimate).

Consistency of cognitive impairment classification with leave one out approach

Use of data from cognitive items in the classification of cognitive impairment followed by the evaluation of the association between cognitive items and the cognitive impairment classification could overestimate the association due to circularity in the analysis. To avoid this circularity, we conducted a quasi-leave one out analysis, wherein we recalculated the classification of cognitive impairment for each association of interest, leaving out data on each specific cognitive item in turn. However, we wanted to ensure that despite leaving out data on individual cognitive items, our classifications remained stable. To evaluate the stability of our classifications, we calculated the accuracy ($\frac{true\ positives + true\ negatives}{total}$) of our classifications in comparison to the set of classifications using all available cognitive items. We found a high level of accuracy (>90%) in all instances, across all studies and items considered.

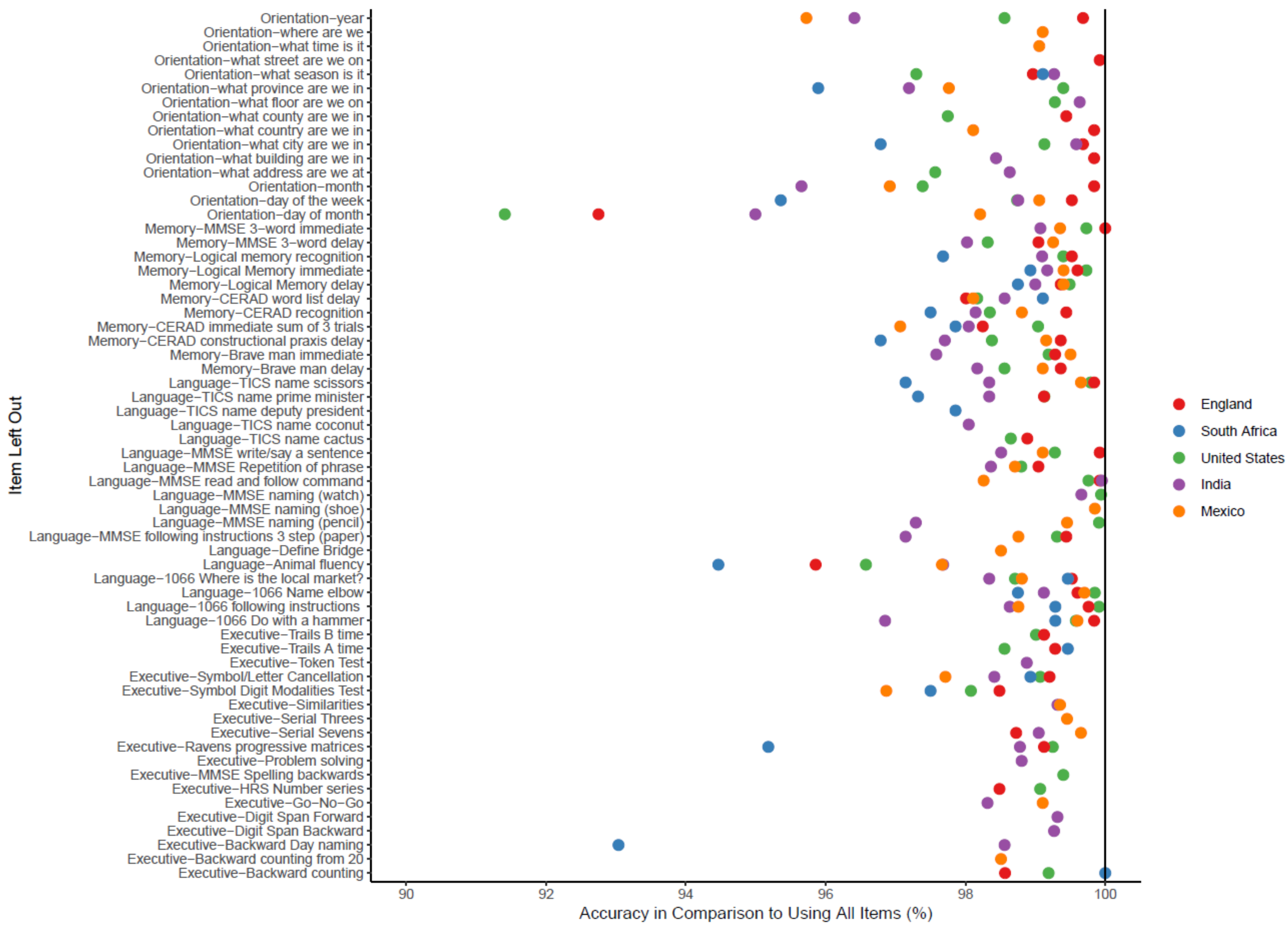


Figure S2. The accuracy of the neuropsychological norms classification algorithm leaving out data on each cognitive item, in comparison to the application of the neuropsychological norms method to all available cognitive data across the Harmonized Cognitive Assessment Protocol Surveys (HCAP) conducted in the United States (N = 3329), England (N = 1255), South Africa (N = 560), India (N = 4095), and Mexico (N = 2011)

Percent of data missing on cognitive items

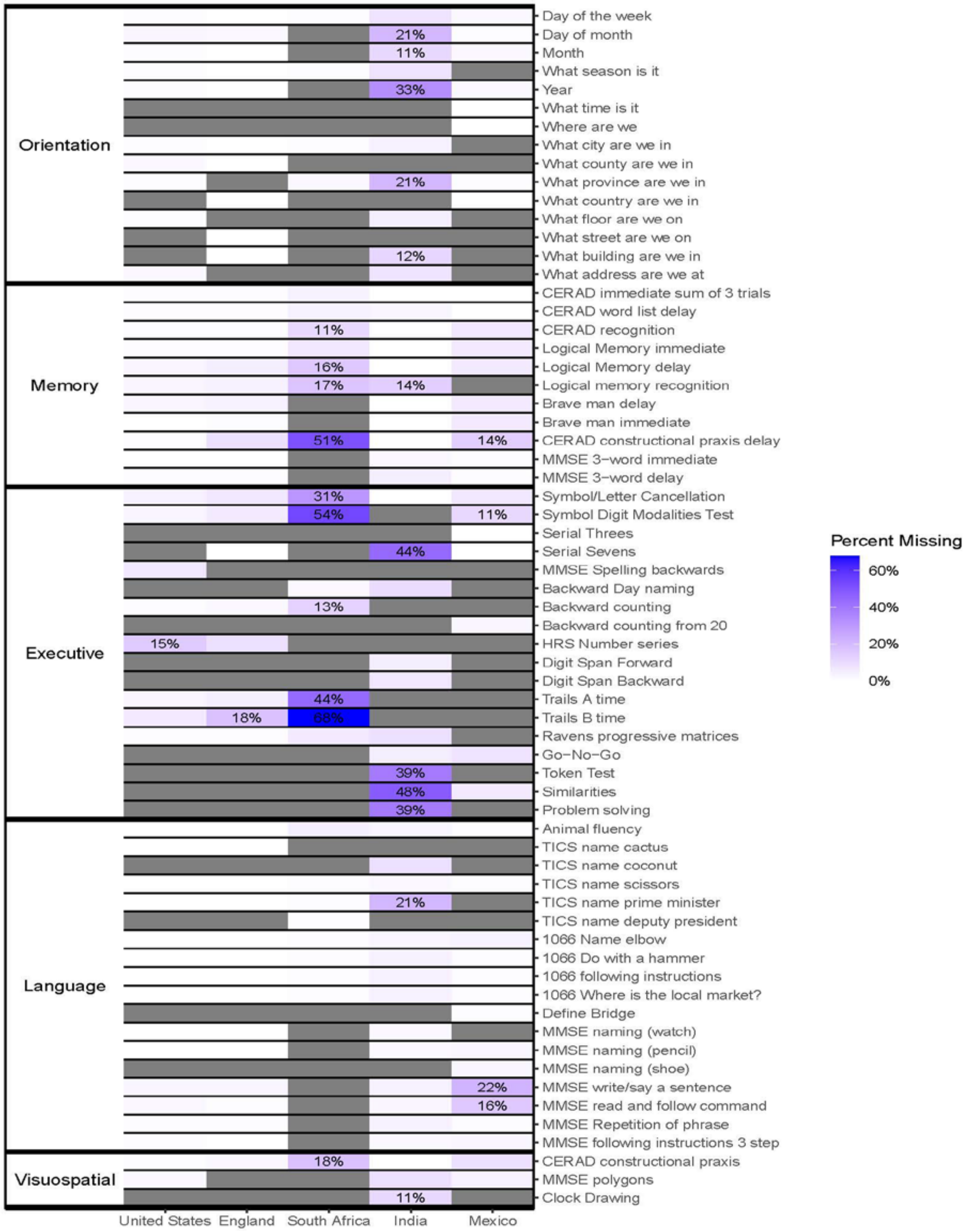


Figure S3. Percentage of data missing on each cognitive item across the Harmonized Cognitive Assessment Protocol Surveys (HCAP) conducted in the United States (N = 3329), England (N = 1255), South Africa (N = 560), India (N = 4095), and Mexico (N = 2011). Numbers are shown when the percent missingness is higher than 10%, grey boxes are shown when the item was not administered in a study.

Items on cognition with suppressed odds ratios

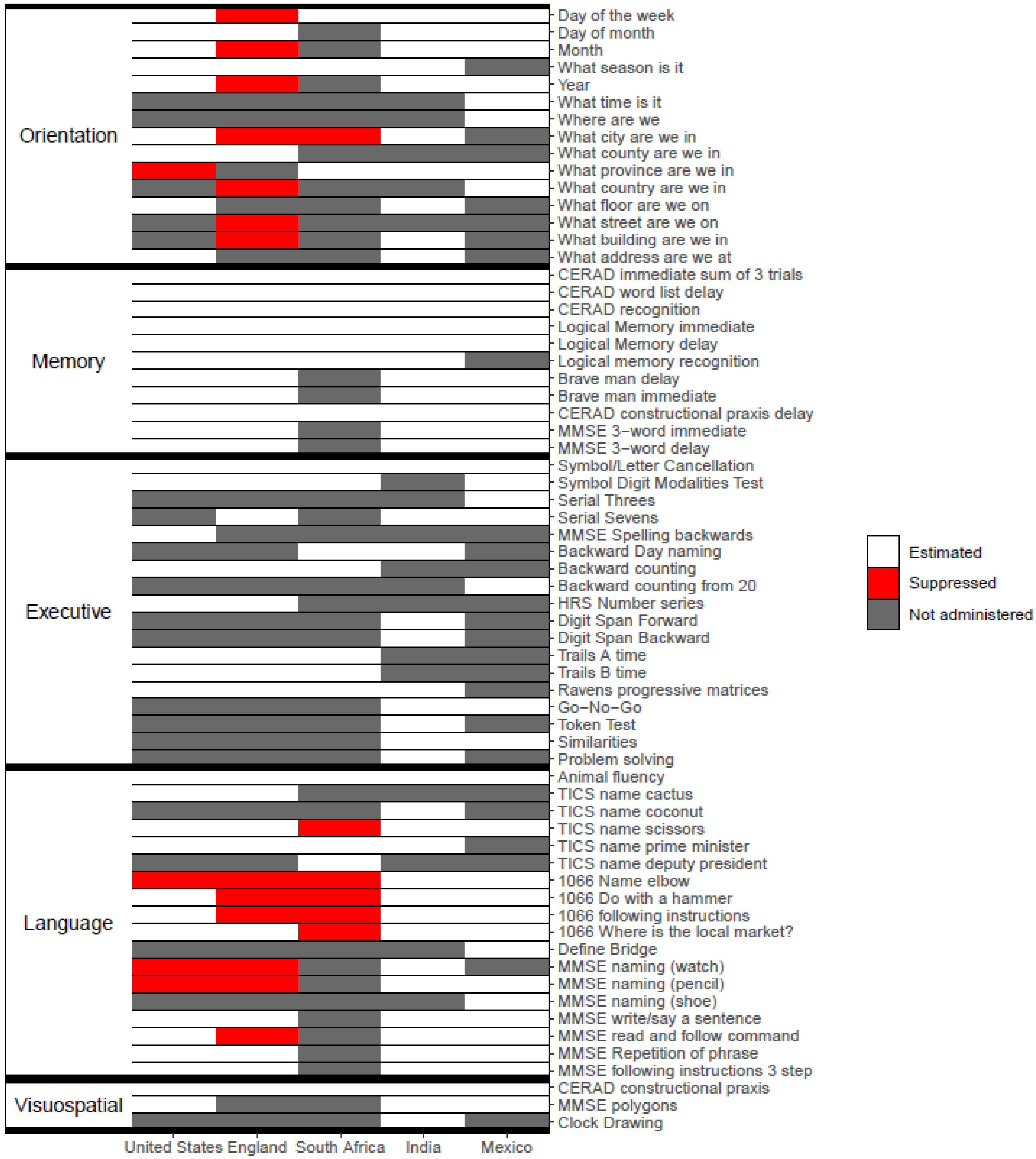


Figure S4. Cognitive items with suppressed odds ratios due to low variability in responses across the Harmonized Cognitive Assessment Protocol Surveys (HCAP) conducted in the United States (N = 3329), England (N = 1255), South Africa (N = 560), India (N = 4095), and Mexico (N = 2011). Grey boxes indicate items that either weren't assessed.

Variability (proportion correct) for binary cognitive items

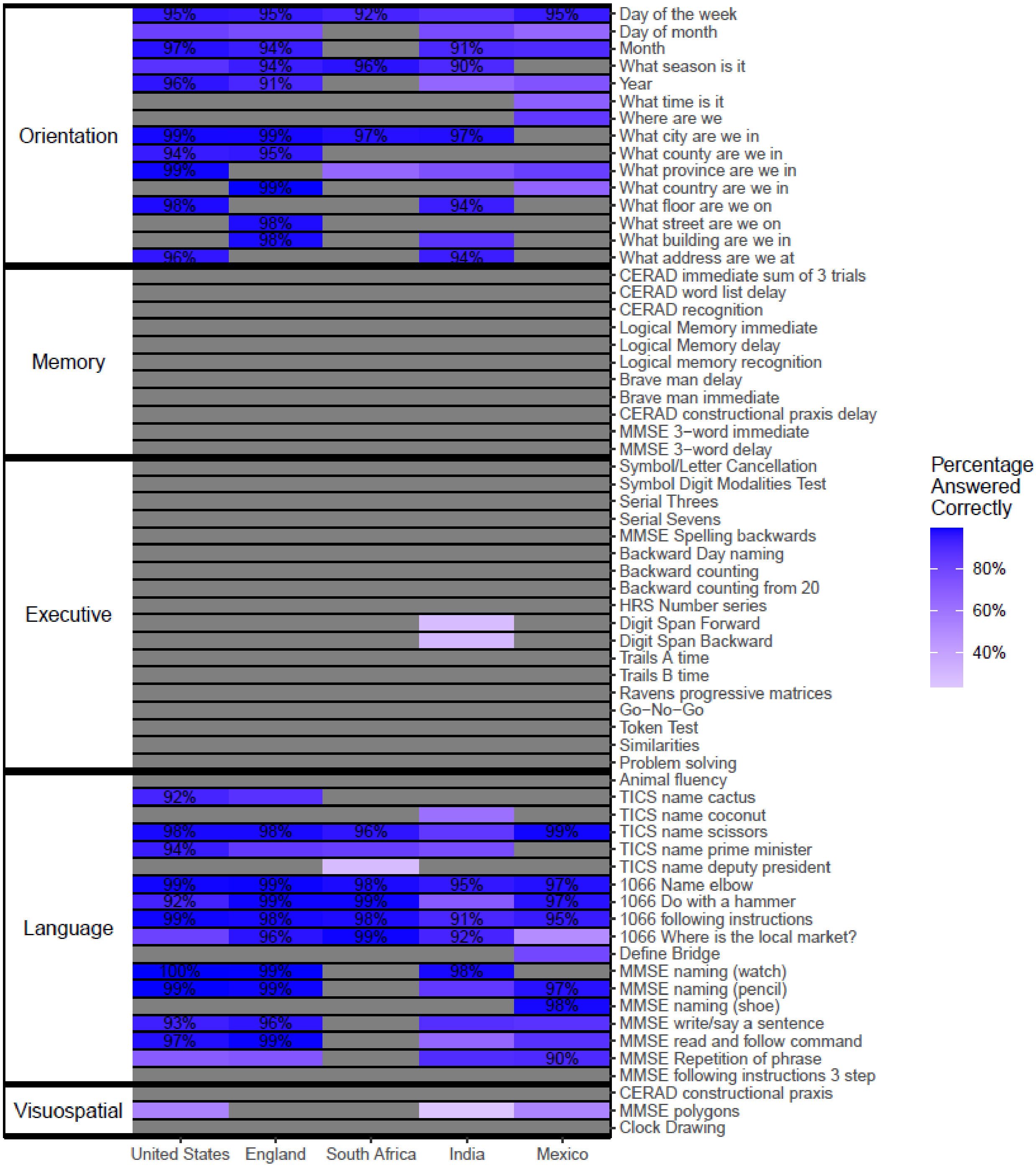


Figure S5. Variability (proportion correct) for binary cognitive items across the Harmonized Cognitive Assessment Protocol Surveys (HCAP) conducted in the United States (N = 3329), England (N = 1255), South Africa (N = 560), India (N = 4095), and Mexico (N = 2011). Grey boxes indicate items that either weren't assessed, or were not binary.

Sensitivity analysis using LCA (binary cognitive items)

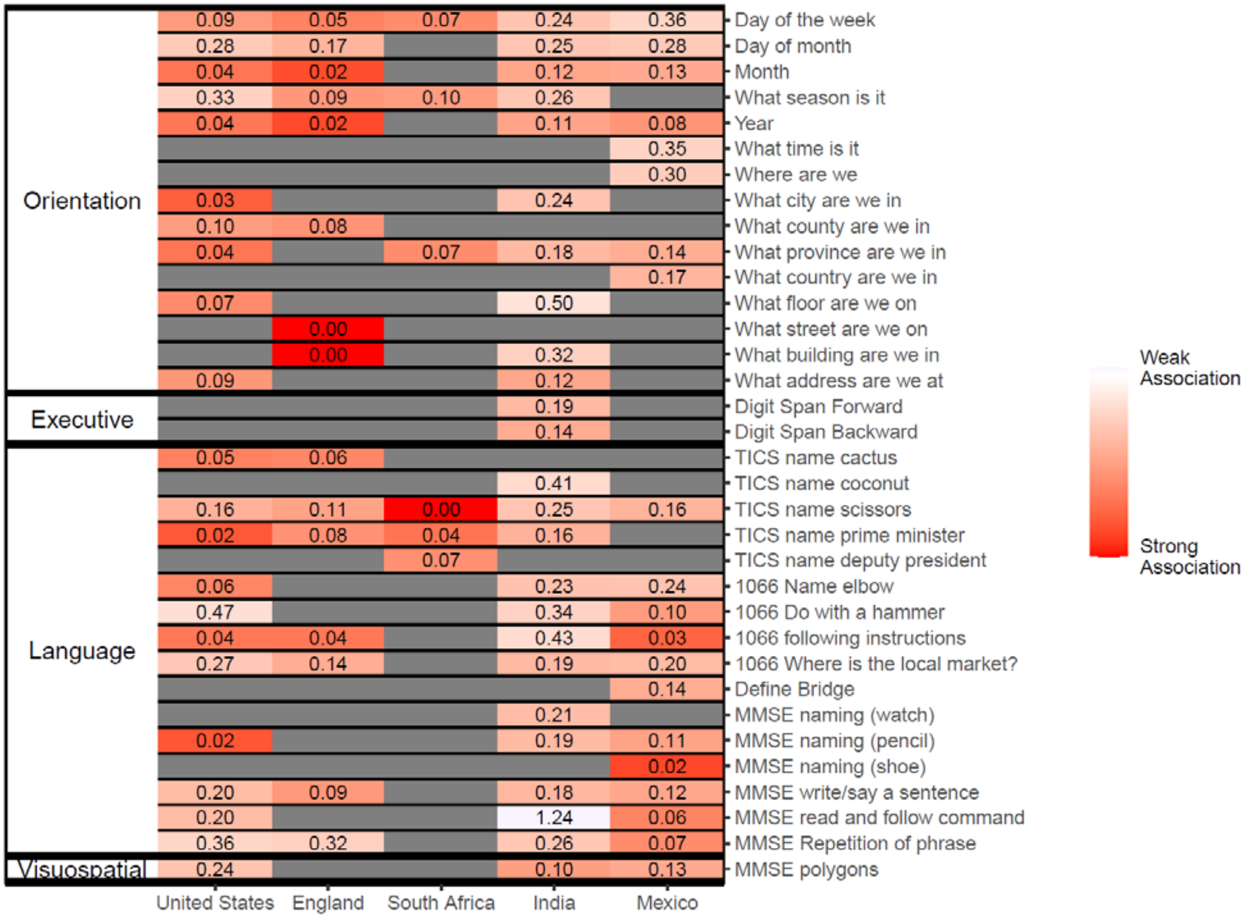


Figure S6. Associations between each binary cognitive test item and cognitive impairment by domain in each Harmonized Cognitive Assessment Protocol Surveys (HCAP) conducted in the United States (N = 3329), England (N = 1255), South Africa (N = 560), India (N = 4095), and Mexico (N = 2011) from latent class analysis. Numbers indicate odds ratios compared individuals in the impaired class to individuals in the unimpaired class. Grey boxes represent instances were an item was not administered or an odds ratio was suppressed due to small cells. Color scale represents log odds ratios.

Sensitivity analysis using LCA (continuous cognitive items)

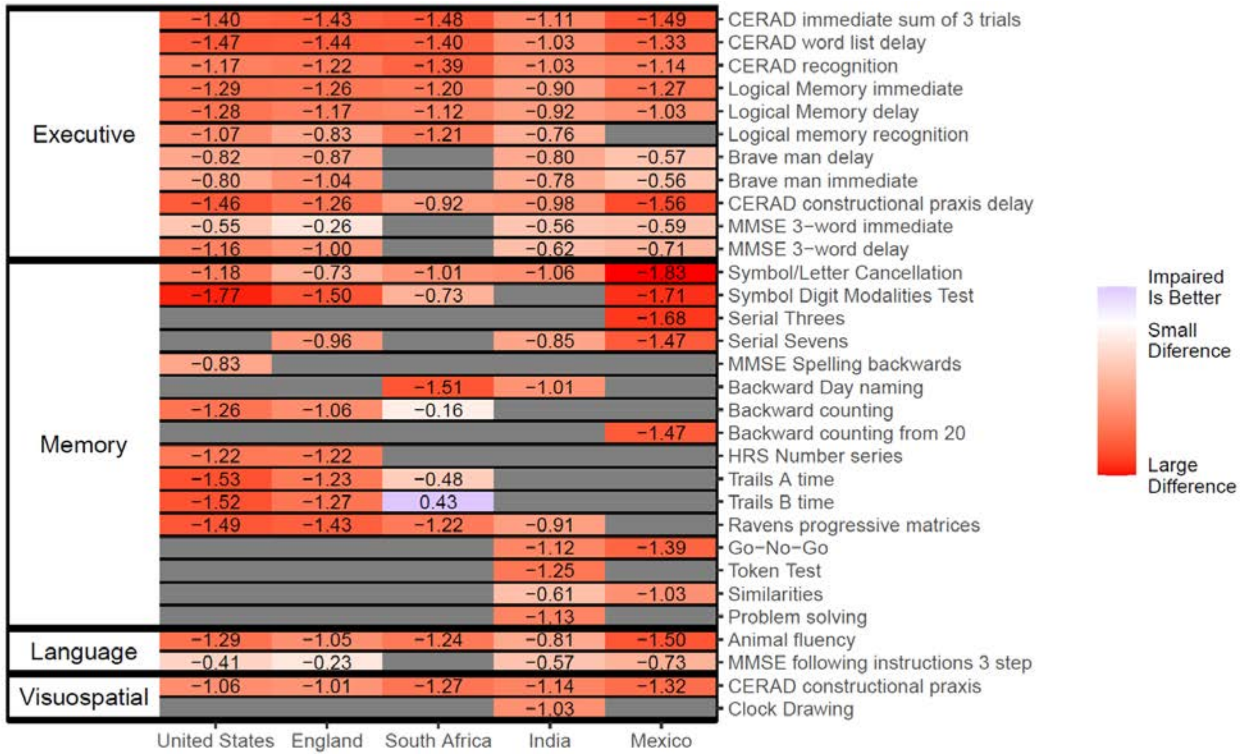


Figure S7. Differences in scores for each cognitive test item comparing the impaired class to the un-impaired class by domain in each Harmonized Cognitive Assessment Protocol Surveys (HCAP) conducted in the United States (N = 3329), England (N = 1255), South Africa (N = 560), India (N = 4095), and Mexico (N = 2011) from latent class analysis. Grey boxes represent instances were an item was not administered or an odds ratio was suppressed due to small cells. Color scale represents log odds ratios.

Sensitivity analysis (restricted to age 65+)

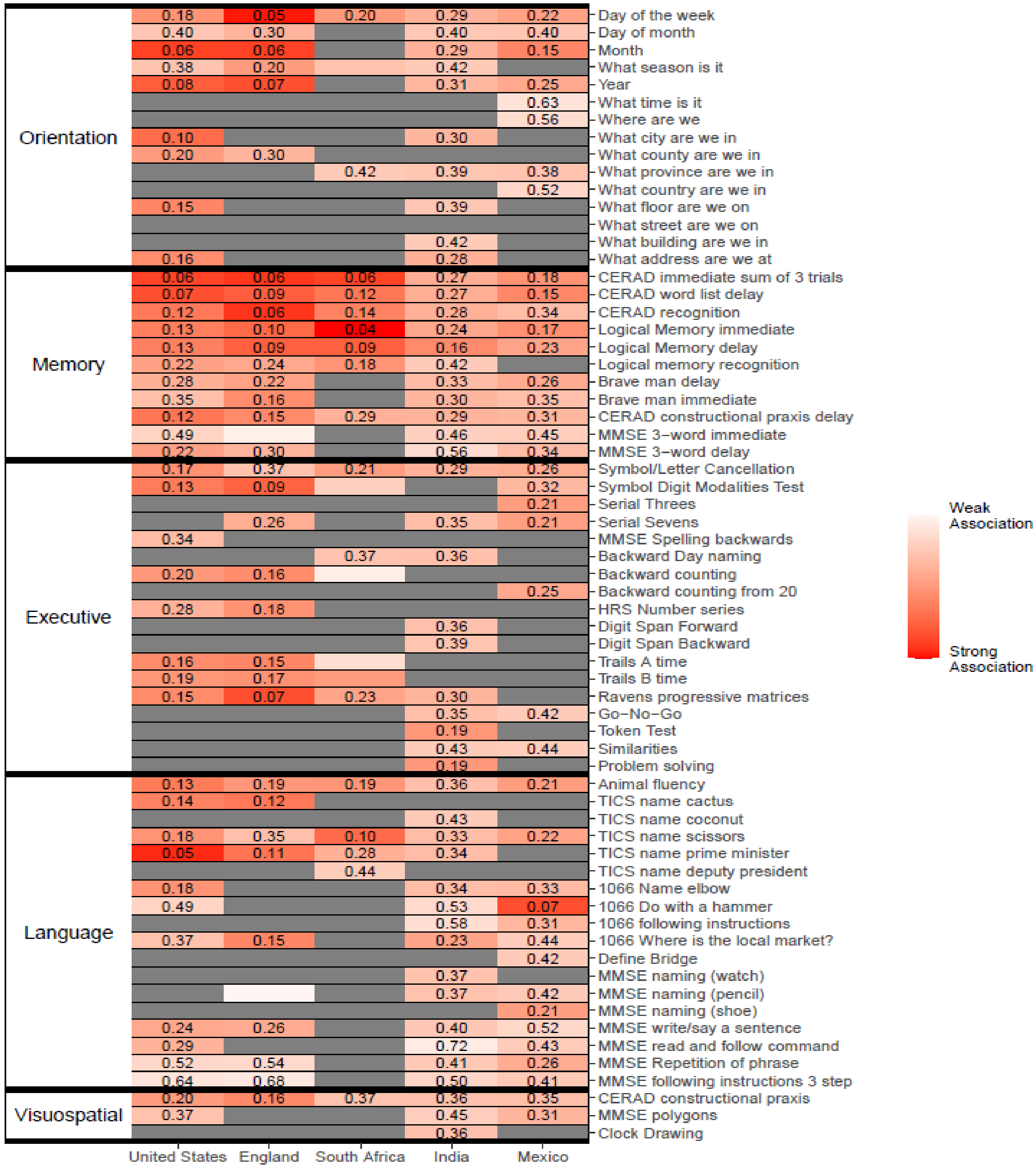


Figure S8. Associations between each cognitive test item and cognitive impairment by domain for participants age 65+ in each Harmonized Cognitive Assessment Protocol Surveys (HCAP) conducted in the United States (N = 3329), England (N = 1255), South Africa (N = 560), India (N = 4095), and Mexico (N = 2011) from logistic regression models, controlling for age and gender. Odds ratios are displayed for significant associations. Grey boxes represent instances were an item was not administered or an odds ratio was suppressed due to small cells. Color scale represents log odds ratios.