**Methods**

**Rasch Analysis**

Rasch analysis was applied to summarize neuropsychological performance; we have described the conceptual and mathematical advantages of this approach over more conventional averaging methods in detail elsewhere [1](#_ENREF_1) .

The extent to which the scores on neuropsychological tests fit a unidimensional, hierarchical model representing the single latent construct of cognitive ability was tested using Rasch analysis. The partial credit model for ordered response categories was used within the Rasch Unidimensional Measurement Model program (RUMM 2030). To fit the data to the Rasch model, the continuous measurement scale for each neuropsychological test (termed items) was converted into unique ordered categories, the number of which depended on the distribution of the outcome.

The first step is to ensure hierarchically ordered thresholds such that a person with a higher level of cognitive ability is more likely to pass thresholds determined to be more difficult. With k ordered categories, the number of thresholds is k-1. Disordered thresholds were iteratively rescored by collapsing adjacent response options. After rescoring, higher values on these tests (items) now represent more cognitive ability.

Fit of the data to the model was evaluated using a *χ*2 goodness of fit test where a probability value >0.05 indicates fit. Two other measures of test fit were also used: item-fit residuals between -2.5 and 2.5 and item Chi-square tests >0.05. Reliability was also estimated, using the person separation index (PSI).

Pairs of items with residual correlation >0.4 were examined and one of the items was removed; the choice was based on both theoretical and statistical considerations. Once the best fitting set of tests was identified, unidimensionality was assessed by the principal component analysis (PCA) of the fit residuals. If the first PCA explains >10% of residual variance, this is an indicator of lack of unidimensionality. Items that remain highly correlated once the effect of the latent trait is removed also indicate lack of unidimensionality. Once a fitting model was obtained, items were examined for differential item functioning (DIF) across sex, age, education, race and HIV status. DIF was considered present for items with *p*-values on a two-way ANOVA that were significant at the level of 0.05 after applying the Bonferroni correction for multiple comparisons. Targeting of patient ability to item difficulty was examined via the person-item distribution plot.

**Statistical Analyses**

Regional brain volume and cortical thickness estimates were modeled by fitting the following linear equation at each voxel:

$$Vol= β\_{0}+β\_{1}∙Age+β\_{2}∙Gender+β\_{3}∙Edu+β\_{4}∙Ethnicity+β\_{5}∙Covariate of interest+ε$$

Here, *Vol* is the value of the Jacobian determinant, tissue density, or cortical thickness estimate, *Covariate of interest* is HIV status, nadir CD4, current CD4, current viral load, viral suppression status (detectable versus undetectable), CPE score, treatment status (treated versus untreated), or cognitive function as summarized by Rasch analysis, and *ɛ* is the residual. Discrete variables (gender, ethnicity, HIV status, viral suppression status, and treatment status) were treated as binary variables. Each linear model was applied in a voxel-wise fashion creating a 3D statistical map allowing the patterns of significance to be visualized. Both positive and negative correlations were tested in all models.

The final linear model was chosen by minimizing the Akaike Selection Criterion (AIC) [2](#_ENREF_2). Increasing the number of parameters in the model improves the fit, but it comes at the cost of increasing complexity. The AIC balances the goodness of fit and the number of parameters by penalizing the number of covariates in the model. The model that has the lowest AIC is suggested to be the most optimal model. While there were no group differences in education and ethnicity, these variables were included in the final linear model because they minimized the AIC.

**Results**

**Rasch Analysis**

All available neuropsychological test scores were entered into a Rasch analysis, to determine if they could be summarized as reflecting a single latent construct. The sample available for the Rasch analysis (N=388; 284 HIV-infected, 104 HIV-uninfected) was larger than the sample available for the imaging analysis; in the latter, participants were excluded if they were missing neuroimaging, demographic or clinical information required for the primary analysis.

Neuropsychological tests scores were initially modelled with between seven and 13 categories; ten of the 16 tests had disordered thresholds. After rescoring, the number of levels per item was reduced to between two and eight. Due to response dependency, three items with correlations greater than 0.4 were removed: Hopkins Verbal Learning Delayed Recall, one measure from the Stroop test, and Grooved Pegboard, Dominant Hand. Grooved Pegboard, Non-Dominant Hand was found to have DIF by age (i.e. suggesting that the item was not measuring the same ability in all participants), and was also dropped.

The final model consisted of 12 NP tests (Table S1). Reliability of the final model was 0.85, regardless of whether extreme values were included. Items and persons fit the model, with fit residuals having a mean 0.10 (SD: 1.06) and -0.29 (SD: 1.03), respectively. (Ideal fit residuals have a mean of 0 and standard deviation of 1.) All item fit residuals fell between -2.5 and 2.5 and none were statistically significant. Global fit was confirmed (nonsignificant global item trait *X*2 of 85.24 on 72 df, p=0.14). Targeting of person ability to item difficulty was good, with a fluid distribution of both persons and items across seven logits, from -3 to 4, though items were available over nine logits, from -4 to 5 (Figure S1).

The final model was used to estimate the performance of each participant on the latent construct of ‘cognitive ability’, and these scores were then entered into the analyses relating brain structural indices to neuropsychological performance described in the main paper.

Table S1: Neuropsychological tests that contributed to the Rasch model (mean raw scores ± SD).

|  |  |  |
| --- | --- | --- |
|  | **HIV-infected**  | **HIV-uninfected** |
| Hopkins Verbal Learning Total Learning Score  | 21.2 ± 5.24 | 24.2 ± 4.37 |
| Symbol Search  | 26.5 ± 8.36 | 33.6 ± 6.81 |
| Letter Number Sequencing | 7.79 ± 3.81 | 9.14 ± 2.71 |
| Digit-Symbol | 8.92 ± 3.23 | 10.79 ± 3.09 |
| Trail-Making Test A  | 36.5 ± 19.7 | 26.1 ± 7.74  |
| Trail-Making Test B | 97.0 ± 50.0 | 72.7 ± 41.3 |
| Stroop 1 Time | 33.6 ± 8.17 | 29.0 ± 5.35 |
| Stroop 3 Time | 76.1 ± 30.5 | 56.6 ± 11.4 |
| Verbal Fluency | 12.4 ± 5.56 | 14.9 ± 6.42 |
| FAS Fluency | 38.1 ± 43.7 | 38.8 ± 12.2 |
| Animal Fluency | 19.0 ± 5.12 | 22.8 ± 4.94 |
| Test of Memory Malingering | 46.0 ± 5.20 | 48.7 ± 1.94 |



Figure S1: Distribution of persons (HIV-infected and HIV-uninfected groups collapsed) located according to position along a continuum of ‘cognitive ability’ (top panel), and the distribution of the items retained in the Rasch model that measure this latent construct (bottom panel). Items were available to assess the full range of ability demonstrated by the sample, with good coverage in the range of ability showed by most of the sample (i.e. items were well-targeted to the ability of the persons in this sample).

**HIV-infected versus HIV-uninfected**



Figure S2: Axial slices of deformation-based morphometry (DBM) in HIV-infected compared to HIV-uninfected subjects. Volume loss was detected in the brainstem and thalamus at trend level (*p<*0.1) in HIV-infected patients. (Image created with MRIcron <http://people.cas.sc.edu/rorden/mricron/index.html>)

**Supplementary Reference**

1. Brouillette M, Mayo N, Fellows LK, et al. A better screening tool for HIV-associative neurocognitive disorders: is it what clinicians need? *AIDS.* 2015;29:895-902.

2. Akaike H. A New Look at the Statistical Model Identification. *IEEE Trans Autom Control.* 1974;AC-19(6):716-723.