Scale Construction

The distributions, correlations with age, and group differences for each primary variable are described in the manuscript (also see Table 3 in manuscript). We also explored ways to reduce the data into a smaller number of scales that could be used by researchers and in clinical trials. Insight was excluded from these analyses because it did not discriminate well between healthy subjects and patients. Several exploratory factor analyses were carried out that suggested that a single factor model might explain the majority of the variance, and that a 3-factor model was also a good fit for the data, with the three working memory scores comprising one factor, the four fluency measures comprising a second factor, and flanker, set shifting, errors, and anti-saccades forming a third. The unstructured task and Social Norms did not load clearly on any of the three factors.

Confirmatory factor analysis (CFA) informed by conceptual models from previous literature and by results of exploratory factor analysis was used to further evaluate dimensional structure and identify homogenous groups of variables that could be used to generate composite scores. These analyses were based on 11 core variables identified in previous analyses: Dot Counting total, d-prime measures from the 1-back and 2-back, total Flanker score, total Set Shifting score, anti-saccade total, total correct responses from each of the verbal fluency trials, and total dysexecutive errors. The planning, insight, and social cognition variables did not fit well with these 11 core variables and were not further considered for scale construction, despite their considerable value as independent procedures ([Krueger et al., 2011](#_ENREF_4); [Shany-Ur & Rankin, 2011](#_ENREF_8)). Two alternative models were tested: 1) a 1-factor (unidimensional) model in which all 11 variables defined a single factor, and 2) a 3-factor model with factors representing cognitive control (Flanker, Set Shifting, anti-saccade, dysexecutive errors), working memory (Dot Counting, 1-back, 2-back), and fluency (phonemic and semantic). The 3-factor model was supported by previous literature and by preliminary results from the exploratory factor analyses. CFA initially was conducted using the adult sample (age 18+) and invariance across age groups was subsequently tested. Patients and healthy normal subjects were combined in these analyses. Results from the initial assessment for English language administrations were used. Residual correlations between variables that shared a common method (category fluency, letter fluency, dprime) or were reaction time based (flanker and set shifting) were allowed as model modifications.

The 1-factor CFA model fit relatively well in the adult sample (χ2[40]=137.4, CFI =0.950, TLI =0.931, RMSEA = 0.065 (90% CI =0.053-0.077), SRMR =0.043). The 3-factor model also yielded excellent fit (χ2[39]=90.2, CFI =0.974, TLI =0.963, RMSEA = 0.048 (90% CI =0.035-0.061), SRMR =0.033). The three factors were highly correlated (fluency with working memory and cognitive control = 0.73, cognitive control with working memory = 0.91). These results suggest that three dimensions very effectively explain intercorrelations of the 11 variables, but that that the 11 variables might be sufficiently unidimensional to be combined into a single score. An additional analysis was performed to evaluate whether the one-factor model was sufficiently unidimensional to support a global executive function composite score based on all eleven variables. A bifactor model was fit in which all eleven variables loaded on a global factor, and in addition, each variable also loaded on a specific factor corresponding to the three-factor model. The global and specific factors were all constrained to be uncorrelated to ensure identifiability of the model. This model showed excellent fit (χ2[32]=69.5, CFI =0.981, TLI =0.967, RMSEA = 0.045 (90% CI =0.031-0.060), SRMR =0.028). Loadings of all 11 variables on the global variable were strong (0.40 – 0.62) and exceeded loadings on the specific factors with the exception of dprime on the 2-back (global factor loading = 0.59, specific factor loading = 0.76). The overall CFA results provide strong support for combining the 11 variables into a single executive function measure, but in addition, suggest that more refined characterization of three sub-components of executive function can be achieved.

Multiple group CFA analyses were used to test invariance of 1-factor and 3-factor models across age groups. In multiple group CFA, a common model for all groups is specified on an a priori basis, and then group differences in individual parameters can be systematically tested. The 1-factor and 3-factor models derived from the 18-64 year old age group were the base models for these analyses. We initially freely estimated loading, intercept, and residual variance parameters in four age groups (3-10 years, 11-17 years, 12-64 years, 65+ years); one loading and intercept for each factor were constrained to equality to identify the models. We then constrained all loadings to be equal in the four groups, and compared the fit of this model with the freely estimated model using the chi-square difference test ([Steiger, Shapiro, & Browne, 1985](#_ENREF_9)). We used modification indices ([Muthén & Muthén, 1998-2012](#_ENREF_7)) to identify non-invariant loadings which subsequently were freely estimated in the different groups. We then constrained all intercepts to equality, and used modification indices to identify non-invariant intercepts. In the next step we constrained residual variances to equality and identified non-invariant parameters with modification indices. This process was repeated iteratively with loadings, intercepts, and residual variances until no significant modifications (p<0.01) were identified.

Results for the 3-factor solution supported partial strong invariance ([Gregorich, 2006](#_ENREF_2); [Meredith & Teresi, 2006](#_ENREF_5); [Widaman & Reise, 1997](#_ENREF_10)). Loadings were invariant across age groups, but there were some difference in intercepts. All of the loadings and 31 of the 44 intercepts were invariant across groups. This level of invariance indicates that the same factors account for intercorrelations of test scores in the different age groups and establishes a common metric for factor scores. The 1-factor solution identified a small number of non-invariant loadings (5 of 44, 4 for the 65+ year age group) and 12 (of 44) non-invariant intercepts. These results also indicate partial strong invariance that means that the same constructs are being measured in the different age groups.

Results of the CFA analyses indicated that: 1) the EXAMINER tests can be well characterized by measures of working memory, fluency, and control, but in addition, 2) a global measure of executive function also supported by psychometric results. We used item response theory (IRT) methods ([Hambleton, Swaminathan, & Rogers, 1991](#_ENREF_3); [Mungas, Reed, & Kramer, 2003](#_ENREF_6)) to generate scores corresponding to these four variables: global executive function, cognitive control, fluency, and working memory. IRT has important invariance properties, and of particular relevance to EXAMINER, examinee scores generated by IRT analysis are invariant to specific items used. Consequently, an IRT score should provide an unbiased estimate of the examinee’s ability even if different variables are used to generate that score. IRT is a special case of latent variable modeling that models item difficult in relation to person ability to maximize measurement precision. Classical test theory scores, which consist of numeric sums of items, have an implicit underlying assumption that all items contribute equally to total score. IRT scores have an advantage in that the contribution of items to scores is weighted by how effectively the item measures at any part of the ability continuum.

Each of the 11 continuous variables was recoded into an ordinal score with up to 20 response categories with at least 10 observations in each response category. This process yielded ordinal scores that roughly matched the distributions of the raw continuous variables. These ordinal scores were then entered into an IRT analyses corresponding to the four scores of interest. The R ltm module was used to fit a two parameter graded response model. Item parameters were calibrated and saved, and examinee scores and standard errors were calculated using Empirical Bayes scoring. Mirroring the CFA analyses, this initially was performed with 18-64 year old English speakers, but subsequent analyses tested for differential item function ([Camilli & Shepard, 1994](#_ENREF_1)) related to age group (3-10, 11-17, 18-64, 65+) and language (English, Spanish). Differential item function means that items measure differently in different groups, and corresponds to measurement non-invariance that was examined (with respect to age) in the CFA analyses. Concretely, DIF is present when loading or threshold parameters differ across groups (age or language). DIF was evaluated in an iterative process in which the baseline model used the same parameters for all groups, but one additional parameter was freely estimated in each subsequent step. This process was empirically guided such that the variable selected for differential parameter estimation in any given step yielded the largest improvement in model fit. This process was continued until further improvement in model fit was not obtained. The score from this last iteration was considered a DIF free standard. This means that measurement differences associated with age or language were explicitly included in the final analytic model, and consequently, the scores obtained in different age or language groups can be directly comparable. Results showed that accounting for DIF related to age group did not change resulting IRT scores, but indicated that it was important to account for language for the global executive function score. Consequently, the same item parameters were used across all age groups, but parameters for the global executive function score differed across language groups for seven of the 11 variables.

References

Camilli, G., & Shepard, L.A. (1994). *Methods for identifying biased test items.* Thousand Oaks:: Sage.

Gregorich, S. E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. *Med Care, 44*(11 Suppl 3), S78-94.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Krueger, C. E., Rosen, H. J., Taylor, H. G., Espy, K. A., Schatz, J., Rey-Casserly, C., & Kramer, J. H. (2011). Know thyself: real-world behavioral correlates of self-appraisal accuracy. *Clin Neuropsychol, 25*(5), 741-756.

Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Med Care, 44*(11 Suppl 3), S69-77.

 Mungas, D., Reed, B. R., & Kramer, J. H. (2003). Psychometrically matched measures of global cognition, memory, and executive function for assessment of cognitive decline in older persons. *Neuropsychology, 17*(3), 380-392.

Muthén, L.K., & Muthén, B.O. (1998-2012). *Mplus User’s Guide. Seventh Edition.* . Los Angeles: Muthén & Muthén.

Shany-Ur, T., & Rankin, K. P. (2011). Personality and social cognition in neurodegenerative disease. *Curr Opin Neurol, 24*(6), 550-555.

Steiger, J. H., Shapiro, A. , & Browne, M. W. . (1985). On the multivariate asymptotic distribution of sequential chi-square statistics. *Psychometrika, 54*, 371-384.

Widaman, K. F., & Reise, S. P. (Eds.). (1997). *Exploring the measurement invariance of psychological instruments: Applications in the substance use domain*. Washington, DC: American Psychological Association.