Global Information for Multidimensional Tests

Katherine G Jonas

Department of Psychological and Brain Sciences

University of Iowa

Author Note

The Memory and Aging Project data collection is supported by NIA grant R01AG17917.

Parts of these analyses were presented at the 2016 International Meeting of the Psychometric Society.

Correspondence may be addressed to Katherine G. Jonas, Stony Brook Health Sciences Center T10-060, 101 Nicholls Road, Stony Brook NY, 11794

email: katherine.jonas@stonybrookmedicine.edu

alternate email: katherine.grace.jonas@gmail.com

phone: 206.484.7010

Abstract

New measures of test information quantify information that could be gained by administering the test to an examinee of unknown ability—termed global test information. Currently, these measures have been developed only for unidimensional tests. This study develops measures of multidimensional global test information, validates them in data are simulated from models of varying dimensionality, and applies them in neuropsychological data collected as part of Rush University's Memory and Aging Project to identify the most informative tests of Alzheimer's disease. These measures allow for direct comparison of complex tests normed in different samples, facilitating test development and selection.

Keywords: test information, reliability, Lindley information, multidimensional item response theory

Global Information for Multidimensional Tests

Introduction

The development and selection of psychological tests is guided by test reliability. However, most measures of reliability are a function of the examinee or sample's ability. A test of algebra, for example, will be less reliable in a sample of elementary school students than a sample of high school students, not because the test is poorly constructed, but because of a mismatch of test difficulty and sample abilities. When test difficulty is low, the test cannot discriminate between two individuals of high ability. Because reliability is a function of test precision, reliability is low in these contexts, too. When test-sample mismatch is obvious, as in the above example, a more appropriate test is easily chosen. In many cases, however, test-sample mismatch is not obvious. In test development, the difficulty of the test is unknown. In an applied assessment setting, the ability of the examinee is unknown. Indeed, this lack of knowledge motivates the assessment.

Recently, measures of reliability have been developed that are solely a function of test parameters. These measures quantify the amount of information that could be gained by administering the test (Markon, 2013; Chang & Ying, 1996), and are referred to as measures of global information. As of yet, global test information has only been defined for unidimensional tests. However, many psychological tests are multidimensional, either intentionally or due to the presence of one of more nuisance traits. The Mini-Mental Status Exam, for example, reflects broad cognitive faculties but also academic achievement (Jones & Gallo, 2000, 2002). The former is relevant to the test's intended purpose as a screening tool for dementia, but the latter can introduce bias. In academic achievement testing, testlet models are often used to distinguish a general trait of interest such as verbal comprehension from knowledge of a specific reading passage. In both cases, it would be useful to decompose total test information into its component traits.

This study defines two measures of global information for multidimensional tests. Existing measures of global information are reviewed first, and then extended to

multidimensional tests. I also define marginal global test information—i.e. test information for a specific trait within a multi-trait test.

Global Test Information

Criterion Information Utility. Global test information quantifies how much administering a test could update the estimate of that person's abilities. This can be conceptualized as the difference between the prior and posterior estimates of the latent trait. Kullback-Leibler divergence, or information utility, is a measure of the difference between two probability distributions (Kullback & Leibler, 1951). When those two probability distributions are the *a priori* estimate of θ and the *a posteriori* estimate, the divergence is referred to as Lindley information (Lindley, 1956):

$$\iota_L(\theta|x) = \int_\theta p(\theta|x) ln \frac{p(\theta|x)}{\pi(\theta)} \tag{1}$$

where $\pi(\theta)$ is the probability distribution of the trait estimate θ prior to administering the test, and $p(\theta|x)$ is the same distribution after administering the test and collecting the data, x. In the multidimensional case, Lindley information is a function of the volume between two multidimensional probability distributions, $\pi(\boldsymbol{\theta})$ and and $p(\boldsymbol{\theta}|x)$.

Lindley information, like other measures of test information, is dependent on θ as well as the observed data x. To divorce information from a given person and test administration, one can define the prior, $\pi(\theta)$, as the distribution that reflects a complete *lack* of information about the trait. This prior can be conceptualized as the distribution that minimizes the informativeness of the prior, or conversely, maximizes the difference between the prior and posterior trait distribution. This prior is called the reference prior, $\pi_r(\theta)$ (Bernardo, 1979b; Berger, Bernardo, & Sun, 2009; Bernardo, 2005; B. S. Clarke & Barron, 1994). The reference prior is a function of test parameters. As a result, calculating Lindley information with regards to the reference prior frees it from the observed distribution of the trait. Numerical calculation of both unidimensional and

multidimensional reference priors is discussed in the study methods.

To calculate Lindley information independent of an observed dataset x, Lindley information is calculated for every possible dataset x, weighting the divergence by the probability of observing the dataset, and summed over all possible datasets. This quantity is called expected information utility (Bernardo, 1979a). When the prior is specified as the reference prior for the trait of interest, $\pi_r(\theta)$, the quantity is called criterion information utility (Markon, 2013):

$$\iota_c = \sum_{x}^{X} \left[p\Big(x|\pi_r(\theta) \Big) \times \iota_L(\theta|x) \right] \tag{2}$$

Since calculating the probability of every possible dataset is computationally intensive, the sum can be approximated by Monte Carlo methods (described in Appendix A).

Normalized Minimum Reduction in Uncertainty (NMRU). Criterion information utility has a natural upper and lower bound. The upper bound, ι_u, is defined by the entropy of the prior, because the test cannot convey more information than what is unknown about the trait (B. S. Clarke & Barron, 1994):

$$\iota_u = H[\pi(\theta)] = -\int \pi(\theta) \ln[\pi(\theta)] d\theta \tag{3}$$

Notably, because the reference prior maximizes missing information about a given trait, it also maximizes entropy and the upper limit of criterion information utility.

The lower bound of criterion information utility, ι_l, is a function of the area under the Fisher information curve:

$$\iota_l = \frac{1}{2} ln \left[\frac{1}{2\pi e} \right] + \ln \int \sqrt{\mathcal{I}(\theta)} d\theta \tag{4}$$

Where $\mathcal{I}(\theta)$ represents the Fisher information curve, and π is the constant. Fisher information is inversely proportional to the expected standard error of measurement (the expected SE is equal to the mean of $\mathcal{I}(\theta)^{-1/2}$),and proportional to the precision of the trait estimate of θ, meaning that the lower bound can be seen as a measure of the test's precision over the range of the trait.

The lower bound divided by the upper bound results in a scaled measure of global information called the normalized minimum reduction in uncertainty (NMRU Markon, 2013):

$$\text{NMRU} = \frac{\iota_l}{\iota_u} \tag{5}$$

NMRU quantifies the precision of a test relative to the entropy of the trait it measures. When NMRU is near 0, administering the test will not significantly update the trait distribution from what is described by the reference prior. To the extent that NMRU nears 1, administering the test will cause the trait distribution to become more peaked and narrow, as the test has conveyed more information about a person's trait standing.

Multidimensional Global Test Information

Multidimensional criterion information utility (md-ι_c). Unidimensional criterion information utility (Equation 2; from here forward called u-ι_c to distinguish it from multidimensional and marginal criterion information) is a function of two components: Lindley information and the probability of the data given the reference prior. Both can be extended to the multidimensional case. Multidimensional Lindley information is a function of the volume between two probability densities given by:

$$\text{md-}\iota_L(\boldsymbol{\theta}|x) = \int_{\theta_d} \dots \int_{\theta_1} p(\boldsymbol{\theta}|x) ln \frac{p(\boldsymbol{\theta}|x)}{\pi_r(\boldsymbol{\theta})} d\theta_1 \dots d\theta_d \tag{6}$$

Where $\pi_r(\boldsymbol{\theta})$ is the multidimensional reference prior for a vector of traits $\boldsymbol{\theta} = (\theta_1, \theta_2 \dots \theta_d)$. By extension, multidimensional criterion information is:

$$\text{md-}\iota_c = \sum_{x}^{X} \Big[p(x|\pi_r(\boldsymbol{\theta})) \times \text{md-}\iota_L(\boldsymbol{\theta}|x) \Big] \tag{7}$$

Multidimensional NMRU (md-NMRU). The components of unidimensional NMRU (u-NMRU), Equations 3 and 4, can also be extended to the multidimensional case. The upper bound of global information is the entropy of the multidimensional reference

prior:

$$\text{md-}\iota_u = H[\pi_r(\boldsymbol{\theta})] = -\int_{\theta_d} \dots \int_{\theta_1} \pi(\boldsymbol{\theta}) ln[\pi(\boldsymbol{\theta})] d\theta_1 \dots d\theta_d \tag{8}$$

And the lower bound is (Bodnar & Elster, 2014):

$$\text{md-}\iota_l = \frac{d}{2} ln\left[\frac{1}{2\pi e}\right] + ln \int_{\theta_d} \dots \int_{\theta_1} \sqrt{|\mathcal{I}(\boldsymbol{\theta})|} d\theta_1 \dots d\theta_d \tag{9}$$

where d is the number of dimensions in $\boldsymbol{\theta} = (\theta_1, \theta_2 \dots \theta_d)$. As in the unidimensional case, multidimensional NMRU (md-NMRU) equals the lower bound divided by the upper bound.

Marginal criterion information utility (m-ι_c). To calculate test information with regard to one of multiple traits, nuisance traits are integrated out of md-ι_c. The reference prior is defined with respect to the trait of interest, θ_k, by integrating over the traits $\boldsymbol{\theta}_{-k}$ (where the subscript -k indicates all traits except k), and multiplying by the marginal likelihood of the data given θ_k. This allows for the calculation of marginal Lindley information (m-ι_L):

$$\text{m-}\iota_L(\theta_k|x) = \int_{\theta_k} p(\theta_k|x) ln \frac{p(\theta_k|x)}{p(\theta_k)} d\theta_k \tag{10}$$

Which quantifies the extent to which the test data changes the estimate of θ_k. Marginal criterion information utility is then:

$$\text{m-}\iota_c = \sum_{x}^{X} \Big[p(x|\pi_r(\theta_k)) \times \text{m-}\iota_L(\theta_k|x) \Big] \tag{11}$$

When the traits $\boldsymbol{\theta} = (\theta_1, \theta_2 \dots \theta_d)$ are uncorrelated, the sum of the marginal informations m-ι_c will approximately equal md-ι_c.

Marginal NMRU (m-NMRU). The upper bound of Equation 11 is the entropy of the trait of interest, θ_k:

$$\text{m-}\iota_u = H[\pi_r(\theta_k)] = -\int_{\theta_k} \pi(\theta_k) ln[\pi(\theta_k)] d\theta_k \tag{12}$$

And the lower bound is (B. Clarke & Ghosal, 2010):

$$\text{m-}\iota_l = \frac{1}{2} ln \left[\frac{1}{2\pi e} \right] + ln \int_{\theta_k} \sqrt{|\mathcal{I}(\theta_k)|} d\theta_k \tag{13}$$

Where $\mathcal{I}(\theta_k)$ is the marginal distribution of the Fisher information matrix with respect to θ_k. Marginal NMRU is the ratio of m-ι_l to m-ι_u. As is the case with marginal criterion information, when traits $\boldsymbol{\theta} = (\theta_1, \theta_2 \dots \theta_d)$ are uncorrelated, the sum of marginal NMRU values equals the multidimensional NMRU. Note that because information is defined with regard to the marginal and not the conditional trait distribution, the values of marginal global information for two correlated traits will also be correlated.

The Present Study

The following analyses validate measures of unidimensional, multidimensional, and marginal global test information in simulated data. The first simulation compares traditional measures of reliability to global test information in cases of test-trait mismatch, and shows that where test-trait mismatch attenuates traditional measures of reliability, global information is constant. The second simulation shows the effect of model misspecification on global test information, and the sensitivity of these measures in tests of uncorrelated traits, as depicted in Figure 2a. The last simulation demonstrates the sensitivity of marginal and multidimensional global test information in tests with multiple, correlated traits, as depicted in Figure 2b.

Study 2 estimates marginal and multidimensional global test information for a number of common neuropsychological tests. Global test information is evaluated as a proxy for neuropsychological tests' criterion validity in the predicting a diagnosis of probable Alzheimer's Disease.

Methods

Study 1: Simulations

All analyses were performed in R (Team, 2010). For all simulations, test response data were simulated according to the 2-parameter logistic model (Birnbaum, 1968) for a sample of 500 respondents. Confirmatory models were fit using Metropolis-Hastings Robbins-Monro estimation in the R package `"mirt"` (Chalmers, 2012). Recovered item parameters were used to calculate a reference prior for each test, from which global test information was calculated. Each simulation condition was replicated 20 times.

Test-trait mismatch. Test-trait mismatch occurs when an individual's abilities are much lower or higher than the range assessed by the trait. This can occur when, for example, an examinee of unusually high or low ability takes an exam, or a patient with severe personality pathology is administered a test of normative personality function. In order to explore how a mismatch between test difficulty and sample ability affects test information, this simulation crossed three tests of varying difficulty with three samples of varying ability. A 16-item, unidimensional test was simulated. Item discriminations were generated from a truncated random normal distribution with a mean of 1.28 (equivalent to an item loading of 0.60 in a factor analytic model), variance of 0.50, and floor of 0.0. Item difficulties were drawn from a normal distribution with variance of 0.5, and means varying from -1.0, 0.0, and 1.0 across conditions. Respondents' abilities were randomly drawn from a normal distribution with variance of 0.5, and means varying from -1.0, 0.0, and 1.0 across conditions. The cross of test difficulty and sample ability resulted in a 3 x 3 design.

Uncorrelated traits, first-order test structure. This simulation tested the effect of item cross-loadings and model misspecification on global test information. A 16-item test was simulated in which the dimensionality of the test varied, with 0, 4, or 8 items cross-loading onto a second trait. The average magnitude of these cross-loadings also varied, from 0.54, 1.28, to 2.27 (or 0.3, 0.6, and 0.8 in a factor analytic parameterization), representing weak, moderate, and strong cross-loadings, respectively. Respondents' abilities

were randomly drawn from a multivariate normal distribution with mean 0.0, variance 1.0, and covariance of 0.0. The cross between the number of cross-loadings and their magnitude resulted in a 3 x 3 design. To test the effect of model misspecification, both a one-factor and two-factor model were fit to each data set.

Correlated traits, second-order test structure. In the second-order model, correlations among first-order traits are represented by a second-order, general trait. In this simulation, three first-order traits were simulated. Each first-order factor was measured by three 16-item tests. The test measuring the first specific trait varied as in the first-order simulation, with 0, 4, and 8 item cross-loadings on the second specific trait. The magnitude of cross-loadings were drawn from a truncated normal distribution with means varying from 0.54 to 1.28, to 2.27 (or 0.3, 0.6, and 0.8 in a factor analytic parameterization), with variance of 0.1 and a floor of 0.0. Item loadings for the second and third tests were also drawn from truncated normal distribution, with mean 1.28, variance 0.5, and floor of 0.0. The loading of the first specific trait onto the general trait varied from 0.54, 1.28, to 2.27 (or 0.3, 0.6, and 0.8 in a factor analytic parameterization), drawn from a truncated normal with variance 0.5 and a floor of 0.0. Loadings of the other two specific traits were drawn from a truncated normal with a mean of 1.28, variance 0.5, and floor of 0.0. Respondents' abilities were randomly drawn from a multivariate normal with means of 0.0, and variances and covariances defined by:

$$\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^T + \boldsymbol{\Psi} \tag{14}$$

Where $\boldsymbol{L}$ represents the vector of first-order factor loadings on the general factor, and Ψ is a diagonal matrix with diagonal elements equal to 1 minus the square of the factor loading. In other words, the covariance matrix of the respondents matched the covariance matrix of the first-order traits. In total, this yielded a 3 x 3 x 3 design.

Calculating reference priors

Calculating reference priors for single traits. The algorithm for estimating the reference prior begins from an arbitrary starting distribution of θ, from which a string of simulated test data, $\boldsymbol{x}$, is generated. The probability $p(\theta|\boldsymbol{x})$ is calculated using Bayes' theorem. Over many simulated samples, mimicking an infinitely long test, $p(\theta|\boldsymbol{x})$ approximates the reference prior, $\pi_r(\theta)$. For uncorrelated traits, the process is as follows (Berger et al., 2009):

1. Define starting values. This includes choosing the number of items to simulate, which is usually a multiple, k, of the number of items in the test. The multiple k is intended to approximate an infinitely long test. Berger, Bernardo and Sun (2009) simulate 500 items. In simulations, k was set to 50, approximating an 800 item test. The same k was used in analysis of the Memory and Aging Project (MAP) data. Because the number of items varied between tests in the MAP dataset, length ranged from 400 upward. Second, the number of samples to be simulated, m, is chosen. M was set to 1,000, which has been effective in a number of applications (Berger et al., 2009). Finally, a starting prior distribution prior was selected (a uniform prior over the range of the latent trait, $p(\theta) = 1$). The initial prior is arbitrary, as the repeated sampling of this procedure asymptotically reaches the reference prior.

2. For each simulation m, simulate response data for k replications of the test, for a given trait value θ. The likelihood of the data given theta, $p(x|\theta)$, is calculated, and divided by $p(x)$, the likelihood of the data integrated over the range of theta.

3. The probability $p(\theta)$ is the ratio in step 2 averaged over the m simulated samples.

4. Repeat steps 2 and 3 for all desired values of θ in order to obtain the reference prior, $\pi_r(\theta)$. For simulations, $\pi_r(\theta)$ was calculated at 60 points spanning the range between -10 and 10. Since tests in the MAP data set measured a wider range of abilities, $\pi_r(\theta)$ for these tests was calculated at 120 points between -20 to 20.

Calculating reference priors for correlated traits. When data depend on more than one correlated trait, calculation of the reference prior proceeded by calculating conditional reference priors, then integrating over the dimensions one by one. For example, in a second-order model, in which items load onto two correlated traits θ_1 and θ_2, the process is as follows (Bernardo, 2005):

1. The single-parameter algorithm above is used to calculate the conditional reference prior $\pi_r(\theta_2|\theta_1)$.

2. The conditional reference prior calculated in step 1 is used to integrate out the nuisance parameter, resulting in a 1-parameter model:

$$p(x|\theta_1) = \int_{\theta_2} p(\boldsymbol{x}|\theta_1, \theta_2)\pi_r(\theta_2|\theta_1)d\theta_2 \tag{15}$$

3. The single-parameter algorithm is applied to the model above to calculate the marginal reference prior $\pi_r(\theta_1)$.

4. The multidimensional reference prior $\pi_r(\theta_1, \theta_2)$ is equal to $\pi_r(\theta_2|\theta_1)$ (from step 1) multiplied by $\pi_r(\theta_1)$ (from step 3).

5. The other marginal reference prior $\pi_r(\theta_2)$ is then:

$$\pi_r(\theta_2) = p(\boldsymbol{x}|\theta_2)\pi_r(\theta_2) = \int_{\theta_1} p(\boldsymbol{x}|\theta_1, \theta_2)\pi_r(\theta_1, \theta_2)d\theta_1 \tag{16}$$

The order of conditioning parameters can be switched (i.e., calculating $\pi_r(\theta_1|\theta_2)$ in step 1 and integrating over θ_1 in step 2) without changing the form of the resulting prior. In the second-order simulations, the mean square difference between priors calculated in opposite directions (from specific factor 1 to specific factor 2, and vice versa) was 1.50e-06. This is small compared to the range of observed probabilities ($p(\theta) = 0.00 - 0.17$).

Treatment of Heywood Cases. When using Monte Carlo methods to calculate reference priors, it is possible to simulate response strings consisting entirely of 1s or 0s. In these cases, the estimated value of θ will be at the extreme upper or lower range over which the probability distribution is calculated. Over many replications, this causes the tails of the probability distribution to tilt upwards. To counter this artifact of the Monte Carlo process, prior distributions were smoothed such that the tail ends of the distribution were forced to asymptotically approach 0 beyond the outermost inflection points using the constrained spline-smoothing package `'cobs'` in R (Ng & Maechler, 2017).

Study 2: Neuropsychological Data

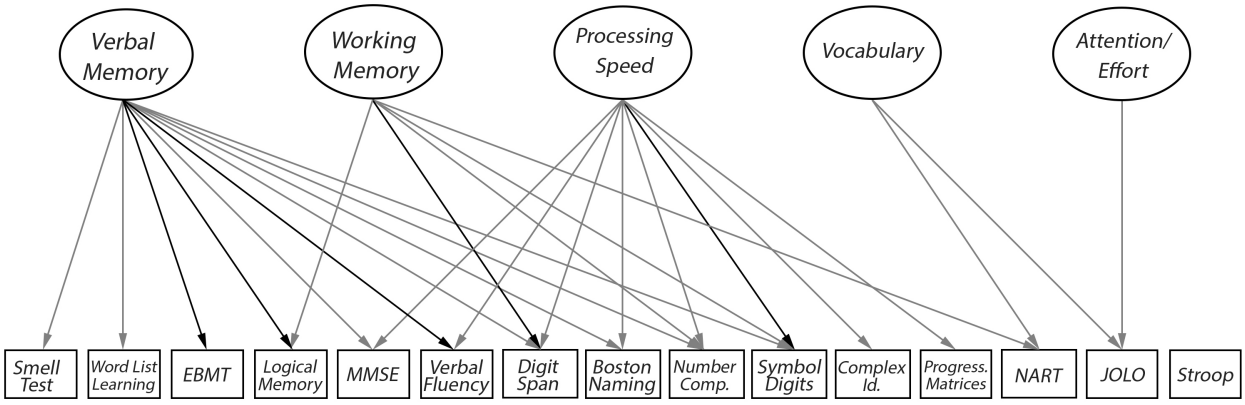
Sample. The Memory and Aging Project (MAP) is a longitudinal study funded by the National Institute of Aging, approved by the Institutional Review Board at Rush University Medical Center (Bennett et al., 2018). The study aims to identify factors that predict cognitive health in those over the age of 65, and to investigate the relationship of those predictors with Alzheimer's neuropathology at the time of death. Participants were recruited from retirement communities, churches, and social service agencies in Northeastern Illinois. Individuals were assessed at home to minimize attrition. The only exclusion criterion for participants was the inability to sign the Anatomical Gift Act, as the study includes the collection of brain and spinal cord tissue after death.

The MAP study is longitudinal, but the following analyses were performed on data from the baseline assessment collected between September 1997 and November 2011, as the sample was largest at this time point (n=1,489). Of those who completed baseline assessment, 73.1% were female and 87.8% were non-Hispanic white. The average age of participants was 80.1 years. Average years of education was 14.4. At baseline evaluation, 5.4% of the sample was considered to have some form of dementia. Diagnosis of probable Alzheimer's or Parkinson's Disease was arrived at by a computerized decision tree based on neuropsychological test scores, the result of which was used by both a neuropsychologist

and a clinician to make a final diagnosis.

The neuropsychological assessment comprised: the Mini Mental Status Examination (MMSE), East Boston Memory Test (EBMT; immediate and delayed recall), Logical Memory (one story, immediate and delayed recall), the Brief Smell Identification Test, Word List Memory (three immediate recall trials, delayed recall, and delayed recognition), Complex Ideational Material, Boston Naming Test (BNT; short form), Category Fluency (fruits, animals), National Adult Reading Test (NART), Digit Span Forward, Digit Span Backward, Digit Sequencing, Symbol Digit Modalities Test (SDMT), Number Comparison, The Stroop Test (word reading and color naming trials), Judgment of Line Orientation (JOLO; abbreviated 15-item version), and Standard Progressive Matrices. Descriptive statistics for each test are reported in Appendix B.

Figure 1. Structure of tests included in the Memory and Aging Project



Note. EBMT = East Boston Memory Test; MMSE = Mini Mental Status Exam; Number Comp. = Number Comparison; Complex Id. = Complex Ideation; Progress. Matrices = Standard Progressive Matrices; NART = National Adult Reading Test; JOLO = Judgment of Line Orientation. Arrow color reflects the number and magnitude of item loadings on factors. Light gray lines represent tests in which fewer than one-third of items load on a factor, or in the case of single-score tests such as the Symbol Digits Modalities Test, a loading less than 0.3. The Stroop Test had no significant loadings on any of the five factors.

Item-level response data were available for 13 of the 19 tests. These responses were scored as correct/incorrect, and modeled by the 2-parameter logistic model. For six tests,

only summary scores were available. These tests were: Logical Memory, the East Boston Memory Test, Verbal Fluency, Symbol Digit Modalities, Number Comparison, and the Stroop Test. Scores on these tests were modeled using the graded response model (Samejima, 1969). The GRM skews measures of test information because category response curves are constrained to have equal discrimination parameters. As a result, test information becomes a function of the *number* of possible scores than the informativeness of those scores. Therefore, while all tests were used to develop a structural model, global information was calculated only for tests with item-level response data.

Structural analyses. Exploratory multidimensional IRT models with 1 to 10 traits were estimated using the R package `"mirt"` (Chalmers, 2012). Model fit was assessed via the Bayesian Information Criterion (BIC), which indicated a five-factor model best fit the data. Model fit is reported in Appendix C. Four rotations of the five-factor model were estimated: a first-order uncorrelated traits rotation, a first-order correlated traits rotation, a second-order rotation, and a bifactor rotation. Since BIC does not differ as a function of rotation, the five-factor uncorrelated traits model was selected on the basis of correlations between factor scores and diagnoses, amount of variance accounted for by the general factor, as well as considerations related to accuracy. Figure 1 is a simplified depiction of the final model's latent structure.

Results

Study 1: Simulations

Test-trait mismatch. The effect of test-sample mismatch on traditional measures of test information is reported in the upper half of Table 1. Reliability was highest when the sample's abilities were matched to the difficulty of the test. Reliability decreased as the gap between difficulty and ability increased. In every case, the change in reliability was significant. The inverse pattern was found for expected standard error. The standard error of trait estimates was lowest when test difficulty matched sample ability, and grew larger as

the mismatch grew. All differences between conditions were significant.

The effect of test-sample mismatch on global information in reported in the lower half Table 1. Across all conditions, global information was remarkably stable. NMRU ranged from 0.31 to 0.32. Criterion information utility ranged between 1.29 and 1.31. None of the differences between conditions were were significant.

Uncorrelated traits, first-order test structure. Table 2 reports the effect of changing dimensionality and model misspecification on global information. Marginal global information for the primary trait (S1) remained consistent as the number and magnitude of cross loadings on the nuisance trait (S2) changed. Marginal global information for S2 approximately doubled as the number of cross-loadings doubled, and increased as the magnitude of those cross-loadings increased. As the number and magnitude of cross-loadings increased, so did multidimensional global information, reflecting the increase in total information that was gained as the test increasingly reflected a second trait.

If the confirmatory model was misspecified as a unidimensional test structure, global information trended lower, and changed very little across conditions. This is likely because when the model is misspecified, the trait that is estimated is a weighted average of the two traits the test actually measures, attenuating cross-loadings' effect on test information.

Correlated traits, second-order test structure. Table 3 reports global information for the general and specific traits in a second-order model as test dimensionality changes. Global information for the primary specific trait (S1) was constant across conditions, as anticipated. Global information for the nuisance trait (S2) increased significantly as both the number and magnitude of loadings on that factor increase, but not as the loading of S1 on G increased. Global information for the general trait (G) increased as the number of cross-loadings increased, reflecting the additional information about G that was provided via a better estimates of S2. In addition, global information for G increased as the magnitude of the loading of S1 on G increased, reflecting the closer relationship between the two traits. These increases were not significant, however. Changes

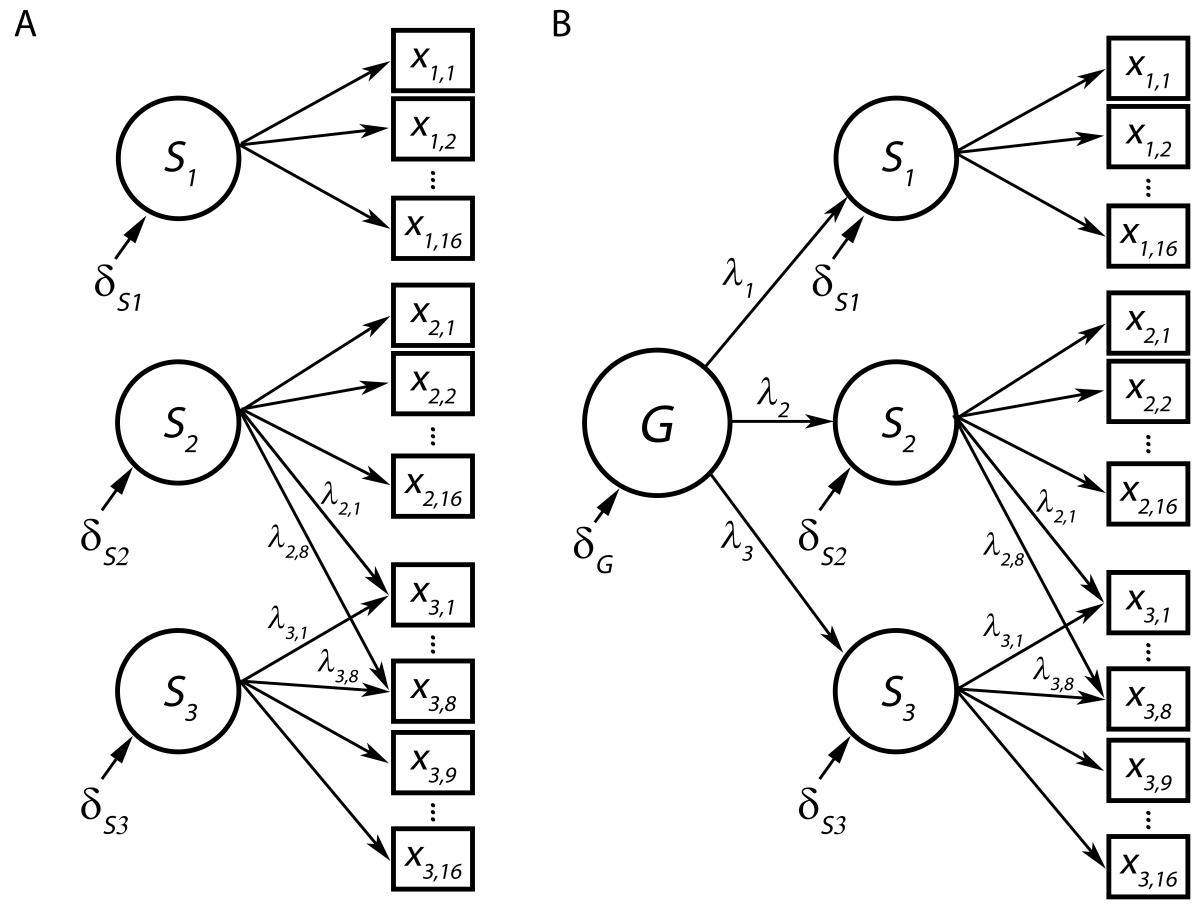


Figure 2. (A) The uncorrelated traits, first-order test structure, and the (B) correlated traits, second-order test structure. Note items cross-loading on a second specific trait. δ is the disturbance (i.e. unique variance) of of traits. λ represents item and trait loadings. Item loadings are only shown for traits and items that crossload, to reduce clutter. Item disturbances are omitted for the same reason.

in item and factor loadings were likely attenuated by the second-order test structure, leading to increases in the standard error of estimates of global information.

Multidimensional global test information increased with the number and magnitude of cross-loadings. This reflects the general increase in information available about both specific traits. Multidimensional information stays stable as the loading of S1 on G increases. This suggests that as the covariance between specific traits increases, the extent to which an estimate of one trait informs that of the other increases, but the total amount of information about both traits remains the same, as the added information about the

second trait is increasingly redundant with that of the first.

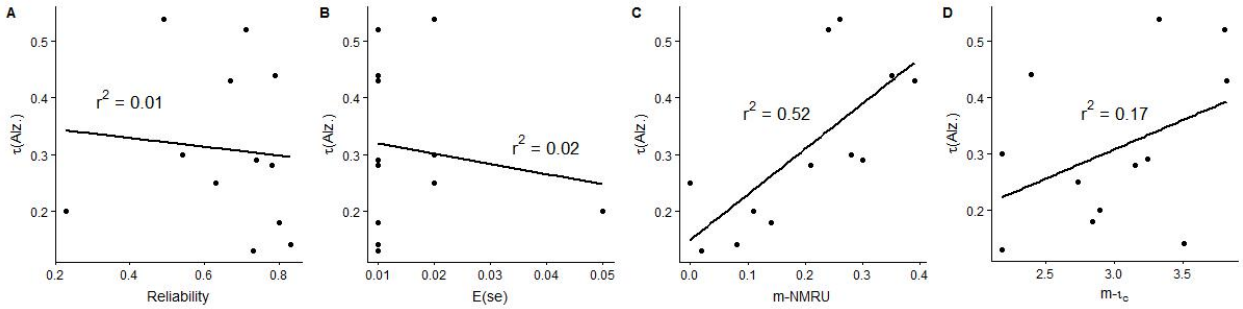
Study 2: Neuropsychological data

Table 4 reports the reliability and expected standard error for each test. In this sample Word List Learning, Judgment of Lines Orientation, and Digit Span were all highly reliable (Chronbach's $\alpha > 0.75$). Expected standard errors were universally low. The second two columns of Table 4 report marginal global test information calculated with regard to verbal abilities, as this trait was most closely associated with Alzheimer's disease. NMRU identifies the Mini Mental Status Exam and Word List Learning (immediate recall condition) as the best tests of verbal abilities (m-NMRU of 0.39 and 0.35, respectively). Criterion information utility is also high for the Mini Mental Status Exam ($\iota_c = 3.82$), as well as Word List Learning (delayed recall condition, $\iota_c = 3.81$). The last column of Table 4 reports the criterion validity of these tests, as reflected in Kendall's rank correlations with diagnoses of Alzheimer's. The best predictors of Alzheimer's were the Mini Mental Status Exam and various trials of Word List Learning.

Notably, the test most strongly correlated with Alzheimer's was Word List Recognition. Word List Recognition was one of the least reliable tests included in the battery (Chronbach's $\alpha > 0.49$), but did not have particularly high values of global information, either (m-NMRU = 0.26 and $\iota_c = 3.33$). Given the age of the Memory and Aging Project (MAP) sample, a reasonable hypothesis is that Word List Recognition tests a lower range of verbal abilities, and in a sample of older adults, performs better in this sample than it would on average. Unless one is very confident that a future examinee or sample will match the MAP sample in its verbal abilities, it is best to base test selection on sample-independent measures of test information. This is demonstrated by Figure 3, which includes four scatterplots demonstrating of criterion validity as a function of reliability (A), expected standard error(B), and marginal global test information (C and D). Marginal NMRU is most closely associated with criterion validity, explaining 50% of the variability

in criterion validity among neuropsychological tests.

Figure 3. Criterion validity as a function of test information



Note. The relationship between the correlation of test scores with probable Alzheimer's and (a) test reliability (b) expected standard error, (C) NMRU calculated with respect to verbal abilities, (D) ι_c calculated with respect to verbal abilities.

Discussion

The multidimensional nature of most psychological tests complicates test selection. These results show that multidimensional measures of global information accurately reflect test structure, and unlike other measures of test information, do not depend on the characteristics of the calibration sample.

Multidimensional global test information can aid test development by allowing for the quantification of test information for specific traits. Information specificity can be maximized by selecting items that maximize marginal global test information. Alternatively, test bias can be minimized by removing items that preferentially measure confounding traits, such as race or native language. While marginal global test information is useful for quantifying test specificity, multidimensional test information reflects broadband informativeness, and is therefore a useful measure of screening tests. Multidimensional test information may be useful for item selection at the outset of a computer adaptive test such as the SAT, for example, whereas marginal global test information may be useful as the test progresses, to select items to hone the estimate of a specific trait.

For clinicians, global information can inform test selection by providing general guidelines for settings in which there is little existing information about the examinee or sample. Marginal global test information can be used to identify tests that best assess traits of interest to the clinic—tests of verbal memory in a clinic that sees many patients with dementia, for example. Since traditional measures of reliability depend on both the sample and the test, use of these measures of test selection can be inefficient when a test is either too hard or too easy relative to the individual's or sample's abilities. However, because global information is a function of the test only, it is stable across samples. Global information can therefore provide guidance in establishing standard test batteries for the average patient.

The analysis of observed neuropsychological test data demonstrates how these measures might be useful in research and clinical assessment. Neuropsychological test data was shown to reflect five traits: verbal ability, sustained attention, executive function, crystallized verbal intelligence, and visuospatial perception. Verbal ability was most closely correlated with probable Alzheimer's. The Mini Mental Status Exam and Word List learning tasks had the greatest marginal global test information for this trait. These were also the tests with the strongest correlation with a diagnosis of probable Alzheimer's disease. Notably, these tests had relatively modest reliability coefficients. It has been noted that maximizing internal consistency does not necessarily improve test validity (Clark & Watson, 1995, 2019). These results suggest an alternative measure of reliability that is more closely associated with validity.

Limitations

The analyses of neuropsychological test data were limited by computational constraints, which required an uncorrelated trait model. This is almost certainly an incorrect model, although the high correlation between corresponding traits from different rotations (all correlations greater than 0.90) suggest the effects of rotation on the results

are minimal. Given continual increases in available computing power, it is likely that computing global test information for correlated traits will be much more feasible in the near future.

These analyses did not specifically investigate the extent to which global information varies due to error in parameter sampling error. Parameter sampling error depends on the specific IRT model (two-, three-, or four-parameter), the size of the calibration sample and length of the test (Hulin, Lissak, & Drasgow, 1982). Future research may reveal how parameter variability translates to variation in global test information.

Finally, like all simulation studies, these results cover only a fixed range of simulation conditions. For example, reference priors were calculated at 60 points along the interval from -10 to 10 standard deviations along the latent trait, in order to capture the extreme range that those tests measured while remaining computationally tractable. However, one could reasonably argue for either a narrower or wider range.

Conclusions

Marginal and multidimensional global information quantify how well a test can inform estimates of one or more traits in a multidimensional test. The development of these metrics will facilitate test development, in that items can be selected so as to maximize information about one or more traits of interest, and to minimize the effect of nuisance traits on responses. Applications to the assessment of probable dementia have been demonstrated, but the findings have applications to achievement testing, and the assessment of personality and psychopathology, too. Especially in contexts where prior information is vague, global information will identify those items or tests which are most likely to be the best measures of the trait of interest.

References

Bennett, D. A., Buchman, A. S., Boyle, P. A., Barnes, L. L., Wilson, R. S., & Schneider, J. A. (2018). Religious Orders Study and Rush Memory and Aging Project. *Journal of Alzheimer's disease: JAD*, *64*(s1), S161–S189. doi:10.3233/JAD-179939

Berger, J. O., Bernardo, J. M., & Sun, D. (2009, April). The formal definition of reference priors. *The Annals of Statistics*, *37*(2), 905–938. doi:10.1214/07-AOS587

Bernardo, J. M. (1979a, May). Expected information as expected utility. *The Annals of Statistics*, *7*(3), 686–690. doi:10.1214/aos/1176344689

Bernardo, J. M. (1979b, January). Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, *41*(2), 113–147. Retrieved March 4, 2015, from http://www.jstor.org/stable/2985028

Bernardo, J. M. (2005). Reference Analysis. In D. K. D. a. C. R. Rao (Ed.), *Handbook of Statistics* (Vol. 25, pp. 17–90). Bayesian Thinking Modeling and Computation. Elsevier. Retrieved March 19, 2015, from http://www.sciencedirect.com/science/article/pii/S0169716105250022

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In *Statistical theories of mental test scores* (pp. 395–479). Oxford, England: Addison-Wesley.

Bodnar, O. & Elster, C. (2014, April). Analytical derivation of the reference prior by sequential maximization of Shannon's mutual information in the multi-group parameter case. *Journal of Statistical Planning and Inference*, *147*, 106–116. doi:10.1016/j.jspi.2013.11.003

Chalmers, R. P. (2012). Mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, *48*(6), 1–29. Retrieved from http://www.jstatsoft.org/v48/i06/

Chang, H.-H. & Ying, Z. (1996, September). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, *20*(3), 213–229. doi:10.1177/014662169602000303

Clark, L. A. & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological assessment*, *7*(3), 309.

Clark, L. A. & Watson, D. (2019, March). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment*. doi:10.1037/pas0000626

Clarke, B. S. & Barron, A. R. (1994, August). Jeffreys' prior is asymptotically least favorable under entropy risk. *Journal of Statistical Planning and Inference*, *41*(1), 37–60. doi:10.1016/0378-3758(94)90153-8

Clarke, B. & Ghosal, S. (2010). Reference priors for exponential families with increasing dimension. *Electronic Journal of Statistics*, *4*, 737–780. doi:10.1214/10-EJS569

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two-and three-parameter logistic item characteristic curves: A Monte Carlo study. *Applied psychological measurement*, *6*(3), 249–260.

Jones, R. N. & Gallo, J. J. (2000, May). Dimensions of the Mini-Mental State Examination among community dwelling older adults. *Psychological Medicine*, *30*(03), 605–618. doi:null

Jones, R. N. & Gallo, J. J. (2002, November). Education and sex differences in the Mini-Mental State Examination: Effects of differential item functioning. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *57*(6), P548–P558. doi:10.1093/geronb/57.6.P548

Kullback, S. & Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, *22*(1), 79–86.

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, *27*(4), 986–1005. Retrieved May 4, 2017, from http://www.jstor.org/stable/2237191

Markon, K. E. (2013). Information utility: Quantifying the total psychometric information provided by a measure. *Psychological Methods*, *18*(1), 15–35. doi:10.1037/a0030638

Ng, P. T. & Maechler, M. (2017). *COBS – Constrained B-splines (Sparse matrix based).* Retrieved from https://CRAN.R-project.org/package=cobs

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, *34*(4, Pt. 2), 100.

Team, R. D. C. (2010). R: A Language and Environment for Statistical Computing. Vienna, Austria. Retrieved from http://www.R-project.org

Table 1
Test-trait mismatch

		Test Difficulty		
	Sample Ability	N(-1.0, 0.5)	N(0.0, 0.5)	N(1.0, 0.5)
Reliability (SD)	N(-1.0,0.5)	0.76* (0.04)	0.72* (0.04)	0.60* (0.05)
	N(0.0,0.5)	0.74* (0.04)	0.77* (0.03)	0.73* (0.04)
	N(1.0,0.5)	0.61* (0.06)	0.71* (0.03)	0.76* (0.03)
E(se) (SD)	N(-1.0,0.5)	0.62* (0.06)	0.81* (0.08)	1.19* (0.13)
	N(0.0,0.5)	0.56* (0.06)	0.51* (0.03)	0.56* (0.06)
	N(1.0,0.5)	1.17* (0.07)	0.81* (0.07)	0.63* (0.05)
ud-NMRU	N(-1.0,0.5)	0.32 (0.01)	0.32 (0.01)	0.32 (0.01)
	N(0.0,0.5)	0.32 (0.01)	0.31 (0.01)	0.32 (0.01)
	N(1.0,0.5)	0.32 (0.01)	0.31 (0.01)	0.31 (0.01)
ud-ι_c	N(-1.0,0.5)	1.31 (0.02)	1.31 (0.02)	1.29 (0.02)
	N(0.0,0.5)	1.29 (0.02)	1.30 (0.02)	1.30 (0.02)
	N(1.0,0.5)	1.29 (0.02)	1.30 (0.02)	1.30 (0.02)

Note. Test difficulty increases from left to right across columns. Sample ability increases from top to bottom down rows within the same information index. E(se) = expected standard error; ud-NMRU = unidimensional normalized minimum reduction of uncertainty; ud-ι_c = unidimensional criterion information utility. * difference between horizontally and vertically adjacent cells is significant at $p < 0.05$.

Table 2
Uncorrelated traits, first order structure

	No. of cross-loadings on S2			Magnitude of cross-loadings		
	0	4	8	0.3	0.6	0.8
S1-NMRU	0.33 (0.01)	0.33 (0.01)	0.34 (0.01)	0.33 (0.01)	0.33 (0.01)	0.34 (0.01)
S2-NMRU	NA	0.12* (0.04)	0.23* (0.03)	0.14* (0.06)	0.18* (0.06)	0.20 (0.06)
md-NMRU	NA	0.69* (0.04)	0.72* (0.04)	0.67* (0.03)	0.70* (0.02)	0.74* (0.03)
ud-NMRU	0.33 (0.01)	0.33 (0.01)	0.34* (0.01)	0.33 (0.01)	0.33 (0.01)	0.34 (0.01)
S1-ι_c	2.38 (0.02)	2.38 (0.02)	2.38 (0.02)	2.38 (0.03)	2.38 (0.02)	2.38 (0.02)
S2-ι_c	NA	1.79* (0.02)	2.06* (0.02)	1.86* (0.02)	1.94* (0.02)	1.98* (0.02)
md-ι_c	NA	4.01* (0.03)	4.15* (0.03)	4.02* (0.03)	4.09* (0.03)	4.13* (0.03)
ud-ι_c	2.40 (0.02)	2.40* (0.02)	2.42* (0.02)	2.40 (0.02)	2.40 (0.02)	2.41 (0.02)

Note. S1 = specific trait 1; S2 = specific trait 2; md = multidimensional; ud = unidimensional; NMRU = normalized minimum reduction of uncertainty; ι_c = criterion information utility. * difference between horizontally adjacent cells significant at $p < 0.05$.

Table 3
Correlated traits, second-order structure

	No. of cross-loadings on S2			Magnitude of cross-loadings			Loading of S1 on G		
	0	4	8	0.3	0.6	0.8	0.3	0.6	0.8
S1-NMRU	0.34 (0.01)	0.34 (0.01)	0.34 (0.01)	0.34 (0.01)	0.34 (0.01)	0.34 (0.01)	0.34 (0.01)	0.34 (0.01)	0.34 (0.01)
S2-NMRU	NA	0.11* (0.02)	0.21* (0.02)	0.14* (0.03)	0.17* (0.02)	0.18 (0.01)	0.17 (0.02)	0.16 (0.02)	0.16 (0.02)
G-NMRU	0.34 (0.09)	0.37 (0.07)	0.40 (0.06)	0.37 (0.06)	0.36 (0.09)	0.37 (0.07)	0.34 (0.08)	0.37 (0.07)	0.40 (0.06)
md-NMRU	NA	0.66* (0.02)	0.68* (0.02)	0.55* (0.02)	0.56* (0.01)	0.57* (0.02)	0.56 (0.01)	0.56 (0.02)	0.56 (0.02)
S1-ι_c	2.40 (0.02)	2.40 (0.02)	2.40 (0.02)	2.40 (0.02)	2.39 (0.02)	2.40 (0.02)	2.40 (0.02)	2.41 (0.02)	2.40 (0.02)
S2-ι_c	NA	1.87* (0.02)	2.09* (0.02)	1.91* (0.02)	2.00* (0.02)	2.03* (0.02)	1.99 (0.02)	1.98 (0.02)	1.97 (0.02)
G-ι_c	2.34* (0.02)	2.40* (0.02)	2.41* (0.02)	2.41* (0.02)	2.36* (0.02)	2.38* (0.02)	2.35* (0.02)	2.39* (0.02)	2.41* (0.02)
md-ι_c	NA	4.14* (0.03)	4.24* (0.03)	4.12* (0.03)	4.21* (0.03)	4.25* (0.03)	4.19 (0.03)	4.17 (0.03)	4.17 (0.03)

Note. S1 = specific trait 1; S2 = specific trait 2; G = second-order trait; md = multidimensional; ud = unidimensional; NMRU = normalized minimum reduction of uncertainty; ι_c = criterion information utility. * difference between horizontally adjacent cells significant at $p < 0.05$.

Table 4
Reliability, expected standard error, and Kendall's rank correlations of test scores with diagnosis of Alzheimer's

Test	Rel.	E(se)	m-NMRU	m-ι_c	τ(Alz.)
Mini Mental Status Exam	0.67	0.01	0.39	3.82	0.43*
Boston Naming Test	0.54	0.02	0.28	2.18	0.30*
Word List Immediate Recall	0.79	0.01	0.35	2.39	0.44*
Word List Delayed Recall	0.71	0.01	0.24	3.81	0.52*
Word List Recognition	0.49	0.02	0.26	3.33	0.54*
Judgement of Line Orientation	0.83	0.01	0.08	3.51	0.14*
Digit Span Forward	0.80	0.01	0.14	2.84	0.18*
Digit Span Backward	0.78	0.01	0.21	3.15	0.28*
Digit Span Sequencing	0.74	0.01	0.30	3.24	0.29*
Complex Ideation	0.23	0.05	0.11	2.89	0.20*
Progressive Matrices	0.63	0.02	0.00	2.73	0.25*
National Adult Reading Test	0.73	0.01	0.02	2.18	0.13*
Smell Test	0.65	0.02	0.25	2.31	0.28*

Note. Rel. = reliability; E(se) = expected standard error; m-NMRU = marginal normalized minimum reduction of uncertainty, calculated relative to Factor 1 of Figure 1; m-ι_c = criterion information utility, calculated relative to Factor 1 of Figure 1; τ(Alz.) = Kendall's rank correlation of test score with Alzheimer's; Alzheimer's is coded as 1 = highly probable, 2 = probable, 3 = possible, and 4 = not present. * denotes correlations significant at $p < 0.05$.

Appendices

Appendix A: Monte Carlo approximation of criterion information utility

Criterion information utility can be transformed in such a way as allow Monte Carlo approximation. The resulting approximation is (Markon, 2013):

$$\hat{\iota_c} = \frac{1}{M} \sum_{m=1}^{M} \ln \left[\frac{p(\theta_m | x_m)}{p(\theta_m)} \right] \tag{17}$$

Where probabilities are calculated with respect to the reference prior, and m indexes the iterations of the Monte Carlo procedure, which is:

1. Choose the number of iterations, m.
2. Randomly generate m values of θ from the reference prior.
3. For each θ_m, randomly generate a response pattern, $\boldsymbol{x}_m$.
4. It can be shown that via Bayes' theorem that the quantity inside the brackets in Equation 31 is equivalent to:

$$\frac{p(x_m | \theta_m)}{p(x_m)} \tag{18}$$

So the approximation to ι_c can be calculated by calculating the quantities in Equation 32 and averaging over the m replications. The standard error of ι_c is equal to:

$$\text{se}(\hat{\iota_c}) = \frac{\text{sd}(\hat{\iota_c})}{\sqrt{M}} \tag{19}$$

Appendix B: Descriptive statistics for neuropsychological tests

	Mean	SD	Median	Skew	Kurtosis	Range
Number Comparison	23.72	7.77	24	-0.29	3.17	0-46
Verbal Fluency, Animals	16.04	5.43	16	0.27	3.32	0-40
Verbal Fluency, Fruits	16.71	5.36	17	-0.12	3.17	0-40
Symbol Digit Modalities	36.65	11.83	38	-0.52	3.34	0-70
Stroop	31.28	11.83	32	-0.55	3.79	-19-70
Logical Memory, Immediate Recall	10.48	4.59	11	-0.18	2.55	0-23
Logical Memory, Delayed Recall	8.70	4.70	9	-0.01	2.36	0-23
East Boston Memory Test, Immediate Recall	9.33	2.14	10	-1.00	4.85	0-12
East Boston Memory Test, Delayed Recall	8.70	2.69	9	-1.52	5.76	0-12
Mini Mental Status Exam	25.38	3.38	26	-3.36	19.62	0-28
Boston Naming Test	13.49	2.29	14	-3.69	20.64	0-15
Word List Immediate Recall, Trial 1	3.75	1.76	4	0.13	2.93	0-10
Word List Immediate Recall, Trial 2	5.94	1.86	6	-0.36	3.27	0-10
Word List Immediate Recall, Trial 3	6.84	1.90	7	-0.70	3.81	0-10
Word List Delayed Recall	4.94	2.59	5	-0.33	2.38	0-10
Word List Delayed Recognition	9.07	2.13	10	-3.01	11.87	0-10
Judgement of Line Orientation	19.45	6.21	20	-0.96	4.14	0-30
Digit Span Forward	8.13	2.11	8	-0.35	3.29	0-12
Digit Span Backward	6.00	2.11	6	0.17	3.23	0-12
Digit Span Sequencing	6.94	1.91	7	-0.93	5.50	0-13
Complex Ideation	3.67	0.63	4	-2.45	10.94	0-4
Progressive Matrices	7.14	2.10	8	-1.55	5.27	0-9
National Adult Reading Test	7.56	2.68	8	-1.22	3.67	0-10
Smell Test	7.36	3.92	9	-0.87	2.41	0-12

Appendix C: Exploratory factor analyses of neuropsychological data

No. traits	BIC
1	353,227.1
2	351,879.9
3	349,763.0
4	349,248.0
5	**348,976.6**
6	349,019.5
7	349,218.9
8	349,761.2
9	35,0766.2
10	352,040.2

Note. BIC = Bayesian Information Criterion. Values in bold indicate the optimal model according to fit index.