Published in final edited form as:
J Clin Exp Neuropsychol. 2010 December ; 32(10): 1041–1049. doi:10.1080/13803391003662728.

Development of a Unidimensional Composite Measure of Neuropsychological Functioning in Older Cardiac Surgery Patients with Good Measurement Precision

Richard N. Jones, ScD[a,b,h,j,#], **James L. Rudolph, MD, SM**[c,e,j,#], **Sharon K Inouye, MD, MPH**[a,b,f,j], **Frances M. Yang, PhD**[a,b,h,j], **Tamara G. Fong, MD, PhD**[b,f,j], **William P. Milberg, PhD**[e,c,j], **Douglas Tommet, MS**[a,b], **Eran D. Metzger, MD**[a,j], **L. Adrienne Cupples, PhD**[l], and **Edward R. Marcantonio, MD, SM**[g,b,j]

[a] Institute for Aging Research, Hebrew SeniorLife, Boston, Massachusetts

[b] Aging Brain Center, Hebrew SeniorLife, Boston, Massachusetts

[c] Geriatric Research, Education, and Clinical Center, VA Boston Healthcare System, Boston, Massachusetts

[e] Division of Aging, Brigham and Women's Hospital, Boston, Massachusetts

[f] Department of Neurology, Beth Israel Deaconess Medical Center, Boston, Massachusetts

[g] Divisions of General Medicine and Primary Care, Beth Israel Deaconess Medical Center, Boston, Massachusetts

[h] Gerontology, Beth Israel Deaconess Medical Center, Boston, Massachusetts

[l] Department of Biostatistics, Boston University School of Public Health, Boston, Massachusetts

[j] Harvard Medical School, Boston, Massachusetts

Abstract

The objective of this analysis was to develop a measure of neuropsychological performance for cardiac surgery and assess its psychometric properties. Older patients (n=210) underwent a neuropsychological battery using nine assessments. The number of factors was identified with variable reduction methods. Item response theory-based factor analysis methods were used to evaluate the measure. Modified parallel analysis supported a single factor, and the battery formed an internally consistent set (coefficient alpha=0.82). The developed measure provided a reliable, continuous measure (reliability >0.90) across a broad range of performance (−1.5 SD units to +1.0 SD units) with minimal ceiling and floor effects.

Keywords

Cognition; cardiac surgery; aged; item response theory; neuropsychology

Corresponding Author (and reprints) James L Rudolph, MD, SM; VABHS GRECC 150 South Huntington Ave Boston MA 02130 Phone: 857 364-6812; Fax: 857 364-6812; jrudolph@partners.org.
[#]The first two authors contributed equally to this work and agreed to share first authorship.

Conflict of Interest:
The authors have no financial conflict of interest to declare

Introduction

Impairment in cognitive function following cardiac surgery was first reported shortly after the advent of cardiopulmonary bypass (Gilman, 1965). Since that time, improvements in surgical techniques, perfusion, anesthesia, and surgical teams have improved cognitive outcomes (Likosky, Nugent, & Ross, 2005). However, postoperative cognitive decline (POCD) remains a significant factor in cardiac surgery occurring in 21-29% at 3-6 months, 31-34% at 1 year, and 42-50% at 5 years (Newman et al., 2001; van Dijk et al., 2007). Despite its frequency, the methods for measuring POCD remain challenging because of a lack of standardization of measurement and definition.

There have been a number of challenges in the study of cognitive function after cardiac surgery. First, the neuropsychological tests used to assess cognitive function in cardiac surgery patients vary widely. In a recent review of POCD after non-cardiac surgery, 72 different neuropsychological tests were used across the literature to assess cognitive function.(Newman, Stygall, Hirani, Shaefi, & Maze, 2007) Second, many patients undergoing cardiac surgery have pre-existing cognitive impairments (Ernest et al., 2006; Rosengart et al., 2005), making them more susceptible to postoperative cognitive changes and increasing the potential for floor effects in cognitive testing. The floor effect is manifested by poor initial performance leading to limited/no possibility of decline. Additionally, insensitive neuropsychological tests have a ceiling effect where most patients perform at the maximum score and thus, differentiating between those who have modest impairments and those who are functioning normally is not possible. Finally, the neuropsychological tests scores are combined into a measure of cognitive decline that is unique to the study. Sometimes even a single study uses more than one measure; for instance, one randomized study of off-pump vs. on-pump surgery reported three methods for defining cognitive change (van Dijk et al., 2007). As a result, no standardized method of assessment or defining change in cognitive function after cardiac surgery currently exists.

A standard measure, or metric, for cognitive function would address these limitations. Neuropsychological testing for a cognitive composite provides more complete coverage of cognitive domains than a screening test and therefore, would be more sensitive to changes postoperatively. Secondly, because floor and ceiling effects are less likely to occur on multiple tests, the composite can describe cognitive function over a broader range of cognitive function than a single test. Additionally, the selection of a standard measure or metric for cognitive function would reduce variability across studies, because the battery of tests and the contribution of each test to the measure would become standardized.

When an intervention or procedure impacts a specific domain of neuropsychological functioning, it is essential that research and clinical attention focus on that specific domain. In general, domain-specific changes can be assessed by considering individual tests or creating composites of tests measuring the same ability.(Newman et al., 2001)

While not a substitute for study of individual neuropsychological domains, a global, unidimensional measure of cognitive functioning is also required in many contexts. First, many interventions, such as cardiac surgery, present many potential insults to cognition (e.g. hypotension, cardiopulmonary bypass, anesthesia, microemboli, hypothermia, etc), thus, it is unlikely that only a single cognitive domain will be affected, and a more general cognitive measure may be preferred to assess overall impact. Second, a unidimensional composite may be preferred to measure change over time, because of more favorable measurement characteristics (e.g. precision, diminished floor/ceiling effects) and minimizing multiple hypothesis testing across multiple domains. Finally, measurement of domain change in relation to global cognition may be more clinically relevant than raw domain change.

(Kanne, Balota, Storandt, McKeel, & Morris, 1998) Thus, a global undimensional composite may provide greater statistical power over evaluating domain specific tests individually, and may reduce risks of spurious inferences.

Another important role of a unidimensional composite is the comparison across studies. At present, thirteen randomized trials have studied the impact of cardiopulmonary bypass on POCD; these studies use a median of 9 neuropsychological tests (range 6-19) and 9 different definitions of POCD. This makes comparison across studies difficult. A global composite may also be a useful and formal way to compare or jointly analyze multiple studies by co-calibrating the composite on a common metric(Crane et al., 2008), even when some of the component neuropsychological tests differ between studies.

To address these issues, we undertook a systematic approach to developing a composite measure of cognitive function using common neuropsychological tests collected as part of a prospective observational study of cognitive outcomes following cardiac surgery (Rudolph et al., 2009). We hypothesized that a) a single, summary measure of neuropsychological performance would be adequate to explain covariation among selected neuropsychological tests, and b) the psychometric properties of the summary measure would be conducive to studying cognitive change with limited floor and ceiling effects. In addition, to improve the ability to study neuropsychological performance across studies, we provide the needed details for creating the neuropsychological composite measure, upon request. Investigators may use this methodology to create comparable scores using somewhat different neuropsychological test batteries and can aggregate results across studies to facilitate cross-study comparisons. Thus, this methodology may represent a substantial advance in the study of the neuropsychological sequelae of cardiac surgery and potentially, in the study of longitudinal cognitive function assessed with neuropsychological measures.

Methods

Participants

The data for this study were collected from patients enrolled in a prospective, observational cohort including patients (targeted age ≥60 years) who were undergoing coronary artery bypass graft (CABG) surgery or combined CABG-valve replacement surgery at two academic medical centers and one Department of Veterans Affairs medical center in Massachusetts. Institutional review boards at the three medical centers approved the study and all patients provided written informed consent. Four hundred and sixty-one adults were screened for participation in the study, 200 refused, 17 were excluded due to additional cardiac procedures, delirium prior to surgery, or inability to complete the preoperative assessment, and 34 were excluded for missing data on more than 2 of 9 neuropsychological tests included in the composite. The analytic sample therefore included 210 adults.

Neuropsychological Assessment

Preoperatively, a 45-minute neuropsychological battery was administered to patients. The battery assessed cognitive domains of memory, learning, attention and executive functioning in accordance with the Statement of Consensus on Assessment of Neurobehavioral Outcomes after Cardiac Surgery (Murkin, Newman, Stump, & Blumenthal, 1995). The Hopkins Verbal Learning Test (HVLT) (Brandt & Benedict, 1991), a 12-item verbal learning and recall measure was administered. The retention percent was calculated as the number of spontaneously recalled items divided by the maximum number of items learned. The Visual Search and Attention Task required patients to identify a specific letter amidst a field of distracting letters and was scored on the number of targets identified in 90 seconds (Trenerry, Crosson, DeBoe, & Leber, 1990). The Trailmaking Test B (*Trail Making Tests A*

and B, 1944), a timed test of shifting attention, involved alternating between a series of numbers and letters. The Digit Symbol Substitution Test is a measure of working memory and attention, where digits were represented by symbols and patients copied a symbol to the corresponding number, and Digit Symbol Copy is a measure where patients copied the symbol. In both tests, the number of correct responses in 90 seconds was recorded. Digit span forward and backward (Wechsler, 1981), tests of working memory, required participants to sustain attention and manipulate information by repeating a series of random digits forward and backwards (Stuss & Levine, 2002), The measure was the number of correct trials for each task. Semantic (category) and phonemic (letter) fluency tasks measured language and knowledge storage patterns by requiring the subject to generate words spontaneously in a category (animals and boys names) or beginning with a specific letter ('f', 'a', and 's') (Benton & Hamsher, 1976). For both, the measure was the number of correct responses. The 30-item Boston Naming Test, a measure of naming, where patients identify line drawings of increasing difficulty, was administered and the number of correct responses was scored (Mack, Freed, Williams, & Henderson, 1992).

Statistical Approach

Overview—The statistical methods to develop the single factor model measure of cognitive performance followed several steps; each step is introduced here and described in more detail later. First, we organized the data into consistent scales across neuropsychological tests using decile cut points. We used parallel analysis to determine the number of factors (Hayton, Allen, & Scarpello, 2004). We estimated parameters of a measurement model using confirmatory factor analysis (CFA) that was consistent with a graded response item response theory (IRT) model and assessed the assumption of local independence. Finally, post-hoc Bayesian methods were utilized to estimate the latent neuropsychological composite scores for individual participants. Univariable and multivariable models were fit using Stata 10 (Stata, Inc. College Station, TX) and multivariate measurement models were fit using Mplus 5.1 (Muthén & Muthén, Los Angeles, CA). We used PARSCALE 4.1 (Scientific Software International, Chicago, IL) to generate test information functions.

Basing the measurement model in IRT offers several advantages. First, and perhaps most importantly, IRT provides a framework to estimate cognitive ability that can be used in the current and future studies; this is superior to factor analysis where the estimates are sample-dependent. Moreover, the IRT ability estimates theoretically have interval scale properties (Stevens, 1946), instead of ordinal scale properties, and thereby makes IRT suitable for analyses of change over time. Finally, IRT allows for rigorous assessment of scale reliability can define regions of variability and thresholds for clinically significant change.

Step 1: Create similar scales between neuropsychological tests—Each neuropsychological test is measured on a scale that may be particular to that individual test. For example, Trailmaking B is a timed test, HVLT is a percentage, and Digit Symbol Substitution Test is scored as the number of correct responses. Additionally, neuropsychological data are often skewed. To address these factors, neuropsychological data were categorized into discrete categories using decile cut points. This provides an efficient tool for normalizing the distributions. If deciles could not be defined because of severe skew, we used the maximum number of categories allowed by the observed data. Deciles were reversed prior to modeling for Trailmaking B so that all measures were in a consistent direction (i.e. a higher decile representing better performance). Missing data were assumed to be missing at random, and all observations were included in analyses using maximum likelihood techniques (Muthen & Kaplan, 1987).

Step 2. Parallel analysis to determine the number of latent factors—Parallel analysis is a robust method for determining the number of latent factors that may account for the covariation among a set of observed variables (Horn, 1965; Kiecolt-Glaser et al., 2003; Zwick & Velicer, 1986). It involves obtaining random data by resampling the observed data set multiple times, computing a correlation matrix and eigenvalues for each random data set, and plotting the 90% confidence interval for the random eigenvalues and the observed eigenvalues. If an observed eigenvalue is less than the 95th percentile of the random data, we conclude that the value of the next eigenvalue is not greater than what would have been observed by chance. The number of eigenvalues that lie above the 95th percentile of random data eigenvalues is the number of factors supported.

Step 3. Estimate a measurement model—We used CFA on the estimated polychoric correlation matrix to estimate a measurement model for the battery. Polychoric correlations represent correlations among ordinal observed variables that are on the scale of Pearson correlation coefficients. CFA on a polychoric correlation matrix is equivalent to a graded response IRT model (Jöreskog & Moustaki, 2001; Mislevy, 1986). Typical IRT models assume unidimensionality, or that a single common factor is sufficient to account for the correlation among the test items. We explore violations of this assumption with a bifactor model. A bifactor model is a measurement model where every indicator is caused by at most two underlying factors, one being a general factor loading in all indicators(McDonald, 1999). Bifactor models, with some restrictions, are equivalent to hierarchical factor models (Chen, West, & Sousa, 2006), but can be more useful in assessing unidimensionality and local independence assumptions. Models were estimated with Mplus software using a limited information multivariate probit regression framework and the mean and variance adjusted weighted least squares estimator (Muthen & Kaplan, 1987). Model fit was assessed with the root mean square error of approximation (range 0-1.0; 0=perfect model fit; 0.06-0.1=acceptable model fit; >0.1 unacceptable model fit (Browne & Cudeck, 1993; Hu & Bentler, 1998)) and the comparative fit index (CFI, range 0-1.0; acceptable model fit ≥0.95 (Bentler & Chou, 1988; Bentler, 1990; Muthen & Kaplan, 1987)).

Step 4: Assess neuropsychological composite reliability and performance—Another benefit of IRT is the ability to measure performance over a larger range of cognitive ability compared to individual neuropsychological tests, which frequently have floor and ceiling effects and demonstrate ordinal, rather than interval scaling properties. We assessed the reliability of the scale using the classical test theory based internal consistency coefficient using Cronbach's alpha (range of 0-1; acceptable reliability: ≥0.80 for group differences and ≥0.90 for individual inferences) (Nunnally & Bernstein, 1994). We also used an IRT concept known as item and scale information which is a method of estimating of the amount of information provided by a scale over the range of cognitive ability. We used the Edwards-Nunnally reliable change method to compute reliable change indices (Atkins, Bedics, McGlinchey, & Beauchaine, 2005; Speer, 1992) and used the exaggerated imprecision method of the reliable change method to account for regression to the mean. The Edwards-Nunnally method centers confidence intervals (±2 standard errors) on the estimated cognitive composite score and considers performance outside the confidence interval as significant change (Speer, 1992).

Results

The cardiac surgery sample (Table 1) represented predominantly older patients (mean age 73 ±7 years) who were well educated (53% >high school education). Consistent with national data (Rosamond et al., 2007), 24% were women. Our sample was racially homogenous; only 5% self-described as a race or ethnicity group other than white. About 1

in 5 had a Geriatric Depression Scale (15-item version) score of ≥5 which is indicative of clinically relevant depressive symptoms (Zalsman, Weizman, Carel, & Aizenberg, 2001).

Scale Neuropsychological Data

The baseline performances on the 10 neuropsychological tests included in this analysis are presented in Table 2 (top panel), and item correlation matrices are also presented (lower diagonal, Pearson correlation coefficients for raw variables, upper diagonal polychoric correlation coefficients for discrete versions). As can be seen, the tests varied in the proportion of individuals completing each task primarily because patients declined additional testing due to fatigue, medical illness, or duration of testing. No tests, except category fluency and digit symbol substitution were normally distributed ($p<.05$) (D'Agostino, Belanger, & D'Agostino Jr, 1990;Royston, 2005) and thus, the decile transformation was justified. The coefficient of internal consistency (Cronbach's alpha) suggested the battery formed an internally consistent set ($\alpha = 0.82$), which is sufficient for group differences research but not sufficiently reliable for individual differences research (Nunnally & Bernstein, 1994).

Parallel Analysis

The first and second eigenvalues extracted from the polychoric correlation matrix were 4.21 and 1.28, respectively. The 95^{th} percentile of the first and second random eigenvalues was 1.51 and 1.34, respectively. Because the second eigenvalue from the observed data falls below the 95^{th} percentile of the random data second eigenvalue, we concluded that the second eigenvalue provided no additional information, which could not be expected by chance alone. Thus, the parallel analysis supported a single neuropsychological composite.

Confirmatory Factor Analysis

Despite the findings of the parallel analysis, the root mean square error of approximation (RMSEA) for the single factor model was 0.142, and the CFI was 0.887, both values outside the accepted range for well fitting models (Table 3). To better explore the covariation in the observed data, we used residuals from the single factor model to postulate a specific bifactor CFA model that fit better (RMSEA=0.062, CFI=0.980). The factor loadings of the single and bifactor models are displayed in Table 3 (right side). Bifactor CFA models can be assumed to support sufficient unidimensionality if the item loadings on the general factor (loading in all items) are greater than the item loadings on the specific factors loadings(McDonald, 1999). For our bifactor CFA model, three pairs of related tests had significant residual correlation (digit symbol copy and substitution, digit span forwards and backwards, and phonemic and semantic fluency). Only digit span forwards and backwards failed this test of sufficient unidimensionality as a pair. Phonemic fluency was loaded essentially equally on the specific and general factor.

At this point, we would be justified in considering whether the inclusion of both digit span forwards and backwards was necessary, or if one could be dropped from the task list (Cattell & Cattell, 1960; Ozer & Reise, 1994). Substantive concerns motivated our retention of both digit span forwards and backwards, including the ease/timing of administration, normal distribution of results, and coverage of the working memory and attention domains. To evaluate the bias in estimated cognitive ability caused by ignoring the multidimensional structure, we compared estimated scores from the unidimensional model and the global ability score from the bifactor model. These estimates were highly correlated (r=0.99), an expected finding and a reflection of the fact that the factor loadings in the unidimensional and for the bifactor-general factor were very similar. This correlation coefficient is not informative about the suitability of a unidimensional or multidimensional measurement model (Reise & Haviland, 2005). As a guide to methodological judgment on this issue, we

examined the scale reliability as implied by the two measurement models and the individual-level bias in latent trait estimation associated with using the overly simplified unidimensional model. We estimated scale reliability based on the factor loadings (Brown, 2006) for the unidimensional factor model ($\rho = 0.83$) and the bifactor model ($\rho = 0.85$), and conclude that impact of the local independence violations on score reliability are trivial. To evaluate individual-level impact, we calculated bias in the unidimensional trait estimate given the multidimensional structure as the difference between the factor score estimate under the unidimensional model and the bifactor model. For a substantial sub-set of the participants (n=57, 25%) the bias was more than trivial (at least |0.2| standard deviation units). The magnitude of bias was related to performance on the digit symbol and trails tasks, but not the digit span tasks. Therefore, we concluded the multidimensionality was ignorable and use a unidimensional measurement model to generate latent trait estimates for the participants.

Assess Performance and Reliability

The distribution of the estimated factor score for the unidimensional neuropsychological composite is presented in Figure 1. We rescaled the factor score to a T-score metric (mean= 50; standard deviation=±10; range=25-75). The neuropsychological composite was normally distributed (p=0.88 for test of deviation from normality) (D'Agostino et al., 1990;Royston, 2005) there was no floor or ceiling in this sample (i.e., no participant performed at the lowest or highest level on all 9 tests). To further gain an appreciation of measurement precision, we estimated item and battery information functions. In the range of 1 SD unit above and below the mean, the neuropsychological performance battery provided excellent measurement precision. The reliability index at this level was at least 0.90 in this ability range which is sufficiently reliable to make inferences of individuals (Nunnally & Bernstein, 1994). The Edwards-Nunnally reliable change indices were calculated to measure the performance of the neuropsychological composite across a range of cognitive ability and reliable change regions are illustrated in Figure 2. Note that reliable change ranges from about 0.5 SD units to 1.5 SD units across the range of the latent trait. In the cognitive score range of 40-60 (67% of subjects), we could reliably measure a decline in cognitive function of 0.5-0.8 SD units.

Model Parameters and Code

Upon request we will provide the specific details of our methodology and code to enable other investigators to extrapolate our methods to other neuropsychological test batteries. Our analyses were conducted with STATA (v10) and Mplus (v5.1). In addition, we have prepared R code for the generation of latent trait estimates using the expected a posteriori method (Bock & Aitkin, 1981) given similarly collected and scored neuropsychological testing data.

Discussion

In this study, we used data from a neuropsychological assessment of 210 patients undergoing cardiac surgery to develop a single measure of cognitive performance. We presented evidence that the measure represents a general cognitive domain with high internal consistency and is relatively free from floor and ceiling effects. To address some of the past challenges in diagnosing cognitive decline after cardiac surgery, we have made the computational algorithm available upon request, so that it can be applied to other data and to stimulate and facilitate the comparison of cognitive data after cardiac surgery across studies.

Factor analysis has been used frequently to create neuropsychological composite scores in patients (Newman et al., 2001; van Dijk et al., 2007). By using IRT to derive the

neuropsychological composite, we address three major concerns about current methods to assess cognitive function 1) floor and ceiling effects, 2) reliable and precise measurement performance over a range of abilities, and 3) comparison of performance among different studies. The neuropsychological composite we calculated demonstrated no ceiling or floor effect and it reliably measured cognitive performance from 1.5 SD below the mean to 1.0 SD above the mean. Ultimately, this neuropsychological composite can be used by other researchers to directly compare results as described below.

The publication of the decile thresholds, item parameters, and programming code allows other researchers to construct a similar neuropsychological composite with their data. Such IRT model comparisons are commonly used in educational testing to compare performance on different versions of a test (McHorney, 2003). The open presentation of our model provides not a definitive solution, but rather an initial step on which we can build a common neuropsychological measure that can address change over time in patients undergoing surgery and in other clinical settings.

Ultimately, with this approach the problem of different assessment batteries becomes a statistical nuisance that can be addressed analytically rather than representing an intractable problem that prevents direct comparison of studies. For this comparison to be possible, at least one identical neuropsychological test needs to be performed within the battery (but the more tests that overlap the greater the confidence in the co-calibration). Common neuropsychological tests provide a basis to build a set of calibration models that could be used to directly compare results of different studies. In the study of POCD, where time before surgery is limited and brevity is essential, investigators may prefer a shorter battery than what we have used in this study. Other investigators may not be satisfied with our reliability estimates and prefer longer batteries. Different batteries may still be linked to our metric so long as at least one common test appears across studies and appropriate modeling constraints are applied. The development of a bank of batteries, decile thresholds, item parameters, and programming code allows expansion of the neuropsychological composite to different populations and in different settings. Ultimately, clinicians could administer a core battery and define a consistent level of cognitive performance across many different clinical settings.

We recognize several limitations of our work. First, this study has a limited sample size. It is possible that if this study had been repeated in a larger sample, different parameter estimates would have been obtained, presumably closer to true population values. We view the neuropsychological composite as an initial solution, the value of which will grow when additional investigators use our results to build composites in other studies of post operative cognitive decline and link in overlapping and non-overlapping neuropsychological performance data. Second, while our single measure displayed good precision, reliable change indices, and was free from floor and ceiling effects, it is important to realize that our neuropsychological composite may be useful in some but not all circumstances. Limitations of composites may arise when the health status change or sub-group differences can be expected to have specific effects on specific aspects of the composite (e.g., administration of a anticholinergic medication impacts memory function (Kay & Granville, 2005)) or the population subgroup is known to have a specific impairment on one aspect of the domain (e.g., those with low education exhibit selective impairment on tests involving computation). These possibilities should be evaluated using established methods for assessing measurement non-invariance or tested formally with experimental or analytic strategies. Third, previous studies have demonstrated some degree of preoperative impairment in the cardiac surgery population (Ernest et al., 2006; Rosengart et al., 2005) which might limit the generalizability of our single measure to an unimpaired population. On the other hand, this baseline impairment may also exist in other medical and surgical samples. Fourth, our study

was limited by the relative homogeneity of the sample with respect to gender and race/ethnicity. However, our sample is distinguished by having one of the highest mean ages of all published cardiac surgery cohorts to date. Finally, the derivation of our measure used only preoperative data, and further study is required to determine our measure's ability to detect change in cognitive function after cardiac surgery.

This work has strengths that deserve comment. First, our battery was chosen to be consistent with the Statement of Consensus on Neurobehavioral Outcomes after Cardiac Surgery, which provides for comparability with related studies (Murkin et al., 1995). Second, while there is some missing neuropsychological data, most tests have >90% completion in our data. We imposed a limit of at most 2 missing items to be included in our analysis, but such a stringent requirement is not necessary from a statistical or measurement point of view under the assumption of missing at random (i.e., the mechanism causing the missingness is not related to the value that would have been observed on the test if it were observed). Finally, the inclusion of the item parameters and software (R syntax) allows other researchers to compare their summary scores to our work.

The measurement of perioperative cognitive function has been characterized by disparate methods for the assessment, scoring, and definition of POCD. This study addresses many of these limitations by developing a single neuropsychological composite which has limited floor and ceiling effects and good reliability across a range of cognitive function. More importantly, we will provide the needed model parameters to allow other investigators to directly compare results from their work to ours (upon request). The methodology developed here may enable statistical comparisons of a variety of neuropsychological test batteries across studies. Future work can build on the foundation provided by this study to greatly advance the understanding of cognitive function after cardiac surgery.

Acknowledgments

The authors acknowledge the contributions of Dr. David Alsop to this paper.

The authors report no financial conflict of interest.

This work was funded by the Harvard Older Americans Independence Center AG08812-14 (ERM, RNJ), R03 AG029861 (JLR), K24 AG00949 (SKI) R21AG025193 (SKI), R21 AG027549 (ERM), R21 AG026566 (ERM), R01 AG 030618 (ERM), R03 AG028189 (ERM),. Dr. Rudolph is supported by a VA Rehabilitation Career Development Award. Dr. Inouye is supported by the Milton and Shirley F. Levy Family Chair.

References

Atkins DC, Bedics JD, McGlinchey JB, Beauchaine TP. Assessing Clinical Significance: Does It Matter Which Method We Use? Journal of Consulting & Clinical Psychology 2005;73(5):982. [PubMed: 16287398]

Bentler, P.; Chou, C. Practical issues in structural equation modeling. In: Long, J., editor. Common problems/proper solutions: Avoiding error in quantitative research. Sage; Newbury Park, CA: 1988.

Bentler PM. Comparative fit indexes in structural models. Psychological Bulletin 1990;107(2):238–246. [PubMed: 2320703]

Benton, A.; Hamsher, K. Multilingual Aphasia Examination. University of Iowa; Iowa City: 1976.

Bock RD, Aitkin M. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika 1981;46(4):443–459.

Brandt, J.; Benedict, RHB. Hopkins Verbal Learning Test-Revised (HVLT). Psychological Assessment Resources, Inc.; Lutz, FL: 1991.

Brown, TA. Confirmatory Factor Analysis for Applied Research. Guilford Publications; 2006.

Browne, M.; Cudeck, R. Alternative ways of assessing model fit. In: Bollen, K.; Long, J., editors. Testing structural equation models. Sage; Thousand Oaks, CA: 1993. p. 136-162.

Cattell, RB.; Cattell, AKS. The individual or group culture fair intellegence test. University of Illinois; Champaign, IL: 1960.

Chen FF, West SG, Sousa KH. A comparison of bifactor and second-order models of quality of life. Multivariate Behavioral Research 2006;41(2):189–225.

Crane PK, Narasimhalu K, Gibbons LE, Mungas DM, Haneuse S, Larson EB, et al. Item response theory facilitated cocalibrating cognitive tests and reduced bias in estimated rates of decline. Journal of Clinical Epidemiology 2008;61(10):1018–1027. e1019. [PubMed: 18455909]

D'Agostino RB, Belanger A, D'Agostino RB Jr. A suggestion for using powerful and informative tests of normality. The American Statistician 1990;44(4):316–321.

Ernest CS, Murphy BM, Worcester MU, Higgins RO, Elliott PC, Goble AJ, et al. Cognitive function in candidates for coronary artery bypass graft surgery. Annals of Thoracic Surgery 2006;82(3): 812–818. [PubMed: 16928490]

Gilman S. Cerebral Disorders after Open-Heart Operations. New England Journal of Medicine 1965;272:489–498. [PubMed: 14250198]

Hayton JC, Allen DG, Scarpello V. Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. Organizational Research Methods 2004;7(2):191.

Horn JL. A rationale and test for the number of factors in factor analysis. Psychometrika 1965;30(2): 179–185. [PubMed: 14306381]

Hu L, Bentler P. Fit indices in covariance structure analysis: Sensitivity to underparameterized model misspecifications. Psychological Methods 1998;4:424–453.

Jöreskog KG, Moustaki I. Factor analysis of ordinal variables: A comparison of three approaches. Multivariate Behavioral Research 2001;36(3):347–387.

Kanne SM, Balota DA, Storandt M, McKeel DW Jr. Morris JC. Relating anatomy to function in Alzheimer's disease: neuropsychological profiles predict regional neuropathology 5 years later. Neurology 1998;50(4):979–985. [PubMed: 9566382]

Kay GG, Granville LJ. Antimuscarinic agents: implications and concerns in the management of overactive bladder in the elderly. Clinical Therapeutics 2005;27(1):127–138. quiz 139-140. [PubMed: 15763613]

Kiecolt-Glaser JK, Preacher KJ, MacCallum RC, Atkinson C, Malarkey WB, Glaser R. Chronic stress and age-related increases in the proinflammatory cytokine IL-6. Proceedings of the National Academy of Science, U S A 2003;100(15):9090–9095.

Likosky DS, Nugent WC, Ross CS. Improving outcomes of cardiac surgery through cooperative efforts: the northern new England experience. Seminars in Cardiothoracic Vascular Anesthesia 2005;9(2):119–121.

Mack WJ, Freed DM, Williams BW, Henderson VW. Boston Naming Test: shortened versions for use in Alzheimer's disease. Journal of Gerontology 1992;47(3):P154–158. [PubMed: 1573197]

McDonald, RP. Test theory: a unified treatment. Lawrence Erlbaum Associates, Inc.; Mahwah, NJ: 1999.

McHorney CA. Ten recommendations for advancing patient-centered outcomes measurement for older persons. Annals of Internal Medicine 2003;139(5 Pt 2):403–409. [PubMed: 12965966]

Mislevy RJ. Recent developments in the factor analysis of categorical variables. Journal of Educational Statistics 1986;11(1):3–31.

Murkin JM, Newman SP, Stump DA, Blumenthal JA. Statement of consensus on assessment of neurobehavioral outcomes after cardiac surgery. Annals of Thoracic Surgery 1995;59(5):1289–1295. [PubMed: 7733754]

Muthen B, Kaplan DH,M. On structural equation modeling with data that are not missing completely at random. Psychometrika 1987;52:531–562.

Newman MF, Kirchner JL, Phillips-Bute B, Gaver V, Grocott H, Jones RH, et al. Longitudinal assessment of neurocognitive function after coronary-artery bypass surgery. New England Journal of Medicine 2001;344(6):395–402. [PubMed: 11172175]

Newman S, Stygall J, Hirani S, Shaefi S, Maze M. Postoperative cognitive dysfunction after noncardiac surgery: a systematic review. Anesthesiology 2007;106(3):572–590. [PubMed: 17325517]

Nunnally, JC.; Bernstein, IH. Psychometric Theory. 3rd ed.. McGraw-Hill College Division; New York: 1994.

Ozer DJ, Reise SP. Personality Assessment. Annual Reviews in Psychology 1994;45(1):357–388.

Reise SP, Haviland MG. Item response theory and the measurement of clinical change. Journal of Personality Assessment 2005;84(3):228–238. [PubMed: 15907159]

Rosamond W, Flegal K, Friday G, Furie K, Go A, Greenlund K, et al. Heart disease and stroke statistics--2007 update: a report from the American Heart Association Statistics Committee and Stroke Statistics Subcommittee. Circulation 2007;115(5):e69–171. [PubMed: 17194875]

Rosengart TK, Sweet J, Finnin EB, Wolfe P, Cashy J, Hahn E, et al. Neurocognitive functioning in patients undergoing coronary artery bypass graft surgery or percutaneous coronary intervention: evidence of impairment before intervention compared with normal controls. Annals of Thoracic Surgery 2005;80(4):1327–1334. discussion 1334-1325. [PubMed: 16181864]

Royston P. Multiple imputation of missing values: update. STATA Journal 2005;5(2):118–201.

Rudolph JL, Jones RN, Levkoff SE, Rockett C, Inouye SK, Selke FW, et al. Derivation and validation of a preoperative prediction rule for delirium after cardiac surgery. Circulation 2009;119:229–236. [PubMed: 19118253]

Speer DC. Clinically significant change: Jacobson and Truax (1991) revisited. Journal of Consulting and Clinical Psychology 1992;60(3):402–408. [PubMed: 1619094]

Stevens SS. On the theory of scales of measurement. Science 1946;103(2684):677–680. [PubMed: 17750512]

Stuss DT, Levine B. Adult clinical neuropsychology: lessons from studies of the frontal lobes. Annual Reviews in Psychology 2002;53:401–433.

Trail Making Tests A and B. War Department, Adjutant General's Office; Washington, DC: 1944.

Trenerry, MR.; Crosson, B.; DeBoe, J.; Leber, WR. Visual Search and Attention Test (VSAT). Odessa, FL: 1990.

van Dijk D, Spoor M, Hijman R, Nathoe HM, Borst C, Jansen EW, et al. Cognitive and cardiac outcomes 5 years after off-pump vs on-pump coronary artery bypass graft surgery. Journal of the American Medical Association 2007;297(7):701–708. [PubMed: 17312289]

Wechsler, D. Manual: Wechsler Adult Intelligence Scale - Revised. Psychological Corp; New York: 1981.

Zalsman G, Weizman A, Carel CA, Aizenberg D. Geriatric Depression Scale (GDS-15): a sensitive and convenient instrument for measuring depression in young anorexic patients. Journal of nervous and mental disease 2001;189(5):338–339. [PubMed: 11379982]

Zwick W, Velicer W. Comparison of five rules for determining the number of components to retain. Psychological Bulletin 1986;99(3):432–442.

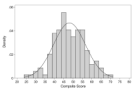

Figure 1. Distribution of Estimated Neuropsychological Composite

The neuropsychological composite has been scaled (T-score) with a mean of 50 and a standard deviation of ±10.

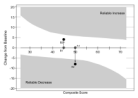

Figure 2. Reliable Change Regions for the Neuropsychological Composite

Shaded areas describe regions of reliable change using the Edwards-Nunnally method (Atkins et al., 2005; Speer, 1992). The asymmetric pattern reflects the accommodation of regression to the mean in the determination of reliable change. The composite is scaled as a T-score, with a mean of 50 and standard deviation of 10 in the sample. Therefore changes of ±6 are reliable changes at a latent trait (θ) level of 50, of six-tenths of a standard deviation. Two hypothetical persons are illustrated: A and B at time 1 and 2. Person A at time 1 (A1) scored 50, and scored 42 (=50-8) at follow-up (A2). This change falls in the “Reliable Decrease” area, meaning a clinically significant decrease in neuropsychological performance has occurred (change in excess of measurement error and regression to the mean). Person B scored 45 at baseline (point B1) and 50 (=45+5) at follow-up (B2). This gain does not fall in the “Reliable Increase” area. Therefore, the gain does not reflect improvement that is in excess of the measurement error of the instrument and taking into account regression to the mean, i.e., the change is not reliable.

Table 1

Participant characteristics (n=210)

Characteristic	Mean (±SD) or n(%)	Range
Age (years)	72 (±7)	48, 87
Men	160 (76%)	
Race		
White	200 (95%)	
All others	10 (5%)	
Education		
<high school	28 (13%)	
high school	68 (33%)	
>high school	114 (54%)	
Mini-Mental State Examination (0-30, 30 best)	27 (±2)	17, 30
Geriatric Depression Scale (0-15, 0 best)	3 (±3)	0, 13

SD, standard deviation;

Table 2

Summary Statistics for Neuropsychological Tests (N=210).

Distributional Features																
Test	N	Mean	SD	Obs Range	Skewness	Kurtosis	P†	Decile Thresholds								
Visual Search & Attention Task	204.0	40.8	11.7	[12, 80]	0.381	3.510	0.032	28	31	35	37	40	43	46	50	55
HVLT Retention Percent	209	68.5	33.8	[0, 250]	0.242	6.379	0.000	14	44	60	67	71	78	86	90	100
Trails B	197	136.4	68.8	[45, 300]	0.951	2.919	0.000	66	78	90	100	113	131	161	199	248
Digit Symbol Substitution	209	31.6	12.8	[0, 75]	0.162	3.219	0.427	16	22	25	28	31	33	37	43	49
Digit Symbol Copy	208	67.5	21.2	[0, 93]	−0.716	3.043	0.001	38	49	59	64	69	75	84	91	92
Digit Span Forwards	210	10.0	2.5	[0, 16]	−0.167	4.142	0.025	7	8	9	–	10	11	–	12	13
Digit Span Backwards	210	5.9	2.1	[0, 13]	0.597	3.953	0.001	4	–	5	–	6	–	7	–	9
Phonemic Fluency	199	31.8	13.1	[0, 98]	0.652	5.700	0.000	16	21	25	29	32	35	39	40	47
Categorical Fluency	196	30.3	7.7	[9, 52]	0.130	2.938	0.741	21	24	26	28	30	32	34	37	41
Boston Naming Test	191	25.6	3.4	[12, 30]	−1.246	4.757	0.000	21	24	25	–	26	27	28	–	29

Correlation Matrix											
	Test	1	2	3	4	5	6	7	8	9	10
1	Visual Search & Attention Task	–	0.17	−0.56	0.56	0.57	0.21	0.16	0.28	0.48	0.22
2	HVLT Retention Percent	0.16	–	−0.33	0.34	0.21	0.02	0.14	0.24	0.34	0.30
3	Trails B	−0.51	−0.23	–	−0.65	−0.54	−0.29	−0.33	−0.34	−0.48	−0.46
4	Digit Symbol Substitution	0.51	0.32	−0.59	–	0.77	0.21	0.23	0.35	0.52	0.47
5	Digit Symbol Copy	0.52	0.21	−0.47	0.71	–	0.22	0.20	0.27	0.46	0.34
6	Digit Span Forwards	0.16	0.00	−0.25	0.18	0.20	–	0.47	0.23	0.25	0.23
7	Digit Span Backwards	0.12	0.17	−0.29	0.21	0.18	0.47	–	0.18	0.07	0.18
8	Phonemic Fluency	0.27	0.22	−0.32	0.36	0.27	0.24	0.19	–	0.49	0.30
9	Categorical Fluency	0.45	0.31	−0.46	0.48	0.43	0.20	0.07	0.48	–	0.45
10	Boston Naming Test	0.20	0.36	−0.41	0.43	0.26	0.21	0.19	0.30	0.43	–

Notes: Obs, observed, SD, standard deviation; P†, D'Agustino et al's test for normality (D'Agostino et al., 1990). P-values less than 0.05 suggest the variable is not normally distributed. Pearson correlation coefficients are presented in the lower diagonal. Polychoric correlations are presented in the upper diagonal.

Table 3

Factor analysis results: Unidimensional and bifactor measurement models

Test	Unidimensional Model	Bifactor Model			
	Common Factor Loading	Common Factor Loading	Specific Factor Loading		
			S1	S2	S3
HVLT Retention Percent	0.66	0.69			
Visual Search & Attention	0.39	0.40			
Task					
Trails B	0.75	0.79			
Digit Symbol Substitution	0.89	0.80	0.41		
Digit Symbol Copy	0.80	0.71	0.49		
Digit Span Forwards	0.37	0.33		0.61	
Digit Span Backwards	0.35	0.30		0.62	
Phonemic Fluency	0.49	0.46			0.47
Categorical Fluency	0.67	0.66			0.39
Boston Naming Test	0.55	0.57			
Model Fit					
CFI	0.887	0.980			
RMSEA	0.142	0.062			
χ^2 (degrees of freedom)	120 (23)	38 (21)			

Notes: HVLT, Hopkins Verbal Learning Task. CFI, Comparative Fit Index. RMSEA, Root mean square error of approximation. An acceptably fitting model is characterized by a CFI of ≥0.95 and an RMSEA <0.10. Bifactor models include a common factor loading in all indicators and one or more specific factors (S1, S2, S3) loading in individual items. Specific factors and the common factors are uncorrelated. Individual items may load on one or two factors in a bifactor model.