Journal of the International Neuropsychological Society (2004), **10**, 578–589.

DOI: 10.1017/S1355617704104049

Increasing the reliability of ipsative interpretations in neuropsychology: A comparison of reliable components analysis and other factor analytic methods

THOMAS W. FRAZIER,[1] ERIC A. YOUNGSTROM,[1] GORDON J. CHELUNE,[2] RICHARD I. NAUGLE,[3] AND TARA T. LINEWEAVER[3]

[1]Department of Psychology, Case Western Reserve University, Cleveland, Ohio
[2]Mellen Center for Multiple Sclerosis Treatment and Research, Cleveland Clinic Foundation, Cleveland, Ohio
[3]Section of Neuropsychology, Cleveland Clinic Foundation, Cleveland, Ohio

(RECEIVED June 13, 2003; REVISED December 30, 2003; ACCEPTED December 30, 2003)

Abstract

Ipsative approaches to neuropsychological assessment typically involve interpreting difference scores between individual test scores. The utility of these methods is limited by the reliability of neuropsychological difference scores and the number of comparisons between scores. The present study evaluated the utility of difference scores using factor analytic methods, including reliable components analysis (RCA), equally weighted composites and individual neuropsychological measures. Data from 1,364 individuals referred for neuropsychological assessment were factor analyzed and the resulting solutions were used to compute composite scores. Reliabilities and confidence intervals were derived for each method. Results indicated that RCA outperformed other factor analytic methods, but produced a slightly different factor structure. Difference scores derived using orthogonal solutions were slightly more reliable than oblique methods, and both were more reliable than those from equally weighted composites and individual measures. Confidence intervals for difference scores were considerably smaller for factor methods relative to those for individual test comparisons, due to the greater reliability of factor based difference scores and the smaller number of comparisons required. These findings suggest that difference scores derived from orthogonal factor solutions, particularly RCA solutions, may improve reliability for clinical assessment purposes. (*JINS*, 2004, *10*, 578–589.)

Keywords: Ipsative, Battery interpretation, Reliable components analysis, Difference scores

Introduction

Norm-referenced, ipsative interpretive approaches are frequently employed in cognitive and neuropsychological assessments (Kaplan et al., 1991; Kaufman, 1994; Matarazzo, 1972; Russell, 2000; Tarter & Edwards, 1986). Ipsative methods usually involve comparisons between test scores within an individual. However, the nature of comparisons and the types of measures compared differ greatly across neuropsychological assessment settings and philosophies. Some clinicians may compare all individual measures in a battery to one another, while others may compare only a few, often theoretically meaningful, groupings of measures to one another. The comparison of "hold" tests to other measures sensitive to neurological impairment is an example of a widespread practice that focuses on theoretically motivated constellations of tests (Johnstone & Wilhelm, 1996; Scott et al., 1997). Similarly, some clinicians may examine only comparisons relevant to the condition of study, such as visual *versus* verbal memory comparisons in medial temporal lobe epilepsy.

Regardless of the specific approach, ipsative methods are based upon the notion that difference scores between measures may provide additional information, not gleaned from the level of performance on individual measures. Supporting this perspective, studies have found neurocognitive difference scores predict effort on testing (Langeluddecke & Lucas, 2003), dementia subtype (Cerhan et al., 2002;

Reprint requests to: Thomas W. Frazier, Department of Psychiatry and Psychology (P57) Cleveland Clinic Foundation, 9500 Euclid Avenue, Cleveland, OH 44195. E-mail: tomfraz@umich.edu, fraziert@ccf.org

Jacobson et al., 2002), and localization of neurological impairment (Keilp et al., 1999; Wilde et al., 2001) at the group level. Unfortunately, there is scant evidence for the utility of ipsative approaches on the individual level (Cerhan et al., 2002; Keilp et al., 1999; Wilde et al., 2001). The few positive findings have been in the area of malingering detection and detection of Alzheimer's disease (for examples, see Jacobson et al., 2002; Millis et al., 1998).

Several competing explanations exist for the lack of evidence supporting the clinical utility of difference scores, including over-reliance on null-hypothesis significance testing at the expense of clinically useful statistics such as sensitivity and specificity (Ivnik et al., 2000)[1], poor predictive validity of difference scores, and/or inadequate reliability of difference scores for making clinical interpretations (Macmann & Barnett, 1997; McDermott et al., 1992; Nunnally & Bernstein, 1994; Streiner & Norman, 1995). The first explanation suggests that difference scores have adequate reliability and validity but that the necessary data have not been reported, whereas the latter two accounts propose that reliability and/or validity may be adequate for group data but not for making decisions about individuals. Inadequate validity may be due to a modest relationship between difference scores and criteria or to insufficient reliability of the difference scores of interest. Thus, establishing sufficient reliability is a precondition to evaluating the clinical utility of neuropsychological comparisons. A .90 level of reliability has been recommended as the minimum level needed to make decisions about individuals (Kelley, 1927; Nunnally & Bernstein, 1994). The present study evaluated whether the level of reliability of difference scores derived from individual cognitive measures met that recommended level and compared these levels of reliability to those derived using factor analytic methods. Given the lack of empirical evidence for the clinical utility of neuropsychological difference scores, it was expected that most difference scores based upon individual tests or subtests would not meet this level of reliability.

Additional methodological and logistical difficulties limit the implementation of ipsative interpretive approaches in neuropsychology. Comparisons of individual subtests in a large test battery results in significant increases in the Type I error rate due to the large number of comparisons. For example, a 20-subtest neuropsychological battery permits as many as 190 comparisons. If all possible comparisons were examined, alpha would need to be reduced to .0003 for each comparison, to maintain the Type I error rate at .05. This correction results in extreme differences being necessary for detecting significant cognitive strengths or weaknesses. Secondly, even in situations where individual measures are highly reliable, difference scores between these measures are often considerably less reliable as a result of moderate to high correlations between parent scores (e.g., WAIS–III VIQ r_{xx} = .97, PIQ r_{xx} = .94, VIQ–PIQ difference score r_{xx} = .82; Wechsler, 1997c). Finally, in many clinical settings, individuals are administered tests that have been normed independently, complicating direct comparisons (Russell, 2000). The creation of large co-normed, neuropsychological test batteries circumvents the latter problem. However, current practice in neuropsychology often involves an individualized, flexible approach to battery administration (Lezak, 1995). The advantages and disadvantages of a flexible approach have been discussed elsewhere (Bauer, 2000). We simply note that a flexible battery approach limits the clinician's ability to perform quantitative neuropsychological comparisons unless the comparisons of interest are derived from tests included in a co-normed battery. Even in cases where neuropsychological test batteries have been concurrently normed, test developers have typically not provided the information necessary to perform psychometrically informed difference score comparisons.

The primary purpose of the present research was to compare several methods for increasing the reliability of neuropsychological comparisons. In particular, this study sought to compare the effectiveness of reliable components analysis (RCA) and other factor analytic methods in producing a small number of reliable neuropsychological variables for computing neuropsychological comparisons. Based upon the aforementioned difficulties with implementing an ipsative approach to neuropsychological data, the application of data reduction techniques was expected to address both the problem of reliability of neuropsychological difference scores and the problem of multiple comparisons by creating a smaller set of highly reliable composite variables.

Reliable Components Analysis

Reliable components analysis (RCA) is an exploratory data reduction technique aimed at forming components that have maximum reliability (Caruso, 2001b; Cliff & Caruso, 1998). It is similar to other component and factor analytic techniques in that one or more uncorrelated composites are formed from the original variables. These composites may then be rotated to maximize interpretive value using orthogonal or oblique rotations. RCA differs from other techniques in that the weights derived for each component maximize the amount of reliable variance in the composite uncorrelated with subsequent composites. For this reason, RCA is an especially attractive technique for maximizing the reliability of neuropsychological difference scores because the reliability of these comparisons is largely dependent upon the reliability of the measures (for a more complete description, see Cliff & Caruso, 1998). RCA differs primarily from other factor methods in that RCA uses reliability coefficients in the diagonal of the correlation matrix whereas PCA uses 1.0s and PAF uses squared multiple correlations in the diagonal of the analyzed matrix. Thus, RCA is a

[1]Clearly hypothesis testing is essential for determining whether obtained results are statistically reliable or likely to have occurred by chance. The point to be made here is simply that clinically useful statistics such as sensitivity and specificity, and positive and negative predictive power, can be reported along with null hypothesis significance tests. A good reference for the disrespect of null-hypothesis significance testing is Soper et al. (1988).

https://doi.org/10.1017/S1355617704104049 Published online by Cambridge University Press

compromise between PCA and PAF in that RCA examines only reliable variance whereas PCA examines all of the variance in each variable and PAF analyzes only common variance.

Several studies have examined the ability of RCA to improve the reliability of cognitive test score comparisons (Caruso, 2001a; Caruso & Cliff, 1998, 1999, 2000; Caruso & Witkiewitz, 2001). These studies have found that RCA produces a similar factor[2] structure to other exploratory factor analytic techniques, but that RCA based composites yield more reliable cognitive difference scores than other techniques. Recent findings from a study by Caruso and Witkiewitz (2002) found that the principle method by which RCA generated reliable difference scores was through the creation of orthogonal composites using varimax rotation of factor weights. This suggests that the uncorrelated nature of factors derived using varimax rotation, regardless of the specific extraction, may be the primary driving force in the higher reliability of difference scores derived from factor analytic methods. The present research sought to examine this possibility by including both orthogonal and oblique rotations from several factor analytic methods.

Based upon the above findings, it was predicted that neuropsychological difference scores derived from RCA, principal components analysis (PCA), and principal axis factoring (PAF) solutions with orthogonal (varimax) rotations would have similar levels of reliability. RCA with varimax rotation was expected to outperform PCA and PAF solutions since this solution was expected to produce more reliable composites. In contrast, RCA, PCA, and PAF solutions with oblique (oblimin) rotations were expected to produce lower levels of reliability of cognitive difference scores due to the moderate to high correlations expected between composites. This expectation is based upon the notion of a positive manifold of correlations between cognitive measures (Carroll, 1993). Factor composition was expected to be similar across factor analytic methods, an important condition for making meaningful comparisons of the reliabilities of cognitive difference scores.

METHODS

Research Participants

Data for the present study were obtained from a de-identified patient registry that has been reviewed and approved by the Institutional Review Board at the Cleveland Clinic Foundation. The database consisted of neuropsychological test data from adolescents and adults referred for neuropsychological assessment at the Cleveland Clinic Foundation (CCF)-Section of Neuropsychology. Approximately 50% of the patients referred for examination come from the Department of Neurology, approximately 25% come from the Department of Psychiatry and Psychology, and approximately 25% come from other sources such as the Departments of Internal Medicine, Neurosurgery, Orthopedic Surgery, Rheumatology, Hematology/Oncology, Cardiology, and other departments or community sources. The most frequently occurring referral questions included assessment for possible dementia, attention-deficit/hyperactivity disorder, or learning difficulties; neurocognitive consequences of stroke, traumatic brain injury, or tumor; and pre-operative epilepsy, deep brain stimulation, or hydrocephalus evaluation. Only data from an individual's initial evaluation were included in the present analyses. In general, the sample appeared representative of the population of individuals seen at this outpatient clinic, with the exception of the exclusion of individuals with moderate to severe dementia that were not able to complete any of the measures of interest. The final sample consisted of 1,364 people (47% female; M_{age} = 51.7, SD = 18.1, range = 16–93). The racial distribution was consistent with that of patients seen at CCF and was similar to the racial distribution of the greater Cleveland area (White 88.9%, African American 8.3%, Hispanic .8%, Asian .5%, Other 1.5%). On average, individuals had 13.7 years of education (SD = 2.9, range = 3–22), and 89.8% were right-handed.

Measures

Table 1 presents descriptive statistics for all neuropsychological measures included in the present study. Measures included Trails A and B (Reitan, 1958) time to completion in seconds; standard scores from the Wisconsin Card Sorting Test (Heaton et al., 1993) for the number of perseverative errors (WCST–perseverative), categories (WCST–categories) and set failures (WCST–set failures); verbal comprehension (WAIS–III VCI), perceptual organization (WAIS–III POI), and processing speed (WAIS–III PSI) standard scores from the Wechsler Adult Intelligence Scale–Third Edition (Wechsler, 1997a); working memory (WMS–III WMI), auditory immediate memory (WMS–III auditory immediate), visual immediate memory (WMS–III visual immediate), auditory delayed memory (WMS–III auditory delayed), visual delayed memory (WMS–III visual delayed), and auditory recognition delayed memory (WMS–III auditory recognition) standard scores from the Wechsler Memory Scale–Third Edition (Wechsler, 1997b); Wide Range Achievement Test–3 reading subtest (WRAT–3 Reading) standard scores (Wilkinson, 1993); average number of taps with the dominant (FT–dominant hand) and non-dominant hand (FT–non-dominant hand) from the finger tapping test (Halstead, 1947; Reitan, 1955; Spreen & Strauss, 1998); total number of seconds to complete the Grooved Pegboard (see Lezak, 1995; Mitrushina et al., 1999) with the dominant (GP–dominant hand) and non-dominant hand (GP–non-dominant hand); total raw score, including spontaneously correct responses and correct responses after

[2]The term factor is used to denote both component methods and common factor methods and refers to any procedure used to reduce the number of variables in a data set to a smaller set of variables that typically explain a large proportion of the variance of the original variables.

https://doi.org/10.1017/S1355617704104049 Published online by Cambridge University Press

Table 1. Descriptive statistics for all neuropsychological measures

	r_{xx}	*M*	*SD*	*SD of M*	*SE*	Median
Trails A	.70	50.51	40.19	0.48	1.09	37.00
Trails B	.78	138.06	90.94	0.93	2.46	104.67
WCST–perseverative errors	.64	89.08	19.26	0.43	0.52	90.67
WCST–categories	.70	3.75	2.22	0.01	0.06	4.00
WCST–set failures	.50	1.12	1.33	0.02	0.04	1.00
WAIS–III VCI	.96	97.18	17.05	0.25	0.46	98.00
WAIS–III POI	.93	93.94	17.79	0.40	0.48	93.67
WAIS–III PSI	.88	88.35	16.40	0.11	0.44	88.00
WMS–III WMI	.86	91.83	17.65	0.38	0.48	99.00
WRAT–3 Reading	.92	97.49	13.69	0.03	0.37	90.67
WMS–III Auditory Immediate	.93	91.32	18.90	0.47	0.51	90.00
WMS–III Visual Immediate	.82	87.86	18.05	0.18	0.49	92.33
WMS–III Auditory Delayed	.87	92.60	18.84	0.42	0.51	88.33
WMS–III Visual Delayed	.83	88.69	18.43	0.26	0.50	95.00
WMS–III Auditory Recognition	.74	94.13	18.56	0.19	0.50	93.00
Finger Tapping–Dominant	.77	41.08	10.56	0.06	0.29	42.37
Finger Tapping–Non-Dominant	.78	38.08	11.98	0.55	0.32	39.07
Grooved Pegs–Dominant	.86	109.22	60.39	0.97	1.64	90.33
Grooved Pegs–Non-Dominant	.86	118.33	64.01	0.41	1.73	99.33
Boston Naming Test	.93	48.28	10.76	0.11	0.29	29.00
Verbal Fluency	.83	29.29	13.09	0.06	0.35	51.33

**Note*. Means, *SD*s, *SD*s of *M*, SEs, and Medians were calculated using all imputed data using the procedures described by Graham and Schafer (1999), therefore median values are actually the average median values from three imputed data sets. *SD* of *M* is computed as the standard deviation of the three means obtained from imputed data sets, whereas *SD* is the standard deviation within each data set averaged across the three data sets.

semantic cue, from the Boston Naming Test (Kaplan et al., 1983); and total number of words produced during three 60-s trials each using either phonemic fluency FAS (Spreen & Benton, 1969) or CFL from the Controlled Oral Word Association Test (Verbal Fluency; Benton et al., 1994).

These 21 measures were selected for analysis based upon several considerations. First, this set represents a core battery of measures given to the majority of outpatients referred for testing. As such, this set provided data on a large number of subjects, with a moderate rate of missing data (27% missing). Second, these measures are derived from tests commonly given in neuropsychological assessment (Lees-Haley et al., 1996). Using data from the survey by Lees-Haley et al., it was estimated that the percentage of neuropsychologists using tests included in the present study ranged from approximately 75% for the WAIS–III (WAIS–R in that study) to approximately 7–10% for the verbal fluency task. Thus, the present results may inform other neuropsychological assessment settings where similar measures are administered. Finally, the measures included in the present study were thought to sample several major domains of cognitive functioning, including language/verbal reasoning abilities, visuoperceptual/constructional skills, attention, executive functions, and memory.

Procedure

Data imputation

In order to maximize sample size and avoid possible introduction of bias due to missing data, we employed the multiple data imputation procedure developed by Graham and Schafer (Graham & Schafer, 1999; Schafer, 2002). The procedure was used to impute three sets of values for each individual. In each data set the missing values are estimated, with slightly different estimations for each missing value in each data set, and complete data remain consistent across data sets. Missing values are estimated using an iterative procedure based upon parameter estimates derived from the EM algorithm. For three imputed data sets, the standard error of imputations will tend to be only 1.04 times as wide as the standard error of an infinite number of data sets (Graham & Schafer, 1999). This indicates that the three data sets in the present study were likely to include highly similar imputed values for all of the missing data from the original data matrix.

Data analyses

Before examining the increase in reliability of neuropsychological comparisons using factor analytic methods, we

https://doi.org/10.1017/S1355617704104049 Published online by Cambridge University Press

first determined the number of factors present in the data set. To make this decision, separate PCA's were computed for each imputed data set, and the resulting eigenvalues were compared to the random values derived by Horn's parallel analysis (HPA; O'Connor, 2000). These analyses were performed separately for each sub-sample within each imputed data set. HPA, used in conjunction with PCA, has been shown through Monte Carlo simulations to be more accurate than conventional methods at judging the number of factors in a data set (Buja & Eyuboglu, 1992; Widaman, 1993; Zwick & Velicer, 1986). Next, the replicability of factor structures was determined for each of the six factor analytic methods examined (PCA–varimax, PCA–oblimin, PAF–varimax, PAF–oblimin, RCA–varimax, and RCA–oblimin) using SPSS (2002) and RCA syntax for SPSS detailed in Caruso (Caruso & Cliff, 2002). Data were randomly divided into two equal sub-samples ($n = 682$) for each imputed data set and all six analytic procedures were performed for each sub-sample. Results for sub-samples were compared using congruence coefficients (Tucker, 1951). In the present study, congruence coefficients were computed in two ways. The first method computed congruency as the correlation between the two independent sets of component loadings. Loadings were obtained from the rotated component matrix for varimax solutions and from the structure matrix for oblique solutions. This method is comparable to the factor congruence coefficients reported by McCrae et al. (1996) . The second method computed congruency as the correlation between the two independent sets of component weights from the rotated weight matrix. The latter method estimates the comparability of factor scores from each sub-sample. After determining the replicability of each solution, sub-samples were recombined and all factor analytic procedures were recomputed and scores for each factor derived from each factor method were retained. The rotated component or structure matrices and weight matrices were examined for each factor method to specify the nature of resulting factor scores. Convergent and discriminant validity of each factor analytic method was examined by computing the correlation between the resulting factor scores.

The reliability of neuropsychological difference scores was determined by first computing the reliability of each composite for each factor solution. To accomplish this, reliabilities for each neuropsychological variable were obtained from published reports (Bowden et al., 1998; Dikmen et al., 1999; Fastenau et al., 1998; Franzen et al., 1995, 1996; Ingram et al., 1999; Tate et al., 1998; Wechsler, 1997c; Wilkinson, 1993). In cases where internal consistency or alternate forms reliability estimates could not be obtained, test–retest reliability was substituted. When multiple reliability estimates were found in the literature, estimates were averaged using Fisher's *r*-to-*z* transformation (Corey et al., 1998). For RCA, the reliability of each composite retained was provided in the output. For the other methods, these values were computed using the formulas provided in Nunnally and Bernstein (1994). The reliabilities of difference scores obtained for each factor solution were computed via the equation provided in Streiner and Norman (1995; p. 168). To demonstrate the increase in reliability using differentially weighted composites, equally-weighted composite scores were also computed using variables loading at or above .55, on average, in the rotated component/structure matrices of all methods.[3] These scores were computed by weighting each variable with a significant loading 1.0 and averaging scores across variables. The resulting difference scores comparing equally weighted composites approximate the clinical interpretive approach of grouping tests into particular cognitive domains and then qualitatively comparing these groupings to determine cognitive strengths and weaknesses. Standard errors of estimation derived from standardized difference scores were computed for each factor analytic solution and for the equally weighted composites (Lord & Novick, 1968). Reliabilities of individual test difference scores were also computed between WAIS–VCI and all other variables and WRAT–Reading and all other variables. These comparisons were chosen since they are commonly employed in clinical practice to compare measures thought to be relatively resistant to brain damage and measures sensitive to neurological impairment (Johnstone & Wilhelm, 1996; Scott et al., 1997). Standard errors of estimation were used to compute 95% confidence intervals for all difference scores. Confidence intervals were computed in standard score units ($M = 100$, $SD = 15$) and were adjusted for the number of comparisons (six comparisons for factor methods and equally weighted composites and 19 comparisons for individual measures). Adjustments were used to balance the increase in Type I error with multiple comparisons. Thus, the resulting 95% confidence intervals indicate that 95% of the time differences between all non-significant score comparisons will be detected as such.

RESULTS

Missing Data Analyses

Approximately 27% of all neuropsychological test data was missing, with the majority of missing data attributed to WCST variables (smallest $n = 881$), Finger Tapping Test variables (smallest $n = 768$), and Grooved Pegboard variables (smallest $n = 889$). Over 65% of the sample had data from at least 16 of the 21 measures and 35% of the sample completed all measures. To determine the influence of missing data on observed and imputed values, correlations between a dichotomous measure specifying whether data was

[3]RCA–oblimin solutions were excluded from the average due to the poor replicability of this method. The .55 value of matrix coefficients was chosen to eliminate the influence of measures with moderate loadings on the reliability of equally weighted composites. WAIS–III POI was not included in the derivation of any of the equally weighted composites because it cross-loaded on two factors. Inclusion would have further diminished the reliability of equally weighted difference scores by increasing the correlation between composites. WMS–WMI and WCST–set failure did not have significant loadings on any factor and therefore were not included in any composite.

https://doi.org/10.1017/S1355617704104049 Published online by Cambridge University Press

complete or imputed was correlated with each variable, separately in each data set. Correlations were then averaged across variables and across imputed data sets. On average, the potential influence of missing data on observed scores was significant, but modest in size (average $r = -.14$, $p < .001$). The data were well suited for multiple imputation procedures, both in terms of completeness and the small relationship between missing data and observed/imputed values.

Table 1 presents descriptive statistics for all neuropsychological variables. Calculations of descriptive statistics and weight matrices derived by factor analytic procedures were performed separately for each imputed data file and then statistics were averaged following the procedures outlined by Graham and Schafer (1999). Very little variation was observed for mean values across imputed data files as evidenced by the small *SD of M* relative to average sample standard deviations of scores (see Table 1 Columns 4 and 5). Similarly, weights derived from PCA and PAF factor solutions tended to be extremely consistent across imputed data sets (PCA–varimax $r = .99$, PCA–oblimin $r = .99$, PAF–varimax $r = .99$, PAF–oblimin $r = .99$). RCA–varimax was slightly less consistent ($r = .93$) and RCA–oblimin displayed poor consistency across data sets ($r = .52$). Given the low level of between data set consistency for RCA–oblimin, no further analyses were performed for this method.

Number of Factors and Replicability

HPA indicated a three-factor solution for each sub-sample in each imputed dataset. In most of these analyses the fourth component was only slightly smaller in magnitude than the random values generated using HPA. Based upon this consideration, interpretability, and previous work suggesting that underfactoring is a more serious error than overfactoring (Fabrigar et al., 1999), four factors were retained in subsequent analyses.

Congruence coefficients from structure or weight matrices were computed within each imputed data file and then averaged across files. Coefficients derived from structure or rotated component matrices were .98 or better for the first two components of each factor analytic method. For the third factor, coefficients ranged from $r = .97$–$.99$, with the exception of RCA–varimax ($r = .94$). Coefficients for the fourth factor were high for PAF methods (PAF–varimax $r = .98$, PAF–oblimin $r = .99$), somewhat lower for PCA solutions (PCA–varimax $r = .88$, PCA–oblimin $r = .93$), and very low for RCA–varimax (.51). For coefficients comparing factor weight matrices, the first three factors were highly replicable for PCA and PAF methods ($r = .96$–$.99$). Coefficients for the fourth factor were high for PAF solutions (PAF–varimax $r = .97$, PAF–oblimin $r = .97$) and somewhat lower for PCA solutions (PCA–varimax $r = .89$, PCA–oblimin $r = .89$). RCA displayed high coefficients for the first three factors ($r = .95$–$.97$), however the coefficient for the fourth factor was considerably lower ($r = .80$). Analyses involving the fourth RCA factor should be viewed cautiously given the variable replicability of this factor.

Factor Composition and Convergent/Discriminant Validity

To examine the nature of factor solutions, coefficients from the structure or rotated component matrices were derived for each factor method using the entire sample, separately in each imputed data set, and then averaged across data sets. Table 2 presents the mean structure/rotated component matrix averaged across methods and imputed data sets. Weight matrices were also examined to aid in interpretation. The first three factors were easily interpreted and were highly consistent across factor methods.[4] The first factor was labeled *Memory* with high loadings from all of the WMS–III indices. The second factor was labeled *Visual Motor* and included high loadings from Trails A and B, FTT–dominant and non-dominant, GP–dominant and non-dominant, and moderate loadings from WAIS–III PSI and WAIS–III POI. The third factor was labeled *Language* and included high loadings from WAIS–III VCI, Boston Naming, and WRAT–3 Reading and moderate loadings from Verbal Fluency, WAIS–III POI, and WMS–III WMI. For the fourth factor, labeled *Executive Functioning*, PCA and PAF had consistently high loadings for WCST–perseverative errors and WCST–categories. However, the fourth factor for RCA–varimax was not consistent with other methods, possibly due to the poor consistency between imputed data sets and inadequate within data set replicability of this RCA factor. The label executive functioning did not apply for RCA–varimax since the largest weights for this factor were a positive weight for WAIS–III POI and a negative weight for Boston Naming Test. For simplicity, this label was retained for all methods. However, comparisons between RCA and other methods for this factor likely involved different constructs.

Table 3 presents convergent and discriminant validity for each method and factor. Correlations between factor scores derived for each factor method and the equally weighted composites were computed to examine the convergent and discriminant validity of these factors. Convergent validity was computed by averaging correlations between factors with similar composition across methods. For example, PCA–varimax Memory factor scores were correlated with Memory factor scores from all other methods. Discriminant validity was determined by averaging correlations between different factors from different methods. For example, PCA–varimax Memory was correlated with all other factors for all other factor methods.

PCA, PAF, and the equally weighted composites displayed good convergent validity for all factors (Memory $r = .94$–$.97$, Visual Motor $r = .83$–93, Language $r = .90$–$.95$,

[4]Factors did not consistently appear as the first, second, third, and fourth factors in each solution. For simplicity, the factor occurring most frequently in these positions is labeled as such.

https://doi.org/10.1017/S1355617704104049 Published online by Cambridge University Press

Table 2. Structure/rotated component matrix averaged across methods and imputed data sets

	Memory	Visual Motor	Language	Executive
Trails A	.36	*.77*	.20	.24
Trails B	.47	*.68*	.30	.41
WCST–perseverative errors	.31	.32	.19	*.63*
WCST–categories	.39	.43	.17	*.66*
WCST–set failures	.05	.04	.05	.37
WAIS–III VCI	.46	.19	*.87*	.26
WAIS–III POI	.44	*.55*	*.55*	.44
WAIS–III PSI	.46	*.62*	.45	.34
WMS–III WMI	.48	.49	.52	.38
WRAT–3 Reading	.27	.15	*.82*	.17
WMS–III Auditory Immediate	*.84*	.26	.46	.28
WMS–III Visual Immediate	*.80*	.32	.26	.17
WMS–III Auditory Delayed	*.86*	.23	.41	.25
WMS–III Visual Delayed	*.81*	.33	.28	.19
WMS–III Auditory Recognition	*.76*	.19	.39	.25
Finger Tapping–Dominant	.15	*.74*	.24	.17
Finger Tapping–Non-Dominant	.09	*.60*	.18	.11
Grooved Pegs–Dominant	.28	*.88*	.13	.16
Grooved Pegs–Non-Dominant	.30	*.85*	.14	.19
Boston Naming Test	.47	.33	*.64*	.14
Verbal Fluency	.39	.42	*.56*	.20

Note. Values greater than .55 are italicized.

and Executive Functioning $r = .79–.83$). RCA showed good convergent validity for the first three factors (Memory $r = .94$, Visual Motor $r = .86$, Language $r = .92$) but poor convergent validity for the fourth factor (Executive Functioning $r = .33$). This is consistent with the previously discussed differences between the nature of the fourth factor for RCA–varimax and the fourth factor for other methods. As expected, based upon the uncorrelated nature of the scores, varimax solutions showed superior discriminant validity to oblique solutions and equally weighted scores.

Reliability of Difference Scores

Prior to examining difference score reliability, composite reliability was computed for each method. All factor solutions produced highly reliable composite scores for the first three factors (Memory $r = .94–.97$, Visual Motor $r = .92–.97$, and Language $r = .94–.97$), with only a slight advantage for RCA–varimax scores. The equally weighted composites also produced highly reliable scores for these factors ($r = .96–.97$). For executive functioning, RCA–varimax produced the most reliable composite ($r = .88$), with PAF–oblimin being slightly less reliable ($r = .84$), and other methods producing the least reliable composite scores (PCA–varimax = .68, PCA–oblimin = .73, PAF–varimax = .71, equal weighting = .80). However, as noted previously, this RCA factor displayed poor convergent validity with other factor solutions.

Table 4 presents reliabilities, standard errors of measurement, and 95% confidence intervals for difference scores

Table 3. Convergent and discriminant validity of factor scores for each method

	PCA–varimax	PCA–oblimin	PAF–varimax	PAF–oblimin	RCA–varimax	Equal weighting
Convergent validity						
Memory	.94	.97	.95	.97	.94	.97
Visual Motor	.87	.84	.87	.83	.86	.93
Language	.94	.95	.94	.94	.92	.90
Executive	.79	.83	.82	.82	.33	.81
Discriminant validity						
Memory	.21	.37	.20	.39	.24	.37
Visual Motor	.20	.32	.18	.31	.25	.32
Language	.17	.32	.17	.34	.13	.44
Executive	.10	.24	.23	.46	.08	.39

https://doi.org/10.1017/S1355617704104049 Published online by Cambridge University Press

Table 4. Reliabilities, standard errors of estimation, and 95% confidence intervals for each difference score by method

	PCA–varimax	PCA–oblimin	PAF–varimax	PAF–oblimin	RCA-varimax	Equal weights
Reliabilities						
Memory–Visual Motor	.93	.92	.93	.92	.95	.91
Memory–Language	.93	.92	.93	.91	.95	.88
Memory–Executive	.80	.77	.82	.76	.91	.76
Visual Motor–Language	.94	.93	.94	.94	.95	.91
Visual Motor–Executive	.81	.78	.82	.76	.92	.69
Language–Executive	.81	.80	.83	.82	.92	.77
Standard errors of estimation						
Memory–Visual Motor	.26	.27	.26	.27	.22	.29
Memory–Language	.26	.27	.26	.29	.22	.32
Memory–Executive	.40	.42	.38	.43	.29	.43
Visual Motor–Language	.24	.26	.24	.24	.22	.29
Visual Motor–Executive	.39	.41	.38	.43	.27	.46
Language–Executive	.39	.40	.38	.38	.27	.42
95% CI						
Memory–Visual Motor	9.36	9.89	9.06	9.65	8.17	10.02
Memory–Language	9.21	9.98	8.90	10.40	8.17	11.68
Memory–Executive	14.38	15.06	13.79	15.39	10.26	15.33
Visual Motor–Language	8.61	8.84	8.61	8.58	7.81	10.31
Visual Motor–Executive	14.12	14.82	13.65	15.32	10.00	16.51
Language–Executive	14.05	14.44	13.58	13.85	10.00	14.99

by factor method. Inspection of this table reveals that the most reliable difference scores were obtained by RCA–varimax. This was especially true for comparisons involving the executive functioning composite, highlighting the increased reliability of the fourth RCA–varimax composite relative to other methods. Other varimax rotated solutions produced the next highest reliability estimates for factor difference scores, and equally weighted difference scores yielded the lowest reliability estimates. In terms of the magnitudes of reliability coefficients, several met the recommended (.90) level of reliability for clinical interpretation. This was especially true for RCA–varimax, where all six estimates exceeded the recommended level. For other methods, comparisons not involving executive functioning all exceeded the recommended level, with the exception of equal weighting, where only two of the three comparisons exceeded .90.

In general, varimax rotated methods produced slightly higher reliabilities than oblimin rotated solutions, with all varimax solutions exceeding .80 and only three of six comparisons for each oblimin solution exceeding .80. Inspection of the standard errors of estimation further supports the advantage of RCA-varimax over other methods and the superiority of varimax solutions over oblimin solutions. All of the standard errors for RCA comparisons were less than 1/3 the standard deviation of scores, while only half of the comparisons for other methods were less than 1/3 the standard deviation of scores. Varimax solutions consistently produced smaller standard errors of estimation than oblique solutions as a result of greater reliability. It should be noted that PCA and PAF analyses using promax rotation were also performed. These analyses were examined since the promax rotation begins with a varimax rotation and then allows factors to correlate, producing a more realistic factor structure for cognitive measures. However, these analyses were not reported because the promax rotation yielded even higher correlations between factors than both PCA and PAF oblimin solutions, and therefore produced less reliable difference scores and were largely redundant with findings for oblimin solutions.

In terms of individual measure difference scores, WAIS–III VCI and WRAT–3 Reading comparisons were generally less reliable than those derived from factor methods and equal weighting (VCI difference score reliabilities $r = .66$–$.88$, average $r = .79$; WRAT–Reading difference score reliabilities $r = .69$–$.86$, average $r = .79$). None of the comparisons reached the recommended .90 level of reliability and only 8 of 19 for WAIS–VCI and 11 of 19 for WRAT–Reading met the .80 level of reliability. Selected difference scores from other comparisons indicated even poorer levels of reliability, with a few comparisons approaching zero reliability.

Table 4 also presents 95% confidence intervals for difference scores derived from each factor method as well as equally weighted difference scores. Individual test comparisons are not included in this table due to the large number of comparisons. However, confidence intervals for these comparisons were also computed. Confidence intervals for each method were adjusted for the number of comparisons required in order to maintain Type I error at .05. In essence,

https://doi.org/10.1017/S1355617704104049 Published online by Cambridge University Press

this creates larger confidence intervals for each comparison but the convention of 95% confidence interval is used to indicate that a Type I error should occur only 5% of the time for all comparisons (for a review and comparison of methods for controlling error, see Williams et al., 1999). As expected, factor methods produced considerably smaller confidence intervals than equally weighted difference scores and individual test comparisons (PCA–varimax ± 8.6 to ±14.4, PCA–oblimin ± 8.8 to ±15.1, PAF–varimax ±8.6 to ±13.8, PAF–oblimin ±8.6 to ±15.4, RCA–varimax ±7.8 to ±10.3, equal weights ±10.0 to ±16.5, VCI individual test comparisons ± 13.8 to ±19.8, WRAT–Reading individual test comparisons ± 14.4 to ±19.4). This was especially evident for comparisons involving only the memory, visual motor, and language factors. RCA consistently produced the smallest confidence intervals.

DISCUSSION

The present study demonstrated the potential clinical utility of orthogonal, factor-analytically derived neuropsychological difference scores. Relative to individual test comparisons, factor methods produced a smaller number of highly reliable comparisons, decreasing the size of the confidence interval necessary for detecting cognitive discrepancies. As expected, RCA–varimax difference scores were slightly more reliable than scores from other varimax solutions, and the latter scores were more reliable than those derived from oblique solutions. Equally weighted composites produced more reliable difference scores than those from individual measures, but were less reliable than scores derived from factor methods. The superiority of factor based difference scores was evident in terms of reliability and, more prominently, in the reduced size of confidence intervals for interpreting cognitive discrepancies. The two major reasons for the latter finding were greater reliability of factor composites and a smaller number of comparisons generated using these methods.

The increased reliability and decreased confidence intervals for neuropsychological difference scores were most prominent for orthogonal factor solutions, supporting the notion that the lack of correlation between factors was the major reason for highly reliable difference scores. This conclusion is also bolstered by the fact that the promax rotation, which begins as a varimax rotation and then allows factors to correlate, yielded very poor reliability for the resulting difference scores as a result of even higher correlations between factors than those obtained for oblimin rotations.

RCA–varimax scores tended to outperform scores from other factor methods, although there were several important caveats to this finding. The replicability of RCA–varimax solutions was lower for the fourth factor than for other methods. This result may be specific to the present data or may indicate that RCA solutions are more sensitive than other methods to small perturbations in observed correlations. Findings also contradicted previous studies suggesting that RCA produces similar factors to other methods (Caruso & Cliff, 2000; Caruso & Witkiewitz, 2002); however this was only the case for the last factor extracted. For PCA and PAF solutions, the last factor was primarily made up of variables with low reliability (WCST–perseverative errors r_{xx} = .64; WCST–categories r_{xx} = .70), but for the RCA–varimax solution the composition of this factor emphasized more reliable variables (WAIS–III POI r_{xx} = .93; Boston Naming Test r_{xx} = .93). This may imply that RCA produces similar initial factors to other techniques, but that the emphasis on reliability in RCA causes subsequent factors to differ considerably from other factor methods.

Alternatively, the present findings may simply have resulted from overfactoring, since more factors were retained than indicated by Horn's parallel analysis (HPA). Monte Carlo studies are needed to determine whether consistent differences are observed between RCA and other factor analytic techniques. These studies should vary the reliability of variables and the number of factors extracted to give a clearer picture of the convergent validity of RCA with other methods. Our findings suggest that use of accurate criteria for determining the number of factors to extract, such as HPA, may limit differences between RCA and other factor solutions. Simulation work should also examine RCA–oblimin solutions to determine whether this technique produces unstable solutions, as it did in the present study. In the event that RCA solutions are found to be less replicable or more sample dependent than other factor methods, the present findings indicate that use of HPA in conjunction with PCA or PAF varimax solutions is likely to yield factor based neuropsychological difference scores that are highly reliable. All difference scores obtained by comparing the first three factors from both PCA and PAF varimax solutions displayed reliabilities at or above .90.

Factor analytic procedures for deriving reliable cognitive discrepancy scores are not a panacea for poorly constructed tests, nor are they the only methods for conceptualizing clinical interpretation of neuropsychological test batteries. Methods such as profile analysis may prove to be even more useful than the Fisherian techniques discussed in the present paper. It should also be noted that, ultimately, the effectiveness of any method in producing reliable comparisons is directly dependent upon the reliability of the individual measures. We were struck by the lack of published reports concerning the reliability of neuropsychological measures employed in the present study. For some tasks only one report concerning either internal consistency or test–retest reliability could be obtained. The lack of reliability data suggests limited attention to the psychometric characteristics of many frequently used neuropsychological measures, an undesirable state of affairs given the ever expanding role of neuropsychological assessment (Fennell, 1995; Ivnik et al., 2000). Examination of the existing literature also indicated that several of the neuropsychological measures included in this study did not meet the recommended level of reliability for clinical use (.90). In particular, the tasks with the poorest reliabilities tended to have small numbers

https://doi.org/10.1017/S1355617704104049 Published online by Cambridge University Press

of items or only a single series of trials. For example, Trails A and B and Grooved Pegboard are one-item tests and Verbal Fluency is a three-item test. WCST consists of a series of responses but each response is dependent upon previous responses and therefore cannot be viewed as a multi-item test.

Additional studies are needed to examine and improve the psychometric characteristics of commonly used neuropsychological measures. Enhancing the reliability of future neuropsychological measures will be especially important for establishing the clinical utility of ipsative interpretive methods in neuropsychological assessment. To accomplish this purpose, test developers should not only focus on enhancing the reliability of individual measures, but should also attempt to measure more specific cognitive processes. Measuring more specific processes would enhance the reliability of factor-analytically derived difference scores by increasing the saturation of factors and decreasing the correlation between factors measuring different cognitive processes. Decreasing the correlation between distinct measures would facilitate the utility of difference scores based upon oblique solutions. Oblique solutions are likely to better represent the true structure of a neuropsychological data set, and having smaller correlations between factors would minimize the undesirable psychometric properties of the resulting difference scores.

The major limitation of the present study was the moderate amount of missing data. We addressed this concern using the best methodological approach available at present, creating multiple imputations using the EM algorithm (Graham & Schafer, 1999). Analyses examining the missing-at-random assumption indicated a small, but significant, negative relationship between missingness and most measures. This suggests that individuals with missing data were generally more impaired than individuals with complete data and presents a problem for inferential analyses. However, the present study investigated the relationships between variables and not inferences about means. Therefore, violation of the missing-at-random assumption may not have been as problematic for the present study. List-wise analyses performed for 478 people with complete data were highly consistent with imputed data analyses, further supporting the notion that missing data had little effect upon the structure of relationships between variables in the present analyses.

Clinical Implications

Findings demonstrated that many commonly employed individual test difference scores have less than desirable levels of reliability. This suggests that these differences should not be the primary or exclusive set of comparisons examined in routine clinical use. In fact, in the present study, individual test difference scores often did not meet the level of reliability recommended for research (.70) and never met the levels recommended for clinical use (.90; Nunnally & Bernstein, 1994). Ipsative approaches relying on these scores are likely to provide a less sensitive evaluation of individuals' cognitive strengths and weaknesses due to the large differences required between individual measures to achieve statistical significance. In contrast to single measure difference scores, orthogonal, differentially weighted factor scores, particularly those derived from RCA, are likely to have sufficient reliability to justify clinical use. Based upon this conclusion, it is recommended that the procedures used in the present study are applied in other neuropsychological assessment settings to increase the utility of ipsative interpretive approaches. Factor-based difference scores may also be useful for increasing the interpretive yield of existing co-normed, neuropsychological batteries. Application of these methods may increase the benefits of fixed *versus* flexible battery approaches to assessment by providing a more sensitive evaluation of cognitive strengths and weaknesses.

Increasingly neuropsychologists are moving away from traditional neuropsychological batteries, such as the Halstead-Reitan, and are moving toward batteries combining one or more cognitive sub-batteries along with traditional neuropsychological tests (e.g., using one or more of the following WAIS–III, WMS–III, Kaufmann Brief Intelligence Test, Woodcock Johnson–III ability and achievement batteries). The sub-batteries included in these assessments typically provide procedures for examining cognitive discrepancies. However, the difference scores examined are limited to the constructs measured by the instrument. Some authors have suggested using multiple sub-batteries to provide more comprehensive coverage of cognitive functioning (Bauer, 2000). While this recommendation accomplishes the intended goal of providing more comprehensive coverage of cognitive constructs, interpretation of these assessments is limited in that difference scores can only be computed within batteries and there are no formal procedures to account for construct overlap.

Application of factor analytic methods, such as RCA, to the entire neuropsychological battery will provide a more sensitive evaluation of the broad cognitive domains assessed by the entire battery than difference scores within sub-batteries. Specifically, factor methods can be useful for specifying the structure of the entire neuropsychological battery. Once the structure is identified, factor-based difference scores can be computed to examine discrepancies between broad domains of functioning. In clinical settings where factor based difference scores cannot be calculated, the present findings suggest that equally weighted composite scores be computed instead of factor based difference scores at this step of interpretation. This is because the difference scores resulting from equally weighted composites are likely to have greater reliability than individual measure differences as long as the correlations between composites are not large. As a second interpretive step, more specific within and between-domain comparisons could be calculated using difference scores computed as part of each sub-battery. Lastly, difference scores between specific measures could be examined to further specify cognitive func-

tioning. Any information gleaned from the latter two steps should be considered tentative, however, given the lower reliability and greater number of comparisons required.

In conclusion, this study was concerned with comparing methods for improving the reliability of neuropsychological difference scores. This research did not attempt to establish the predictive validity of the resulting scores. By accomplishing the former goal, the present research demonstrated that orthogonal, factor-based differences have the potential to yield clinically useful information regarding cognitive strengths and weaknesses. However, studies of the predictive validity of orthogonal factor-based difference scores are necessary to firmly establish the usefulness of these scores and justify the extra time required to compute and interpret them.

REFERENCES

Bauer, R.M. (2000). The flexible battery approach to neuropsychological assessment. In R.D. Vanderploeg (Ed.), *Clinician's guide to neuropsychological assessment* (pp. 419–448). Mahwah, NJ: Lawrence Erlbaum Associates.

Benton, A.L., Hamsher, K., & de S. Sivan, A.B. (1994). *Multilingual aphasia examination*. Iowa City, IA: AJA Associates.

Bowden, S.C., Fowler, K.S., Bell, R.C., Whelan, G., Clifford, C.C., Ritter, A.J., & Long, C.M. (1998). The reliability and internal validity of the Wisconsin Card Sorting Test. *Neuropsychological Rehabilitation*, *8*, 243–254.

Buja, A. & Eyuboglu, N. (1992). Remarks on parallel analysis. *Multivariate Behavioral Research*, *27*, 509–540.

Carroll, J.B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York: Cambridge University Press.

Caruso, J.C. (2001a). Increasing the reliability of the Fluid/Crystallized Difference Score from the Kaufman Adolescent and Adult Intelligence Test with reliable component analysis. *Assessment*, *8*, 155–166.

Caruso, J.C. (2001b). Reliable components analysis of the Stanford-Binet: Fourth Edition for 2- to 6-year olds. *Psychological Assessment*, *13*, 261–266.

Caruso, J.C. & Cliff, N. (1998). The factor structure of the WAIS–R: Replicability across age groups. *Multivariate Behavioral Research*, *33*, 291–308.

Caruso, J.C. & Cliff, N. (1999). The properties of equally and differentially weighted WAIS–III factor scores. *Psychological Assessment*, *11*, 198–206.

Caruso, J.C. & Cliff, N. (2000). Increasing the reliability of Wechsler Intelligence Scale for Children–Third edition difference scores with reliable component analysis. *Psychological Assessment*, *12*, 89–96.

Caruso, J.C. & Cliff, N. (2002). *Reliable component analysis: A new/old method for exploratory data reduction*. Unpublished manuscript, University of Montana at Missoula and University of Southern California at Los Angeles.

Caruso, J.C. & Witkiewitz, K. (2001). Memory and reasoning abilities assessed by the Universal Nonverbal Intelligence Test: A reliable component analysis (RCA) study. *Educational and Psychological Measurement*, *61*, 5–22.

Caruso, J.C. & Witkiewitz, K. (2002). Increasing the reliability of ability–achievement difference scores: An example using the Kaufman Assessment Battery for Children. *Journal of Educational Measurement*, *39*, 39–58.

Cerhan, J.H., Ivnik, R.J., Smith, G.E., Tangalos, E.C., Petersen, R.C., & Boeve, B.F. (2002). Diagnostic utility of letter fluency, category fluency, and fluency difference scores in Alzheimer's disease. *Clinical Peuropsychologist*, *16*, 35–42.

Cliff, N. & Caruso, J.C. (1998). Reliable component analysis through maximizing composite reliability. *Psychological Methods*, *3*, 291–308.

Corey, D.M., Dunlap, W.P., & Burke, M.J. (1998). Averaging correlations: Expected values and bias in combining Pearson *r*'s and Fisher's *z* transformation. *Journal of General Psychology*, *125*, 245–262.

Dikmen, S.S., Heaton, R.K., Grant, I., & Temkin, N.R. (1999). Test–retest reliability and practice effects of expanded Halstead-Reitan Neuropsychological Test Battery. *Journal of the International Neuropsychological Society*, *5*, 346–356.

Fabrigar, L.R., Wegener, D.T., MacCallum, R.C., & Strahan, E.J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*, 272–299.

Fastenau, P.S., Denburg, N.L., & Mauer, B.A. (1998). Parallel short forms for the Boston Naming Test: Psychometric properties and norms for older adults. *Journal of Clinical and Experimental Neuropsychology*, *20*, 828–834.

Fennell, E.B. (1995). The role of neuropsychological assessment in learning disabilities. *Journal of Child Neurology*, *10*, S36–S41.

Franzen, M.D., Haut, M.W., Rankin, E., & Keefover, R. (1995). Empirical comparison of alternate forms of the Boston Naming Test. *Clinical Neuropsychologist*, *9*, 225–229.

Franzen, M.D., Paul, D., & Iverson, G.L. (1996). Reliability of alternate forms of the Trail Making Test. *Clinical Neuropsychologist*, *10*, 125–129.

Graham, J.W. & Schafer, J.L. (1999). On the performance of multiple imputation for multivariate data with small sample size. In R.H. Hoyle (Ed.), *Statistical strategies for small sample research* (pp. 1–29). Thousand Oaks, CA: Sage Publications, Inc.

Halstead, W.C. (1947). *Brain and intelligence*. Chicago: University of Chicago Press.

Heaton, R.K., Chelune, G.J., Talley, J.L., Kay, G.G., & Curtiss, G. (1993). *Wisconsin Card Sorting Test manual: Revised and expanded*. Odessa, FL: Psychological Assessment Resources.

Ingram, F., Greve, K. W., Ingram, P. T. F., & Soukup, V. M. (1999). Temporal stability of the Wisconsin Card Sorting Test in an untreated patient sample. *British Journal of Clinical Psychology*, *38*, 209–211.

Ivnik, R.J., Smith, G.E., Petersen, R.C., Boeve, B.F., Kokmen, E., & Tangalos, E.G. (2000). Diagnostic accuracy of four approaches to interpreting neuropsychological test data. *Neuropsychology*, *14*, 163–177.

Jacobson, M.W., Delis, D.C., Bondi, M.W., & Salmon, D.P. (2002). Do neuropsychological tests detect preclinical Alzheimer's disease: Individual–test versus cognitive-discrepancy score analyses. *Neuropsychology*, *16*, 132–139.

Johnstone, B. & Wilhelm, K.L. (1996). The longitudinal stability of the WRAT–R reading subtest: Is it an appropriate estimate of premorbid intelligence? *Journal of the International Neuropsychological Society*, *2*, 282–285.

Kaplan, E., Fein, D., Morris, R., & Delis, D.C. (1991). *The WAIS–R as a neuropsychological instrument. Manual*. San Antonio, TX: Psychological Corporation.

Kaplan, E., Goodglass, H., & Weintraub, S. (1983). *The Boston Naming Test*. Philadelphia: Lea and Febiger.

Kaufman, A.S. (1994). *Intelligent testing with the WISC–III*. New York: Wiley.

Keilp, J.G., Gorlyn, M., Alexander, G.E., Stern, Y., & Prohovnik, I. (1999). Cerebral blood flow patterns underlying the differential impairment in category vs. letter fluency in Alzheimer's disease. *Neuropsychologia*, *37*, 1251–1261.

Kelley, T.L. (1927). *Interpretation of educational measurements*. Yonkers, NY: World Books.

Langeluddecke, P.M. & Lucas, S.K. (2003). Quantitative measures of memory malingering on the Wechsler Memory Scale–Third Edition in mild head injury litigants. *Archives of Clinical Neuropsychology*, *18*, 181–197.

Lees-Haley, P.R., Smith, H.H., Williams, C.W., & Dunn, J.T. (1996). Forensic neuropsychological test usage: An empirical survey. *Archives of Clinical Neuropsychology*, *11*, 45–51.

Lezak, M.D. (1995). *Neuropsychological assessment* (3rd ed.). New York: Oxford University Press.

Lord, F.M. & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Macmann, G.M. & Barnett, D.W. (1997). Myth of the master detective: Reliability of interpretations for Kaufman's "intelligent testing" approach to the WISC–III. *School Psychology Quarterly*, *12*, 197–234.

Matarazzo, J.D. (1972). *Measurement and appraisal of adult intelligence*. Baltimore: Williams & Wilkins.

McCrae, R.R., Zonderman, A.B., Costa, P.T., Bond, M.H., & Paunonen, S.V. (1996). Evaluating Replicability of factors in the revised NEO Personality Inventory: Confirmatory factor analysis versus procrustes rotation. *Journal of Personality and Social Psychology*, *70*, 552–566.

McDermott, P.A., Fantuzzo, J.W., Glutting, J.J., Watkins, M.W., & Baggaley, A.R. (1992). Illusions of meaning in the ipsative assessment of children's ability. *Journal of Special Education*, *25*, 504–526.

Millis, S.R., Ross, S.R., & Ricker, J.H. (1998). Detection of incomplete effort on the Wechsler Adult Intelligence Scale–Revised: A cross-validation. *Journal of Clinical and Experimental Neuropsychology*, *20*, 167–173.

Mitrushina, M.N., Boone, K.B., & D'Elia, L.F. (1999). *Handbook of normative data for neuropsychological assessment*. New York: Oxford University Press.

Nunnally, J.C. & Bernstein, I.H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill, Inc.

O'Connor, B.P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods Instruments and Computers*, *32*, 396–402.

Reitan, R.M. (1955). Investigation of the validity of Halstead's measures of biological intelligence. *Archives of Neurology and Psychiatry*, *73*, 28–35.

Reitan, R.M. (1958). Validity of the trail making test as an indicator of organic brain damage. *Perceptual and Motor Skills*, *8*, 271–276.

Russell, E.W. (2000). The cognitive-metric, fixed battery approach to neuropsychological assessment. In R.D. Vanderploeg (Ed.), *Clinician's guide to neuropsychological assessment* (pp. 449–482). Mahwah, NJ: Lawrence Erlbaum Associates.

Schafer, J.L. (2002). Multiple imputation with PAN. In L. Collins & G. Aline (Eds.), *New methods for the analysis of change* (pp. 357–377). Washington, DC: American Psychological Association.

Scott, J.G., Krull, K.R., Williamson, D.J.G., Adams, R.L., & Iverson, G.L. (1997). Oklahoma Premorbid Intelligence Estimation (OPIE): Utilization in clinical samples. *Clinical Neuropsychologist*, *11*, 146–154.

Soper, H.V., Cicchetti, D.V., Satz, P., Light, R., & Orsini, D.L. (1988). Null hypothesis disrespect in neuropsychology: Dangers of alpha and beta errors. *Journal of Clinical and Experimental Neuropsychology*, *10*, 255–270.

Spreen, O. & Benton, A.L. (1969). *Neurosensory Center Comprehensive Examination for Aphasia: Manual of directions*. Victoria, British Columbia, Canada: Neuropsychology Laboratory, University of Victoria.

Spreen, O. & Strauss, E. (1998). *A compendium of neuropsychological tests: Administration, norms, and commentary* (2nd ed.). New York: Oxford University Press.

SPSS, Inc. (2002). SPSS Professional Statistical Package (Version 11.0). Chicago: Author.

Streiner, D.L. & Norman, G.R. (1995). *Health measurement scales: A practical guide to their development and use* (2nd ed.). New York: Oxford University Press.

Tarter, R.E. & Edwards, K.L. (1986). Neuropsychological assessment. In T. Incagnoli, G. Goldstein, & C.J. Golden (Eds.), *Clinical application of neuropsychological test batteries* (pp. 135–153). New York: Plenum Press.

Tate, R.L., Perdices, M., & Maggiotto, S. (1998). Stability of the Wisconsin Card Sorting Test and the determination of reliability of change in scores. *Clinical Neuropsychologist*, *12*, 348–357.

Tucker, L.R. (1951). *A method for synthesis of factor analysis studies (Personnel Research Section Report No. 984)*. Washington, DC: Department of the Army.

Wechsler, D. (1997a). *Manual for the Wechsler Adult Intelligence Scale–Third Edition*. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1997b). *Manual for the Wechsler Memory Scale–Third Edition*. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1997c). *WAIS–III WMS–III technical manual*. San Antonio, TX: The Psychological Corporation.

Widaman, K.F. (1993). Common factor analysis versus principal component analysis: Differential bias in representing model parameters? *Multivariate Behavioral Research*, *28*, 263–311.

Wilde, N., Strauss, E., Chelune, G.J., Loring, D.W., Martin, R.C., Hermann, B.P., Sherman, E.M.S., & Hunter, M. (2001). WMS–III performance in patients with temporal lobe epilepsy: Group differences and individual classification. *Journal of the International Neuropsychological Society*, *7*, 881–891.

Wilkinson, G.S. (1993). *The Wide Range Achievement Test–3*. San Antonio, TX: The Psychological Corporation.

Williams, V.S.L., Jones, L.V., & Tukey, J.W. (1999). Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. *Journal of Educational and Behavioral Statistics*, *24*, 42–69.

Zwick, W.R. & Velicer, W.F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, *99*, 432–442.