Journal of the International Neuropsychological Society (2008), **14**, 1069–1073.

doi:10.1017/S1355617708081228

BRIEF COMMUNICATION

On the reliability and standard errors of measurement of contrast measures from the D-KEFS

JOHN R. CRAWFORD,[1] DAVID SUTHERLAND,[1] AND PAUL H. GARTHWAITE[2]

[1]School of Psychology, University of Aberdeen, Aberdeen, United Kingdom
[2]Department of Mathematics and Statistics, The Open University, Milton Keynes, United Kingdom

(RECEIVED February 1, 2008; FINAL REVISION June 25, 2008; ACCEPTED June 27, 2008)

Abstract

A formula for the reliability of difference scores was used to estimate the reliability of Delis-Kaplan Executive Function System (D-KEFS; Delis et al., 2001) contrast measures from the reliabilities and correlations of their components. In turn these reliabilities were used to calculate standard errors of measurement. The majority of contrast measures had low reliabilities: of the 51 reliability coefficients calculated in the present study, none exceeded 0.7 and hence all failed to meet any of the criteria for acceptable reliability proposed by various experts in psychological measurement. The mean reliability of the contrast scores was 0.27, the median reliability was 0.30. The standard errors of measurement were large and, in many cases, equaled or were only marginally smaller than the contrast scores' standard deviations. The results suggest that, at present, D-KEFS contrast measures should not be used in neuropsychological decision making. (*JINS*, 2008, *14*, 1069–1073.)

Keywords: Executive functioning, Neuropsychological assessment, Difference scores, Psychometrics, Measurement error, Quantitative methods

INTRODUCTION

The publication of the Delis-Kaplan Executive Function System (D-KEFS; Delis et al., 2001) is a positive development in the assessment of executive functioning. The D-KEFS gathers together an extensive range of some of the best available measures of executive functioning, offers a carefully standardized administration, and provides norms based on a large, stratified, census-matched, sample. Some reviewers, however, have expressed concern over the reliability of D-KEFS scores (Baron, 2004; Schmidt, 2003; Strauss et al., 2006). Other reviewers (Homack et al., 2005; Shunk et al., 2006), and the authors of the test (Delis et al., 2004), have been more sanguine. Shunk et al. (2006), for example, conclude that the D-KEFS is "psychometrically sound" (p. 275) and notes that low reliability "has been a popular criticism of the D-KEFS system but does not pose serious concern" (p. 277).

Correspondence and reprint requests to: Professor John R. Crawford, School of Psychology, College of Life Sciences and Medicine, King's College, University of Aberdeen, Aberdeen AB24 3HN, United Kingdom. E-mail: j.crawford@abdn.ac.uk

To date, the debate over the reliability of D-KEFS scores has focused on the reliability coefficients presented in the D-KEFS test manual. The present paper is concerned, not with existing reliability information, but with the reliability of the D-KEFS contrast scores. Contrast scores allow neuropsychologists to examine the discrepancies between related measures (for example, an individual's category fluency score can be compared to her/his letter fluency score). Like other D-KEFS measures, they are standardized to have a mean of 10 and standard deviation of 3.

The provision of contrast scores is in keeping with the process approach to neuropsychological assessment advocated by the test's authors (Delis et al., 2001). However, there are at least two reasons to be cautious about the use of contrast scores in neuropsychological decision making. First, although the rationale for the use of discrepancy scores may appear compelling, empirical studies of their ability to differentiate between healthy and impaired populations have often produced disappointing results (Smith et al., 2008).

Second, there is the danger that the reliability of discrepancy scores/contrast scores will be unacceptably low. When, as is the case with the D-KEFS contrast scores, the two

components used to form the contrast have the same standard deviation, the reliability of the difference score is a simple function of the correlation between the two components and their reliabilities. Thus there are two sources of measurement error present in a contrast score rather than the one present in a simple score. Also, as contrast scores typically compare measures of two related constructs, the correlation between the components will often be moderate and may approach the reliabilities of the components in its magnitude. In this situation the variance of the contrast score will predominantly be measurement error variance.

In view of the foregoing concerns over contrast scores, it is important that neuropsychologists have access to information on the reliability and standard errors of measurement of the D-KEFS contrast scores. Unfortunately no such information is provided in the D-KEFS manual. This omission runs counter to expert advice. For example, Standard 2.1 of the authoritative Standards for Educational and Psychological Testing (1999) states, "For each total score, subscore, or combination of scores that is to be interpreted, estimates of relevant reliabilities and standard errors of measurement . . . should be reported." (p. 31). Standard 2.3 makes it explicit that difference scores (i.e., contrast scores) should not be regarded as exempt from these requirements. It states, "When test interpretation emphasizes differences between two observed scores of an individual . . . , reliability data, including standard errors, should be provided for such differences." (p. 32). The primary aim of the present study is to attempt to provide the information recommended earlier by estimating the reliability of the contrast scores and thereby also providing accompanying standard errors of measurement.

METHOD

This research was approved by the School of Psychology Ethics Committee, University of Aberdeen, and was conducted in accordance with the Helsinki Declaration.

Reliability of Contrast Scores

When the two components have a common standard deviation, the formula for the reliability of a difference score (e.g., Crocker & Algina, 1986) is simply

$$r_{DD} = \frac{0.5(r_{XX} + r_{YY}) - r_{XY}}{1 - r_{XY}}, \qquad (1)$$

where r_{XX} and r_{YY} are the reliabilities of the two components entering into to the difference (i.e., contrast) score, and r_{XY} is the correlation between them. The reliability coefficients thus obtained can then be used, along with the standard deviation of contrast scores (which is 3 in all cases) to obtain standard errors of measurement.

Ideally, the reliabilities and correlation between the components of difference scores should be obtained from the same test session (i.e., internal consistency coefficients should be used). However, as the authors of the D-KEFS point out, the nature of many of the subtests preclude calculation of these coefficients (Delis et al., 2001) and therefore they used the test-retest method to estimate reliability for the majority of subtests.

The test-retest reliabilities of the two components of each of the contrast scores along with their intercorrelations were obtained from the D-KEFS technical manual for each of the three standardization sample age bands (8–19, 20–49, and 50–89 years). However, for the Verbal Fluency and Sorting subtests it was also possible to use internal consistency data on the pairs of components used to derive the contrasts. For these latter estimates, some pre-processing of the data was required: The internal consistency estimates in the test manual were presented for a finer gradation of age bands (16 in all) than that used elsewhere in the manual. These reliability estimates were averaged (*via* Fisher's z transformation) to obtain the averaged reliability within each of the three principal age groups.

In the case of the D-KEFS Color-Word Interference test, one of the contrast scores compares Inhibition/Switching with combined Naming + Reading. Test-retest reliability data are not presented for combined Naming + Reading. However, internal consistency coefficients are presented and so the reliability was estimated using the averaged internal consistency coefficients within each of the three age groups.

RESULTS AND DISCUSSION

Reliabilities for the Trail Making Test

It can be seen from Table 1 that the estimated reliabilities for the five Trail Making Test contrast scores show considerable variability. However, they are generally very low, ranging from −0.23 to .49. Note that many of the reliability coefficients take negative values. The term in the numerator for the reliability of a difference score (Formula 1) consists of the average of the reliabilities of the two components minus the correlation between them. Thus, when the correlation between the components exceeds the average reliability, the reliability of the difference score will take a negative value. When this occurs it is taken as an indication that the true reliability of the difference score is either zero or, at best, low.

There are two main factors that contribute to the possibility of obtaining negative estimates. First, when the samples used to estimate the reliability of the scores and their correlations are modest in size, the sample estimates of the true reliabilities and correlations will be subject to considerable error. Thus, for example, if the sample reliability coefficients for the components are lower than the true reliabilities, the reliability of the contrast score will be underestimated; the underestimation will be particularly marked if, by chance, the sample correlation between the two components overestimates the true correlation.

Table 1. Estimated reliability of the 14 D-KEFS contrast scores in the three standardization sample age bands: The reliabilities of the components were based on test-retest reliability coefficients (for three of the contrast scores reliabilities were also estimated using internal consistency data, these are indicated by the suffix IC)

	Reliability in each age band		
Subtest: Contrast measure	8–19	20–49	50–89
TMT: Number-Letter Switching *vs*. Visual Scanning	0.145	0.335	0.406
TMT: Number-Letter Switching *vs*. Number Sequencing	0.096	−0.058	−0.080
TMT: Number-Letter Switching *vs*. Letter Sequencing	−0.118	−0.234	0.107
TMT: Number-Letter Switching *vs*. (Number + Letter Seq)	0.000	−0.136	−0.063
TMT: Number-Letter Switching *vs*. Motor Speed	0.364	0.424	0.493
VF: Letter Fluency *vs*. Category Fluency	0.300	0.476	0.659
VF: Category Switching *vs*. Category Fluency	0.387	0.314	0.460
VF: Letter Fluency *vs*. Category Fluency **(IC)**	0.356	0.415	0.477
VF: Category Switching *vs*. Category Fluency **(IC)**	0.292	0.294	0.339
DF: Switching *vs*. (Filled Dots + Empty Dots)[1]	0.034	0.164	0.373
CWI: Inhibition *vs*. Color Naming	0.696	0.552	0.021
CWI: Inhibition/Switching *vs*. Color Naming	0.653	0.367	0.250
CWI: Inhibition/Switching *vs*. Word Reading	0.629	0.048	0.263
CWIInhibition/Switching *vs*. Inhibition	0.651	−0.041	−0.107
CWI: Inhibition/Switching *vs*. (Naming + Reading)	0.557	0.222	0.425
ST: Sort Recognition *vs*. Free Sorting Description	0.061	−0.547	0.179
ST: Sort Recognition *vs*. Free Sorting Description (IC)	0.256	0.328	0.449

Note. TMT = Trail Making Test; VF = Verbal Fluency; DF = Design Fluency; CWI = Color-Word Interference; ST = Sorting Test. [1]The reliabilities for Filled Dots + Empty Dots in each age band were calculated by the present authors using the formula for the reliability of a composite.

Negative estimates of reliabilities can also be obtained when between-component items genuinely share more variance than do within-component items (the correlation will then exceed the averaged reliability). In this latter case the problem is one of construct validity rather than reliability. Both of these factors may contribute to the negative values obtained for the D-KEFS contrast scores. Certainly the sample sizes used to estimate the test-retest reliabilities were very modest (28, 35, and 38 for the 8–19, 20–49, and 50–89 age bands respectively). Moreover, some of the pairs of components have highly similar task demands. The differences between the tasks that the tests' authors believe to be crucial may in fact not be, whereas, within each task, the cognitive demands may change as the task progresses and this recruitment of different cognitive processes may be common to both components (hence the between-component variance may exceed the within-component variance).

Reliabilities of the Remaining D-KEFS Contrast Scores

The reliabilities of Verbal Fluency contrast scores are generally higher than those obtained for the Trail Making test (and none are negative). However, the reliabilities are still, in most cases, disappointingly low; they range from .29 to .66. The results obtained when the test-retest reliabilities of the components were used as inputs *versus* those obtained using the internal consistency coefficients do not differ dramatically, although (with one exception) the test-retest reliabilities yielded higher reliabilities.

Design Fluency has only one contrast score: the reliability is low in all three age groups (ranging from .3 to .37). For the five Color-Word Interference contrast scores the results are again disappointing. The most striking feature of these latter results is that the reliabilities of the contrast scores are generally low for the two older age groups (reliabilities range from − .11 to .55), whereas they are much higher for the youngest age group (where they range from .56 to .70).

The Sorting Test has only one contrast score (but two estimates of its reliability as internal consistency data were available). Although both sets of estimates are low, it can be seen that, in this case, the results are particularly poor for the test-retest reliabilities.

Averaged Reliability of D-KEFS Contrast Scores and Standards for Reliability

The average reliability (across all contrast scores and age bands) was 0.27 with a median of .30. Various systems for classifying the adequacy of the reliability of psychological tests have been proposed by experts in measurement. Of the 51 reliability coefficients calculated for the contrast scores in the present study, none met Nunnally and Bernstein's

(1994) requirement of reliabilities of .90 or above. Moreover, none of the contrast scores can be classified as "reliable" (>.70) according to Sattler (2001). Similarly, the reliabilities of all of the contrast scores were "unacceptable" according to Cicchetti's (1994) classification system, that is, all fell below .70.

Standard Errors of Measurement for the D-KEFS Contrast Scores

The standard errors of measurement (SEM) for the contrast scores are presented in Table 2. The measurement error variance of test scores cannot exceed the total variance. Therefore, contrast scores with negative estimated reliabilities were assumed to have zero reliability and the SEM was set equal to the standard deviation of obtained scores (3) in such cases.

The standard errors of measurement in Table 2 can be used to set confidence intervals on D-KEFS contrast scores. Experts on psychological measurement are unanimous in recommending that test scores should be accompanied by confidence intervals. These intervals serve the general purpose of reminding us that scores are fallible (i.e., they avoid reifying the observed score) and serve the specific and practical purpose of quantifying the effects of such fallibility (Crawford & Garthwaite, 2008).

As can be anticipated from the foregoing results, the confidence intervals will be wide for most of contrast scores. That is, a high degree of uncertainty over an individual's true contrast score will be the rule rather than the exception. For example, the median reliability for the contrast scores was 0.30 (for the Letter Fluency *vs*. Category Fluency contrast in the 8–19 age group based on test-retest reliabilities). The accompanying standard error of measurement was 2.51. Suppose an individual's contrast score is 10. The SEM multiplied by 1.96 is 4.92: adding and subtracting this quantity from the obtained score (and rounding) gives a 95% confidence interval for the score of 5 to 15. Expressing this interval in the form of percentile ranks (as recommended by Crawford & Garthwaite, 2008) the limits are from the 5th percentile to the 95th percentile.

Should the D-KEFS Contrast Scores be Interpreted?

The reliability of a difference score is constrained by the reliabilities of its components. Therefore, given that the reliabilities of many of the D-KEFS subtests are themselves modest (Schmidt, 2003; Strauss et al., 2006), it was not expected that the contrast scores would be very reliable. However, the results were particularly disappointing. Indeed, based on the present analysis the majority of D-KEFS contrast scores should be considered to be uninterpretable. A reliability coefficient is an estimate of the proportion of test variance that is true variance and most of the D-KEFS contrast scores this proportion is low. That is, the indications are that most of the variance of these scores is simply measurement error variance.

Table 2. Standard errors of measurement (SEM) for the 14 D-KEFS contrast scores in the three standardization sample age bands: Where the reliability of a contrast score was estimated to be negative the SEM was set equal to the standard deviation of the score (SEMs based on use of internal consistency data are indicated by the suffix IC)

	SEM in each age band		
Subtest: Contrast measure	8–19	20–49	50–89
TMT: Number-Letter Switching *vs*. Visual Scanning	2.77	2.45	2.31
TMT: Number-Letter Switching *vs*. Number Sequencing	2.85	3.00	3.00
TMT: Number-Letter Switching *vs*. Letter Sequencing	3.00	3.00	2.84
TMT: Number-Letter Switching *vs*. (Number + Letter Seq)	3.00	3.00	3.00
TMT: Number-Letter Switching *vs*. Motor Speed	2.39	2.28	2.14
VF: Letter Fluency *vs*. Category Fluency	2.51	2.17	1.75
VF: Category Switching *vs*. Category Fluency	2.35	2.49	2.21
VF: Letter Fluency *vs*. Category Fluency (IC)	2.41	2.30	2.17
VF: Category Switching *vs*. Category Fluency (IC)	2.52	2.52	2.44
DF: Switching *vs*. (Filled Dots + Empty Dots)	2.95	2.74	2.38
CWI: Inhibition *vs*. Color Naming	1.65	2.01	2.97
CWI: Inhibition/Switching *vs*. Color Naming	1.77	2.39	2.60
CWI: Inhibition/Switching *vs*. Word Reading	1.83	2.93	2.58
CWI: Inhibition/Switching *vs*. Inhibition	1.77	3.00	3.00
CWI: Inhibition/Switching *vs*. (Naming + Reading)	2.00	2.65	2.28
ST: Sort Recognition *vs*. Free Sorting Description	2.91	3.00	2.72
ST: Sort Recognition *vs*. Free Sorting Description (IC)	2.59	2.46	2.23

Note. TMT = Trail Making Test; VF = Verbal Fluency; DF = Design Fluency; CWI=Color-Word Interference; ST = Sorting Test.

We turn now to briefly consider of the types of evidence that might lead to a modification or even rejection of these preliminary, pessimistic conclusions. One possibility is that future reliability studies will provide more encouraging results. For example, an alternative approach to estimating the reliability and standard errors for the D-KEFS contrast scores would simply be to correlate the contrast scores obtained on two occasions; that is, to calculate test-retest reliabilities for the contrast scores (and thereby also obtain standard errors) just as was done for the other D-KEFS scores.

The estimates of reliabilities obtained using this approach are liable to be somewhat higher than those reported here because it does not involve mixing the estimate of the correlation between tests obtained from one test session with the estimates of reliability obtained from two sessions. However, the reliability estimates provided by this alternative approach are not liable to be dramatically different from those reported here and, unfortunately, for most of the contrast scores, dramatic differences would be required before the reliabilities could be considered adequate.

A demonstration that contrast scores have large effect sizes (or high sensitivity and specificity) when cognitively intact samples are compared to various clinical populations (e.g., patients with focal frontal lesions) would also provide support for the use of contrast scores. Such results are not impossible: It has been suggested that, paradoxically, unreliable scores can still, in theory, possess sufficient power to detect group differences (see Strauss, 2001 for a brief commentary on this controversial topic). However, classic treatments (e.g., Chapman & Chapman, 1973) of the effects of the reliability of measures on the ability to differentiate between groups would argue against the likelihood of strongly positive outcomes.

In summary, at present it would be imprudent to use D-KEFS contrast scores in arriving at a formulation of an individual's cognitive strengths and weaknesses. The neuropsychological community should be willing to be convinced otherwise but, in view of the present results, the burden of proof must lie firmly with those who would advocate their use.

ACKNOWLEDGMENTS

The first author (JRC) undertakes consultancy for the Psychological Corporation (publishers of the D-KEFS). This work was undertaken while one of the authors (PHG) was a visiting academic at the University of New South Wales, Sydney, and was conducted without the support of a research grant or contract.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Baron, I.S. (2004). Delis-Kaplan Executive Function System. *Child Neuropsychology*, *10*, 147–152.

Chapman, L.J. & Chapman, J.P. (1973). Problems in the measurement of cognitive deficit. *Psychological Bulletin*, *79*, 380–385.

Cicchetti, D.V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*, 284–290.

Crawford, J.R. & Garthwaite, P.H. (2008). Percentiles please: The case for expressing neuropsychological test scores and accompanying confidence limits as percentile ranks. *Clinical Neuropsychologist* (in press).

Crocker, L.M. & Algina, J. (1986). *Introduction to classical and modern test theory*. Fort Worth, TX: Holt, Rhinehart & Winston.

Delis, D.C., Kaplan, E., & Kramer, J. (2001). *Delis Kaplan Executive Function System*. San Antonio, TX: Psychological Corporation.

Delis, D.C., Kramer, J.H., Kaplan, E., & Holdnack, J. (2004). Reliability and validity of the Delis-Kaplan executive function system: An update. *Journal of the International Neuropsychological Society*, *10*, 301–303.

Homack, S., Lee, D., & Riccio, C.A. (2005). Test review: Delis-Kaplan executive function system. *Journal of Clinical and Experimental Neuropsychology*, *27*, 599–609.

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

Sattler, J.M. (2001). *Assessment of children: Cognitive applications* (4th ed.). San Diego, CA: Sattler.

Schmidt, M. (2003). Hit or miss? Insight into executive functions. *Journal of the International Neuropsychological Society*, *9*, 962–964.

Shunk, A.W., Davis, A.S., & Dean, R.S. (2006). Test review of the Delis-Kaplan Executive Function System. *Applied Neuropsychology*, *13*, 275–279.

Smith, G.E., Ivnik, R.J., & Lucas, J. (2008). Assessment techniques: Tests, test batteries, norms and methodological approaches. In J. E. Morgan & J. H. Ricker (Eds.), *Textbook of clinical neuropsychology* (pp. 41–60). New York: Taylor and Francis.

Strauss, E., Sherman, E.M.S., & Spreen, O. (2006). *A compendium of neuropsychological tests: Administration, norms and commentary* (3rd ed.). New York: Oxford University Press.

Strauss, M.E. (2001). Demonstrating specific cognitive deficits: A psychometric perspective. *Journal of Abnormal Psychology*, *110*, 6–14.