Article

Applied Psychological Measurement
1–7

DOI: 10.1177/0146621615577361
apm.sagepub.com
SAGE

An Evaluation of the Consistency and Reliability of the Defense Automated Neurocognitive Assessment Tool

C. R. Russo[1] and C. E. Lathan[1]

Abstract

A durable, portable, and field-hardened computerized neurocognitive test (CNT) called the Defense Automated Neurobehavioral Assessment (DANA) tool was recently developed to provide a practical means to conduct neurological and psychological assessment in situ. The psychometric properties of the DANA have been previously described. This present work discusses the test–retest reliability of the DANA Rapid test battery, as administered to a homogeneous population of U.S. Air Force Academy football team players (N = 162) across the duration of the season. The intraclass correlation coefficient (ICC) metric of the DANA is compared with that from two different CNTs recently reported in Cole et al., and the implications of using the metric to interpret comparative test reliability among different CNTs are discussed.

Keywords

reliability, cognitive assessment, screening, traumatic brain injury, concussion, sports concussion, CNT

Introduction

Although the efficiency and utility of a neurocognitive assessment tool is determined by its internal consistency, and the measure of test–retest reliability is typically described as the foundation on which a test's validity is established (Cole et al., 2013), a consistent methodology for quantifying reliability is not clearly demonstrated in the current neuropsychology literature. Disconcertingly, recent works have found that the reliability coefficients of computerized neurocognitive tests (CNTs) tend to fall below the bar of ''clinical acceptability'' (Broglio, Ferrara, Macciocchi, Baumgartner, & Elliot, 2007; Cole et al., 2013; Resch et al., 2013; Segalowitz et al., 2007). Differences in the characteristics of test batteries, differences in the design of test–retest studies, and insufficiently explained and non-standardized methods of analysis have all served to confound the matter of clearly defining a quantifiable measure of reliability (Christie, Kamen, Boucher, Inglis, & Gabriel, 2010; Weir, 2005).

[1]AnthroTronix Inc., Silver Spring, MD, USA

Corresponding Author:
C. R. Russo, AnthroTronix Inc., 8737 Colesville Rd., Suite L203, Silver Spring, MD 20910, USA.
Email: clementina.russo@atinc.com

The definition of the reliability coefficient is both contextual and application-specific (e.g., Baumgartner, 1969; Feldt & McKee, 1958; Streiner & Norman, 1995), and as a result, any given reliability coefficient is not a universal measure of reliability. For example, there is still a pervasive reporting of Pearson's *r* to assess reliability even though its use is actively discouraged due to the model's inherent disregard for systematic error (Baumgartner, 2000; Bedard, Martin, Krueger, & Brazil, 2000; Kroll, 1962; Ludbrook, 2002; Safrit, 1976) and is therefore an inappropriate metric for certain applications.

In neuropsychology, the intraclass correlation coefficient (ICC; as formalized by Shrout & Fleiss, 1979, and then updated by McGraw & Wong, 1996) is often reported as the stand-alone metric of test–retest reliability. By definition, the ICC calculation entails six different possible configuration parameters by which the coefficient is determined, and each model's estimate is unique. A side-by-side comparison of the methods of recent works (such as Broglio et al., 2007; Cole et al., 2013; Resch et al., 2013) reveals that the ICC approach is often not applied in a standardized way, possibly because the ICC model itself is not well-understood. Implicit disagreement between research groups pertaining to which ICC model most accurately describes the test–retest design may stem from confusion around the applicability of the ICC model as developed for inter-rater reliability rather than for test–retest reliability (Weir, 2005).

These two types of reliability describe distinctively different situations: Inter-rater reliability tests the hypothesis that a heterogeneous group of judges (the raters) similarly rate the same set of subjects (the ratees) across multiple testing sessions. Test–retest reliability tests the hypothesis that the subjects themselves (the ratees) perform the same way across the sessions and assumes the raters (the computers) are the same. As a result, the ICC was designed to be relatively insensitive to within-subject, session-to-session variability, and is thus only informative of test–retest reliability *by proxy*. Owing to this detail, other fields (e.g., exercise and sports science, sports medicine, and physical therapy) report the ICC along with a precision metric provided by the standard error of the measurement (SEM) that offers an absolute bound on the measurement of interest (Denegar & Ball, 1993; Learmonth, Dlugonski, Pilutti, Sandroff, & Motl, 2013). It is crucial to note that for the purpose of comparing the reliability of different CNTs, it is particularly important to report the reliability coefficient accompanied by a precision metric.

The SEM carries the same units as the measurement of interest (e.g., throughput or reaction time) and it is informative of within-subject reliability. It can be obtained from the reliability coefficient (this method provides the coefficient's precision), or independently of the reliability coefficient, from the square root of the mean square error. In either case, the minimum difference (MD) is directly constructed from the SEM and describes the minimum amount of change in results required to be considered a real effect and not an artifact of associated error (Weir, 2005).

This work describes consistent test and retest measurements of the Defense Automated Neurobehavioral Assessment (DANA) test as administered to a homogeneous population of U.S. Air Force Academy football team players ($N = 162$) through time points across the season. The ICC metric of the DANA is compared with that from two different CNTs recently reported in Cole et al. (2013), and the implications of using the metric to interpret comparative reliability among different CNTs are discussed.

Method

The DANA Rapid Tests and Administration

The psychometric properties of the DANA test batteries have been previously described and evaluated (Lathan, Spira, Bleiberg, Vice, & Tsao, 2013). Data for this study were collected under the U.S. Air Force Academy performance improvement protocol. The DANA Rapid test

Table 1. Summary Statistics for Each Testing Session Date (Denoted as $T_{1,2}$).

		<TP>		*SD* <TP>			*R*		
Battery	Subtest	T_1	T_2	T_1	T_2	*n*	LB	UB	MD
DANA	PRT	102.9	106.8	13.02	12.57	89	0.75		18.04
							0.60	0.84	
	RT	192.4	198.3	26.77	23.54	87	0.81		32.34
							0.69	0.87	
ANAM4	PRT	97.62	103.56	13.10	10.31	50	0.51		25.42
							0.28	0.69	
	RT	86.88	91.38	11.46	12.76	50	0.60		22.37
							0.39	0.75	
ImPACT	RT	0.60	0.61	0.08	0.09	44	0.53		0.17
							0.28	0.71	

Note. Reported mean throughput (<TP>) is the mean throughput of *n* total subjects (matched for test and retest), and *SD* <TP> is the associated standard deviation of the mean throughput of *n* total subjects. For the DANA data, *R* was calculated by the way of Equation 2, presented with LB and UB (constructed in the 95% confidence interval). With respect to the ANAM4 and ImPACT data listed here from Cole et al. (2013), the ICC model used to calculate *R* which was not reported in that work. Note that the reported metric for ImPACT is not <TP>, but rather ImPACT's composite score. The MD is calculated from Equations 3 and 4, using the greater standard deviation between the two time points per subtest, per test battery. TP = throughput; LB = lower bound; UB = upper bound; MD = minimum difference; DANA = Defense Automated Neurobehavioral Assessment; ANAM = Automated Neurobehavioral Assessment; ImPACT = Immediate Post-Concussion Assessment and Cognitive Testing; RT = Reaction Time; PRT = Procedural Reaction Time.

battery consists of three cognitive tests given in succession, each of which measures reaction time (see Table I in Lathan et al., 2013). On a given testing date, U.S. Air Force cadets participating on the Air Force Academy football team were administered the DANA Rapid along with a demographic survey that were both loaded onto a collection of identical mobile devices (Trimble Nomads, model 900S). The test administration time totaled about 5 min.

Data were collected at the beginning of the season on August 22 to 24, 2012, in the middle of the season on November 6 to 7, 2012, and at the end of the season on April 30 to May 1, 2013. If a subject took more than one administration of a test in any given testing session, then only the first administration was included in the following analysis. In addition, a subject must have correctly responded to more than 66% of the test stimuli. Test–retest reliability was calculated from the scores of the first test (or only test) administered per testing date, tabulated for the same subjects across the season. The test–retest reliability (reported ICC) of the DANA is compared with that from two of the CNTs reported in Cole et al. (2013), for comparable neurocognitive subtests. The retest duration specified in Cole et al. was 21 to 42 days post-baseline, and thus for better comparison, the shortest period retest of the DANA that was administered is reported, approximately 77 days post-baseline (August-November administrations).

For each of *n* subjects, a subject's mean throughput (<TP>, with units of min^{-1}) was calculated from correctly answered mean response time data (<$\text{RT}_{\text{correct}}$>, with units of milliseconds),

$$<\text{TP}> = \frac{\frac{\text{total}_{\text{correct}}}{\#\text{trials}}}{<\text{RT}_{\text{correct}}>} \times 60,000. \quad (1)$$

The factor of 60,000 converts milliseconds to minutes. To assess consistency in <TP> in the DANA data across time points, a multi-way repeated-measures ANOVA was performed for all of the subjects across testing sessions, per subtest.

The ICC

In essence, the ICC is a relative (unitless) measure of test error with regard to between-subject variability. The interpretation of the ICC is that it represents the score variance that is attributable to the variability between subjects and the remainder of the variance is attributable to error. Because systematic differences can explain variability in the data, the choice of ICC model parameters is dependent on the study design and the characteristics of the test subjects. Test–retest designs lend themselves to two-way model analyses, and for situations in which the testing devices are not themselves variable (identical and electronically stable), use of the average measures parameter is appropriate as it will not over-compensate for error and thus artificially deflate the resulting coefficient (Christie et al., 2010; Weir, 2005). The two-way, average measures model is given as,

$$\mathrm{ICC}\{2,k\} = \frac{\mathrm{MS}_S - \mathrm{MS}_E}{\mathrm{MS}_S + \frac{k(\mathrm{MS}_T - \mathrm{MS}_E)}{n}}, \tag{2}$$

where $\mathrm{MS}_{S,E,T}$ denote the mean squares (derived from the sum of squares resulting from a repeated-measures ANOVA calculation) of subjects (S), systematic error (E), and trials (T); k is the number of test administrations and n is the number of subjects. On the DANA data presented in this work, the ICC was tabulated using a *MATLAB* script (Salarian, 2008) that follows the formalism in McGraw and Wong (1996).

SEM and MD

The SEM is an absolute index of reliability and provides insight into the session-to-session noise in a given set of data. It carries the same units as the measurement of interest and can be interpreted as the reliability within individual subjects (Shrout, 1998; Weir, 2005). The SEM can be found both from the ICC, or estimated independently from the ICC as the square root of the mean square error (MS_E; Eliasziw, Young, Woodbury, & Fryday-Field, 1994; Hopkins, 2000; Stratford & Goldsmith, 1997; Weir, 2005). Following the formalism in Weir (2005), from the ICC, the SEM is written as,

$$\mathrm{SEM}_{\mathrm{ICC}} = SD\sqrt{1 - \mathrm{ICC}}, \tag{3}$$

where SD is the standard deviation of subjects' scores.

Alternatively, the ICC-independent form of the SEM is calculated using the MS_E, related to the sum of squares error, SS_E, found from the ANOVA calculation, $\mathrm{MS}_E = \mathrm{SS}_E/((n-1)(k-1))$, where n is the number of subjects and k is the number of test administrations. The ICC-independent form of the SEM is the square root of the MS_E ($\mathrm{SEM} = \sqrt{\mathrm{MS}_E}$).

In either case, the SEM is the basis of the MD index or the minimum increment of observable change that warrants consideration as a real change in score and likely not attributable to error:

$$\mathrm{MD} = \mathrm{SEM} \times z \times \sqrt{2}, \tag{4}$$

where the $\sqrt{2}$ is an artifact of the standard error of the difference of two score results from test and retest administrations. In Equation 4, z is the distribution score used to construct the confidence interval. In this work, the MD is reported as an absolute index for reliability, constructed in the 95% confidence interval for which $z = 1.96$.

Results and Discussion

The results from all tested batteries are captured in Table 1. The reliability coefficient measured for DANA, for matching subjects across test and retest sessions, is found to be higher than those from both the Automated Neurobehavioral Assessment Metric (ANAM), and the Immediate Post-Concussion Assessment and Cognitive Testing (ImPACT) tools, and comparatively, the DANA exhibits a reliability coefficient within tighter upper and lower bounds than both ANAM and ImPACT. A multi-way repeated-measures ANOVA was performed for all of the subjects per DANA subtest, grouped by testing session, and the results (not reported, for brevity) showed that the marginal means of <TP> in the retest sessions are not significantly different from those in the first test sessions ($p > .05$). For a homogeneous, non-clinical population, the average throughput per subtest is not expected to change across testing sessions; in other words, the expectation for test reliability as performed on healthy and unvaried subjects is that the results of the retest administration should show no statistically significant change from the results of the first test administration. The result of the ANOVA is interpreted as a demonstration of consistency across test and retest sessions.

The tabulated MD for each subtest is based on the reported ICC coefficient for each test and battery, and the greater reported standard deviation between the two time points. In this context, the MD can be interpreted as the minimum amount of change in the population's score necessary to alert a statistically significant decrease or increase in test performance. The MD for DANA per subtest is approximately 17% of the <TP>, whereas the MD of both ANAM and ImPACT are approximately 24% and 27% of the <TP>, suggesting that, for the timescale reported here, DANA is a more sensitive measure to changes in test performance. In other words, a smaller proportional MD suggests that the measurement it describes is more sensitive to change and less confounded by error than a measurement with a proportionally larger MD.

The reliable change index (RCI) has been suggested as an alternative method to test–retest reliability, and several methods for determining the RCI have been published previously (e.g., Bruggemans, de Vijver, & Huysmans, 1997; Chelune, Naugle, Luders, Sedlack, & Awad, 1993; Jacobson & Truax, 1991; Moritz, Iverson, & Woodward, 2003; Temkin, Heaton, Grant, & Dikmen, 1999). Each of these methods distinctly relies upon the calculation of a reliability coefficient, arrived at by way of the ICC or by Pearson's r, or both. The aforementioned issues that arise from a non-standardized approach to arriving at the reliability coefficient serve to confound the interpretation of a RCI, specifically if more than one measure of reliability is calculated for any given test. For example, two measures of reliability are presented in Cole et al. (2013) and the standard error of the difference (S_{diff}, a precursor to the RCI) is tabulated using whichever measure is greater. When distinct disagreement exists between each measure within a given test, and/or when comparing tests that were measured using different reliability models, this approach is misleading. Under such circumstances and if the ICC is the calculated reliability measure, the ICC-dependent form of the SEM can provide an absolute bound on the reliability coefficient, or alternatively, the ICC-independent form of the SEM provides a bound on the measurement error. In either case, the resulting MD is found to be a consistent index for reliable change.

Conclusion

The DANA Rapid test battery was administered to cadets of the U.S. Air Force Academy football team before the commencement of the season and again during the season. The test–retest reliability of the DANA tool was evaluated and it was found that the test is consistent between test and retest sessions administered within approximately 77 days. Owing to the non-standardized practices of measuring test–retest reliability, both reporting the ICC along with a

precision metric and utilizing alternative methods to arrive at a RCI are suggested, such as with the MD approach, which is described in the ''SEM and MD'' section.

Acknowledgments

The authors would like to thank the collaborators Dr. Gerald McGinty, Dr. C. Dain Allred, Dr. Darren E. Campbell, Capt. Jack Tsao, Dr. James Spira, and Dr. Joseph Bleiberg for their intellectual contributions to this work, and the Technical Engineer James Drane for his effort in collecting these data.

Declaration of Conflicting Interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: AnthroTronix Inc. is the developer of the Defense Automated Neurobehavioral Assessment (DANA) tool.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was funded by AnthroTronix Inc. and the U.S. Navy Bureau of Medicine and Surgery (BUMED).

References

Baumgartner, T. (1969). Estimating reliability when all test trials are administered on the same day. *Research Quarterly for Exercise and Sport*, *40*, 222-225.

Baumgartner, T. (2000). Estimating the stability reliability of a score. *Measurement in Physical Education and Exercise Science*, *4*, 175-178.

Bedard, M., Martin, N., Krueger, P., & Brazil, K. (2000). Assessing reproducibility of data obtained with instruments based on continuous measurements. *Experimental Aging Research*, *26*, 353-365.

Broglio, S., Ferrara, M., Macciocchi, S., Baumgartner, T., & Elliot, R. (2007). Test-retest reliability of computerized concussion assessment programs. *Journal of Athletic Training*, *42*, 509-514.

Bruggemans, E., de Vijver, F. V., & Huysmans, H. (1997). Assessment of cognitive deterioration in individual patients following cardiac surgery: Correcting for measurement error and practice effects. *Journal of Clinical and Experimental Neuropsychology*, *19*, 543-559.

Chelune, G., Naugle, R., Luders, H., Sedlack, J., & Awad, I. (1993). Individual change after epilepsy surgery: Practice effects and base rate information. *Neuropsychology*, *7*, 41-52.

Christie, A., Kamen, G., Boucher, J., Inglis, J., & Gabriel, D. (2010). A comparison of statistical models for calculating reliability of the Hoffmann Reflex. *Measurement in Physical Education and Exercise Science*, *14*, 164-175.

Cole, W. R., Arrieux, J. P., Schwab, K., Ivins, B., Qashu, F. M., & Lewis, S. (2013). Test-retest reliability of four computerized neurocognitive assessment told in an active duty military population. *Archives of Clinical Neuropsychology*, *28*, 732-742.

Denegar, C. R., & Ball, D. W. (1993). Assessing reliability and precision measurement: An introduction to intraclass correlation and standard error of measurement. *Journal of Sport Rehabilitation*, *2*, 35-42.

Eliasziw, M., Young, S., Woodbury, M., & Fryday-Field, K. (1994). Statistical methodology for the concurrent assessment of interrater reliability: Using goniometric measurements as an example. *Physical Therapy*, *74*, 777-788.

Feldt, L., & McKee, M. (1958). Estimation of the reliability of skill tests. *Research Quarterly for Exercise and Sport*, *29*, 279-293.

Hopkins, W. (2000). Measures of reliability in sports medicine and science. *Sports Medicine*, *30*, 375-381.

Jacobson, N., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *59*, 12-19.

Kroll, W. (1962). A note on the coefficient of intraclass correlation as an estimate of reliability. *Psychological Bulletin*, *33*, 313-316.

Lathan, C., Spira, J., Bleiberg, J., Vice, J., & Tsao, J. (2013). Defense Automated Neurobehavioral Assessment (DANA)—Psychometric properties of a New Field-Deployable Neurocognitive Assessment Tool. *Military Medicine*, *178*, 365-372.

Learmonth, Y., Dlugonski, D., Pilutti, L., Sandroff, B., & Motl, R. (2013). The reliability, precision and clinically meaningful change of walking assessments in multiple sclerosis. *Multiple Sclerosis Journal*, *19*, 1784-1791.

Ludbrook, J. (2002). Statistical techniques for comparing measures and methods of measurement: A critical review. *Clinical and Experimental Pharmacology and Physiology*, *29*, 527-536.

McGraw, K., & Wong, S. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*, 30-46.

Moritz, S., Iverson, G., & Woodward, T. (2003). Reliable change indexes for memory performance in schizophrenia as a means to determine drug-induced cognitive decline. *Applied Neuropsychology*, *10*, 115-120.

Resch, J., Driscoll, A., McCaffray, N., Brown, C., Ferrara, M., Macciocchi, S., & . . . Walpert, K. (2013). ImPact Test-Retest Reliability: Reliably unreliable? *Journal of Athletic Training*, *48*, 506-511.

Safrit, M. (1976). *Reliability theory*. Washington, DC: American Alliance for Health, Physical Education, and Recreation.

Salarian, A. (2008). *ICC & ANOVA*. Retrieved from http://www.mathworks.com/matlabcentral/fileexchange/22099-intraclass-correlation-coefficient-icc/content/anova_rm.m

Segalowitz, S., Mahaney, P., Santesso, D., MacGrigor, L., Dwyan, J., & Willer, B. (2007). Retest reliability in adolescents of a computerized neuropsychological battery used to assess recovery from concussion. *NeuroRehabilitation*, *22*, 243-251.

Shrout, P. (1998). Measurement reliability and agreement in psychiatry. *Statistical Methods in Medical Research*, *7*, 301-317.

Shrout, P., & Fleiss, J. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *36*, 420-428.

Stratford, P., & Goldsmith, C. (1997). Use of standard error as a reliable index of interest: An applied example using elbow flexor strength data. *Physical Therapy*, *77*, 745-750.

Streiner, D., & Norman, G. (1995). *Measurement scales: A practical guide to their development and use* (2nd ed.). Oxford, UK: Oxford University Press.

Temkin, N., Heaton, R., Grant, I., & Dikmen, S. (1999). Detecting significant change in neuropsychological test performance: A comparison of four models. *Journal of the International Neuropsychological Society*, *5*, 357-369.

Weir, J. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *Journal of Strength and Conditioning Research*, *19*, 231-240.