Test-retest reliability on the Cambridge Neuropsychological Test Automated Battery: Comment on Karlsen et al. (2020)

Caroline Skirrow, Nathan Cashdollara, Kiri Grangera, and Francesca Cormackb,d

Cambridge Cognition, Cambridge, UK; School of Psychological Science, University of Bristol, Bristol, UK; Cambridge Cognition, Cambridge, MA, USA; Department of Psychiatry, University of Cambridge, Cambridge, UK

ABSTRACT
Test-retest reliability is essential to the development and validation of psychometric tools. Here we respond to the article by Karlsen et al. (Applied Neuropsychology: Adult, 2020), reporting test-retest reliability on the Cambridge Neuropsychological Test Automated Battery (CANTAB), with results that are in keeping with prior research on CANTAB and the broader cognitive assessment literature. However, after adopting a high threshold for adequate test-retest reliability, the authors report inadequate reliability for many measures. In this commentary we provide examples of stable, trait-like constructs which we would expect to remain highly consistent across longer time periods, and contrast these with measures which show acute within-subject change in response to contextual or psychological factors. Measures characterized by greater true within-subject variability typically have lower test-retest reliability, requiring adequate powering in research examining group differences and longitudinal change. However, these measures remain sensitive to important clinical and functional outcomes. Setting arbitrarily elevated test-retest reliability thresholds for test adoption in cognitive research limits the pool of available tools and precludes the adoption of many well-established tests showing consistent contextual, diagnostic, and treatment sensitivity. Overall, test-retest reliability must be balanced with other theoretical and practical considerations in study design, including test relevance and sensitivity.

Comment
Karlsen et al. (2020) recently published a paper in the journal "Applied Neuropsychology: Adult" on the test-retest reliability of the Cambridge Neuropsychological Test Automated Battery (CANTAB) in 75 healthy individuals assessed twice over a period of three months. The authors define outcome measures achieving acceptable test-retest reliability as those meeting correlation coefficients between testing occasions of at least (Pearson’s) $r = .75$. Karlsen et al. (2020) report that three of fourteen CANTAB outcome measures reach this threshold (21%) and describe inadequate reliability for the remainder of outcome measures (Pearson’s $r$ range .39–.73). In the current response to this article we would like to introduce a broader debate around the interpretation of these findings and discuss the application of reliability measurements in cognitive assessments for clinical research.

The test-retest reliability coefficients documented by Karlsen et al. (2020) are broadly in keeping with reliability levels previously reported for CANTAB (Barnett et al., 2010; Cacciamani et al., 2018; Feinkohl et al., 2020; Fowler et al., 1995; Gonçalves et al., 2016; Lowe & Rabbitt, 1998). As noted by Karlsen et al. (2020) themselves, test-retest reliability below $r = .75$ is not unique to CANTAB but is commonly reported for many traditional neuropsychological measures. This includes well-established traditional neuropsychological tests across a range of cognitive domains, including planning, inhibition and memory (Calamia et al., 2013; Köstering et al., 2015; Soveri et al., 2018), as well as in cognitive tests assessed using other computerized cognitive test batteries (Cole et al., 2013). In a meta-analysis by Calamia et al. (2013), the average test-retest reliability of many common cognitive neuropsychological tests on immediate retest was estimated at $r = .71$, and only around a third of test outcomes managed to reach thresholds of $r = .75$ and above.

Test-retest reliability is essential to the development of psychometric tools and refers to the reproducibility of two or more measurements using the same tool for the same person under the same conditions, where we would not expect the individual to have changed on the outcome (Aldridge et al., 2017). Good test-retest reliability can give
us confidence that the tool accurately represents the individual characteristics of a person at given time, which can allow us to use tools to support diagnostics, differentiation of participant cohorts, and identify genuine change (Aldridge et al., 2017). Given the greater difficulty with reliably detecting thresholds of impairment and change over time with assessments that have lower test-retest reliability, and the need for larger samples to detect group differences for such measures (Charter & Feldt, 2001; Chelune et al., 1993; Soveri et al., 2018), scientists are reasonably inclined to seek out assessments with a high degree of test-retest reliability for their research.

Thresholds for adequate reliability diagnostics vary from textbook to textbook, and paper to paper. The threshold adopted by Karlsen et al. (2020) is one of a range of short-hand rules of thumb (also see Cicchetti, 1994; Hopkins et al., 1990), which each provide differing thresholds for defining acceptable reliability. These are often provided as a matter of opinion without further justification or qualification, and with little or no theoretical basis (Charter & Feldt, 2001). However, when specifying the minimally acceptable threshold for test-retest correlations, it is important to consider the underlying properties of the construct that a given assessment measures.

Henry (1959) describes three sources of variance within reliability coefficients: (1) “true” individual difference variance, (2) variation within the individual as well as variation in response to the test situation, which ordinarily cannot be separated, and (3) experimental error. Although reliability coefficients are commonly interpreted as the degree to which a test is free from measurement error, they also incorporate true within-person variability. This is reflected in the decline in test-retest reliability with longer duration between tests (Calamia et al., 2013), likely reflecting an increase in true change over time.

With stable trait-like constructs, test-retest assessments using appropriately sensitive measures are likely to yield higher test-retest reliabilities, since within-individual variance is minimized. Taking an example from the physical health domain, a person's Body Mass Index (BMI) is modifiable in the longer term through exercise, dietary changes and growth, but with accurate measurement is likely to show only modest change over time. As a result, BMI typically shows high test-retest reliability on repeated short-term and longer-term assessment (Intra-class correlations of .95-.96 (Brisson et al., 2018; Leatherdale & Laxer, 2013)). By contrast, blood pressure varies acutely depending on short-term changes in exercise, stress and anxiety, caffeine, nicotine and alcohol consumption, and body position (standing, seated, lying down), as well as longer-term changes in exercise habits, dietary changes and growth. As a result, test-retest reliability of blood pressure measures are typically much lower ($r = .41-.76$ (Schechter & Adler, 1988; Stergiou et al., 2002)). However, despite this variability in measurements, blood pressure remains an integral part of routine clinical assessments for gauging patient health, and is used as an outcome measure in research and clinical trials, where increased test-retest variability is typically overcome by boosting sample size (Golomb et al., 2008).

Likewise, some cognitive measures are more resistant to variation over time and are more likely to meet high thresholds of test-retest reliability. In the domain of task-assessed cognition, measures of crystallized intelligence (e.g. vocabulary, semantic knowledge) tend to be more stable. In a meta-analysis of cognitive assessments the highest test-retest correlations were obtained for tests of vocabulary and information on the Wechsler Adult Intelligence Scale ($r = .90-.91$), both tests of semantic knowledge (Calamia et al., 2013).

However for more fluid cognitive domains, such as memory, attention, and processing and response speed, research suggests that cognitive function can vary acutely and systematically in relation to a variety of within-subject changes in response to contextual factors. For example, CANTAB tests tapping into these cognitive domains have been shown to be sensitive to the consumption of caffeinated drinks (Durlach, 1998), sleep-wake cycles (Oosterman et al., 2009), ambient temperature and humidity (Trezza et al., 2015) and acute anxiety induction (Savulich et al., 2019). Whilst these effects provide evidence for the sensitivity of these measures to changes in internal, external and psychological factors, they also represent a challenge when examining test-retest reliability. In test-retest studies contextual control is therefore important (for example, by completing repeated testing at similar times in the day, or maintaining consistency across test occasions in the consumption of stimulating substances (caffeine, nicotine) prior to research participation (Barnett et al., 2010)).

Any psychometric tool should aim to capture the construct of interest with the greatest fidelity, and as little measurement error as possible. However, for many cognitive constructs we may need to be realistic and accept that heightened sensitivity to within-subject factors is likely to increase measurement variability, which comes at a cost to test-retest reliability. As discussed by Guyatt et al. (1987), the usefulness of instruments to detect change in individuals over time not only relies on the test-retest reliability of these instruments, but also on their responsiveness, or their ability to detect differences and change. This distinction is of paramount importance when deploying cognitive assessments in clinical trials aimed to optimize the fidelity of a particular metric in order to be sensitive enough to detect performance changes due to therapeutic intervention.

Test-retest reliability matters since it affects the power of a clinical trial to detect a significant treatment effect, and in group comparison studies to detect significant group differences. However, this can be overcome by designing studies with larger samples that absorb increased outcome variability, whilst maintaining sensitivity to group differences or longitudinal change. Test–retest reliability is important, but it is not everything (Barnett et al., 2010), and should be balanced with other theoretical and practical considerations in study design, including test relevance and sensitivity. Just as for measures of physical health, setting arbitrarily elevated test-retest reliability thresholds for test adoption in cognitive research will limit the pool of available tools, and preclude
the adoption of many which are well established and show consistent contextual, diagnostic, and treatment sensitivity.

**Disclosure statement**

All authors are employed by Cambridge Cognition, a neuroscience technology company, with products which include the cognitive test battery CANTAB.

**ORCID**

Caroline Skirrow [http://orcid.org/0000-0001-8692-7787]
Nathan Cashdollar [http://orcid.org/0000-0003-3847-7261]
Kiri Granger [http://orcid.org/0000-0003-3749-3323]
Sally Jennings [http://orcid.org/0000-0002-9863-2623]
Elizabeth Baker [http://orcid.org/0000-0001-8008-7035]
Jennifer Barnett [http://orcid.org/0000-0002-4851-5949]
Francesca Cormack [http://orcid.org/0000-0002-4413-177X]

**References**

Aldridge, V. K., Dovey, T. M., & Wade, A. (2017). Assessing test-retest reliability of psychological measures: Persistent methodological problems. *European Psychologist, 22*(4), 207–218. https://doi.org/10.1027/1016-9040.000298

Barnett, J. H., Robbins, T. W., Leeson, V. C., Sahakian, B. J., Joyce, E. M., & Blackwell, A. D. (2010). Assessing cognitive function in clinical trials of schizophrenia. *Neuroscience and Biobehavioral Reviews, 34*(8), 1161–1177. https://doi.org/10.1016/j.neubiorev.2010.01.012

Brisson, N. M., Stratford, P. W., & Maly, M. R. (2018). Relative and absolute test-retest reliability of biomechanical risk factors for knee osteoarthritis progression: Benchmarks for meaningful change. *Osteoarthritis and Cartilage*, 26(2), 220–226. https://doi.org/10.1016/j.joca.2017.11.003

Cacciamani, F., Salvadori, N., Eusie, P., Lisetti, V., Luchetti, E., Calabresi, P., & Parnetti, L. (2018). Evidence of practice effect in CANTAB spatial working memory test in a cohort of patients with mild cognitive impairment. *Applied Neuropsychology: Adult, 25*(3), 237–248. https://doi.org/10.1080/23279095.2017.1286346

Calamia, M., Markon, K., & Tranel, D. (2013). The robust reliability of neuropsychological measures: Meta-analyses of test-retest correlations. *The Clinical Neuropsychologist, 27*(7), 1077–1105. https://doi.org/10.1080/13854066.2013.809795

Charter, R. A., & Feldt, L. S. (2001). Meaning of reliability in terms of individual change after epilepsy surgery: Practice effects and base-rate information. *Neuropsychology, 7*(1), 41–52. https://doi.org/10.1037/0894-4105.7.1.41

Chelune, G. J., Naugle, R. L., Lüders, H., Sedlak, J., & Awad, I. A. (1993). Individual change after epilepsy surgery: Practice effects and base-rate information. *Neuropsychology, 7*(1), 41–52. https://doi.org/10.1037/0894-4105.7.1.41

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normalized and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284–290. https://doi.org/10.1037/1040-1594.6.4.284

Cole, W. R., Arieux, J. P., Schwab, K., Ivins, B. J., Qashu, F. M., & Lewis, S. C. (2013). Test-retest reliability of four computerized neuropsychological assessment tools in an active duty military population. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists, 28*(7), 732–742. https://doi.org/10.1093/arclin/act040

Durlach, P. J. (1998). The effects of a low dose of caffeine on cognitive performance. *Psychopharmacology, 140*(1), 116–119. https://doi.org/10.1007/s002130050746

Feinkohl, I., Borchers, F., Burkhardt, S., Krampe, H., Kraft, A., Speidel, S., Kant, I. M. J., van Montfort, S. J. T., Aarts, E., Kruppa, J., Slooter, A., Winterer, G., Pischon, T., & Spies, C. (2020). Stability of neuropsychological test performance in older adults serving as normative controls for a study on postoperative cognitive dysfunction. *BMC Research Notes, 13*(1), 3–8. https://doi.org/10.1186/s13104-020-4919-3

Fowler, K. S., Saling, M. M., Conway, E. L., Semple, J. M., & Louis, W. J. (1995). Computerized delayed matching to sample and paired associate performance in the early detection of dementia. *Applied Neuropsychology, 2*(2), 72–78. https://doi.org/10.1207/s13324836anp02_2

Golomb, B. A., Dimsdale, J. E., White, H. L., Ritchie, J. B., & Criqui, M. H. (2008). Reduction in blood pressure with statins. *Archives of Internal Medicine, 168*(7), 721–727. https://doi.org/10.1001/archinte.168.7.721

Gotenval, M. M., Pinho, M. S., & Simões, M. R. (2016). Test-retest reliability analysis of the Cambridge Neuropsychological Automated Tests for the assessment of dementia in older people living in retirement homes. *Applied Neuropsychology: Adult, 23*(4), 251–263. https://doi.org/10.1080/23279095.2015.1053889

Guyatt, G., Walter, S., & Norman, G. (1987). Measuring change over time: Assessing the usefulness of evaluative instruments. *Journal of Chronic Diseases, 40*(2), 171–178. https://doi.org/10.1016/0021-9681(87)90069-5

Henry, F. M. (1959). Reliability, measurement error, and intra-individual difference. *Research Quarterly of the American Association for Health, Physical Education and Recreation, 30*(1), 21–24. https://doi.org/10.1080/10671188.1959.1061303

Hopkins, K. D., Stanley, J. C., & Hopkins, B. R. (1990). *Educational and Psychological measurement and evaluation* (7th ed.). Prentice-Hall.

Karlsen, R. H., Karr, J. E., Sakvik, S. B., Lundingvold, A. J., Hjemdal, O., Olsen, A., Iversen G. L., & Skandsen, T. (2020). Examining 3-month test-retest reliability and reliable change using the Cambridge Neuropsychological Test Automated Battery. *Applied Neuropsychology: Adult*. https://doi.org/10.1080/23279095.2020.1722126

Köstering, L., Nitschke, K., Schumacher, F. K., Weiller, C., & Kaller, C. P. (2015). Test-retest reliability of the Tower of London planning task (TOL-F). *Psychological Assessment, 27*(3), 925–934. https://doi.org/10.1037/pas0000097

Leatherdale, S. T., & Laxer, R. E. (2013). Reliability and validity of the weight status and dietary intake measures in the COMPASS questionnaire: Are the self-reported measures of body mass index (BMI) and Canada’s food guide servings robust? *International Journal of Behavioral Nutrition and Physical Activity, 10*(1), 42. https://doi.org/10.1186/1479-5868-10-42

Lowe, C., & Rabitt, P. (1998). Test/re-test reliability of the CANTAB and ISPOCD neuropsychological batteries: Theoretical and practical issues. Cambridge Neuropsychological Test Automated Battery. *International Study of Post-Operative Cognitive Dysfunction. Neuropsychologia, 36*(9), 915–923. https://doi.org/10.1016/S0028-3932(98)00085-0

Oosterman, J. M., Van Someren, E. J. W., Vogels, R. L. C., Van Harten, B., & Sderer, E. J. A. (2009). Fragmentation of the rest-activity rhythm correlates with age-related cognitive deficits. *Journal of Sleep Research, 18*(1), 129–135. https://doi.org/10.1111/j.1365-2869.2008.00704.x

Savulich, G., Hezeman, F. H., van Ghesel Grothe, S., Dafflon, J., Schulen, B., Brühl, A., Sahakian, B. J., & Robbins, T. W. (2019). Acute anxiety and autonomic arousal induced by CO2 inhalation impairs prefrontal executive functions in healthy humans. *Translational Psychiatry, 9*(1), 296. https://doi.org/10.1038/s41398-019-0634-z

Schechter, C. B., & Adler, R. S. (1988). Bayesian analysis of diastolic blood pressure measurement. *Medical Decision Making: An International Journal of the Society for Medical Decision Making, 8*(3), 182–190. https://doi.org/10.1177/0272989X8800800306

Soveri, A., Lehtonen, M., Karlsson, L. C., Lukasik, K., Antfolk, J., & Laine, M. (2018). Test-retest reliability of five frequently used executive tasks in healthy adults. *Applied Neuropsychology: Adult*. https://doi.org/10.1080/3932(98)00086-0
Stergiou, G. S., Baibas, N. M., Gantzarou, A. P., Skeva, I. I., Kalkana, C. B., Roussias, L. G., & Mountokalakis, T. D. (2002). Reproducibility of home, ambulatory, and clinic blood pressure: Implications for the design of trials for the assessment of antihypertensive drug efficacy. *American Journal of Hypertension, 15*(2 Pt 1), 101–104. https://doi.org/10.1016/S0895-7061(01)02324-X

Trezza, B. M., Apolinario, D., De Oliveira, R. S., Busae, A. L., Luiz, F., Gonçalves, T., Hilário P., Saldiva N., & Jacob-Filho, W. (2015). Environmental heat exposure and cognitive performance in older adults: A controlled trial. *AGE, 37*(3), 9783. https://doi.org/10.1007/s11357-015-9783-z