Received: 20 September 2022 | Revised: 13 January 2023 | Accepted: 19 January 2023

DOI: 10.1002/dad2.12406

Alzheimer's & Dementia
Diagnosis, Assessment & Disease Monitoring

RESEARCH ARTICLE

Reliability of the NACC Telephone-administered Neuropsychological Battery (T-cog) and Montreal Cognitive Assessment for participants in the USC ADRC

Rebecca Sitra Howard[1] | **Terry E. Goldberg**[2] | **James Luo**[1] | **Cynthia Munoz**[1] | **Lon S. Schneider**[1]

[1]Keck School of Medicine of USC, Los Angeles, California, USA

[2]Department of Psychiatry, Columbia University Medical Center, New York, New York, USA

Correspondence
Lon S. Schneider, Keck School of Medicine of USC, Los Angeles CA 90089, USA.
Email: lschneid@usc.edu

Funding information
The Della Martin Foundation; Alzheimer's Disease Research Center, Grant/Award Number: P30 AG066530; Novel Cognitive and Functional Measure for Alzheimer's Disease, Grant/Award Number: R01 AG051346

Abstract

Introduction: Restrictions during the COVID-19 pandemic necessitated remote administration of neuropsychological testing. We assessed the test–retest reliability for a telephone-administered cognitive battery, recommended for use in the National Institute on Aging Alzheimer's Disease Research Center (ADRC).

Methods: 64 participants in the University of Southern California ADRC clinical core underwent repeat telephone evaluation using the T-cog Neuropsychological Battery. Reliability was measured by intraclass correlation coefficient (ICC) for continuous variables and weighted Kappa coefficient for categorical variables. Mean scores for Montreal Cognitive Assessment (MoCA) total and Craft Story 21 Immediate and Delayed Recall were compared using paired *t* tests.

Results: Mean age was 74.8 (8.3 standard deviation); 73.4% were female. ICCs ranged from 0.52 to 0.84, indicating moderate test–retest reliability except for number span backward, which showed poor reliability. Weighted Kappa for MoCA items ranged from −0.016 to 0.734; however, relatively good observed agreement was seen across all items (70.3% to 98.4%). Although MoCA total scores did not significantly change, Craft Story 21 Immediate and Delayed Recall mean scores increased between first and second administrations ($P < 0.0001$).

Discussion: Test–retest reliability for the T-cog Neuropsychological Battery is adequate. The variation seen in testing is similar to results seen from face-to-face testing, with Craft Story 21 recall showing modest and expected practice effects.

KEYWORDS
Alzheimer's disease, Alzheimer's Disease Research Center, category fluency, cognitive impairment, craft story, Montreal Cognitive Assessment, neuropsychological testing, number span, story recall, test–retest reliability, verbal fluency

Highlights

- Moderate test–retest reliability is seen in most measures of the National Alzheimer's Coordinating Center Neuropsychological Test Battery and the Montreal Cognitive Assessment (MoCA).
- Intraclass correlation coefficients ranged from 0.52 to 0.84, except for number Span backward.
- Weighted Kappa for MoCA items varied, but good observed agreement was seen.
- MoCA total mean score did not change significantly between administrations.
- Craft Story 21 Immediate and Delayed Recall means increased on repeat testing ($P < 0.0001$).

1 | INTRODUCTION

Research in Alzheimer's disease and other dementias is dependent on accurate and reliable measurement of neuropsychological performance over a long disease course. Restrictions introduced in the wake of the COVID-19 pandemic, however, presented an unprecedented challenge for delivering these tests. One research lesson learned was the need for remote administration of a variety of rating scales and cognitive instruments. Telephone administration is a feasible form of remote administration, with the American Psychological Association, and other organizations around the world, publishing guidelines[1] to move toward tele-assessment while restrictions limited face-to-face testing. In the context of older participant samples, telephone administration is also attractive, given their broad access to phones in addition to their, sometimes-compromised, ability to use technology for video conferences or smartphone test administration.[2]

However, it should not be assumed once appropriate tests are selected, given the nature of telephone communication, that the reliability of telephone test administration follows reliability in face-to-face administration. Environmental variables, usually controlled for in a laboratory setting, for example distractors, hearing impairments due to poor signal and the absence of lip reading, play a significant role in comprehension of everyday face-to-face conversation[3] and may negatively impact participant scoring. Alternatively, without tester supervision, the participant may also rely on aids such as clocks, notepads, or other persons, improving their score. With this in mind, we elected to conduct a study of test–retest reliability for Uniform Data Set (UDS)-recommended telephone measures.

2 | METHODS

2.1 | Participants

Patients, ranging from cognitively unimpaired to diagnosed dementia, from the University of Southern California Alzheimer's Disease Research Center (ADRC) clinical core were seen for yearly evaluation. Inclusion criteria for this study were age > 70 and having been previously registered and followed in the clinical core. Eligibility for the clinical core includes having a vascular or metabolic risk factor for cognitive impairment or dementia, willingness to donate one's brain upon death, or participation in ADRC-affiliated studies. In recent years, these eligibility considerations skew clinical core participants toward the cognitively unimpaired to early mild cognitive impairment (MCI) range, tending to not include participants with dementia.

We planned to include 60, English-speaking participants who underwent telephone-based neuropsychological evaluation, performed and obtained using the UDS.[4] Consent was obtained prior to participants' annual, remote visit and the T-cog Neuropsychological Battery. The battery was intended to be repeated approximately 2 weeks after the initial test. One version of the telephone UDS (T-UDS) was used for both administrations. Two neuropsychological technicians administered the T-cog, and the same tester administered the first and repeated test to each participant.

2.2 | T-cog neuropsychological battery

The modified telephone version of the UDSv3 Neuropsychological Test Battery[4–6] is intended to support continued data collection considering the limitations on in-person study visits during the COVID-19 pandemic. The battery includes the Blind/Telephone Montreal Cognitive Assessment (MoCA[7]), Craft Story 21 Recall–Immediate,[5-8] Number Span Test: Forward and Backward,[5] Craft Story 21 Recall–Delayed,[5-8] and Category and Verbal Fluency.[5]

2.2.1 | Blind/Telephone Montreal Cognitive Assessment (MoCA[7])

The MoCA is a face-to-face screening instrument composed of eight sections including visuospatial/executive, naming, memory, attention, language, abstraction, delayed recall, and orientation, generating a summative score out of 30. In the Blind/Telephone version, naming, cube drawing, and clock drawing are omitted allowing for a maximum score of 22.

2.2.2 | Craft Story 21 Recall–Immediate[5-8]

The Craft Story 21 Recall (Immediate) assesses the participant's ability to recollect a short story. The participant is asked to retell the story from memory immediately after hearing it. The participant is asked to repeat the story after a 20-minute delay and is also cued to remember the story for later. The participant is scored separately on both verbatim (maximum score 44) and paraphrase (maximum score 25) recollection. Instructions for face-to-face and remote administration do not differ.

2.2.3 | Number Span Test: Forward and Backward[5]

The Number Span Test is a test of working memory, in which participants are asked to repeat sequences of numbers exactly, immediately after hearing them, with sequences presented in ascending order of difficulty. Participants are given two trials of each sequence length and continue until failing two consecutive trials of the same length. In the first instance, participants must remember the number sequences in the order presented to them (Number Span Forward) and then must recall different number sequences in reverse order in the second set of trials (Number Span Backward). Both sets of trials are scored by (1) total number of correct trials and (2) longest sequence repeated correctly. Face-to-face and remote administration instructions are the same.

2.2.4 | Craft Story 21 Recall–Delayed[5-8]

The Craft Story 21 Recall (Delayed) assesses the delayed recall (episodic memory) of the story that was read to the participant during the immediate recall earlier in the testing session. It is administered 20 minutes after the immediate recall. As with immediate recall, the participant is scored on both verbatim and paraphrased response, with the same number of maximum points available for both. Remote administration instructions do not differ from face-to-face.

2.2.5 | Category Fluency[5]

Category Fluency measures the participant's semantic memory. The participant is given 60 seconds to name as many distinct items of a given category (animals or vegetables) as they can and is then scored by the number of unique responses generated. Face-to-face and remote test administration is the same.

2.2.6 | Verbal Fluency[5]

Verbal Fluency measures the participant's speeded word retrieval to phonemic cues. The participant is given 60 seconds to name as many distinct items that begin with a certain letter of the alphabet (F and L) as they can and is then scored by the number of unique responses generated.

RESEARCH IN CONTEXT

1. **Systematic Review**: The test–retest reliability of the cognitive tests that comprise the National Alzheimer's Coordinating Center (NACC) Neuropsychological Test Battery show moderate–good reliability in face-to-face administration. However, there are limited published studies on test–retest reliability of remote methods.
2. **Interpretation**: Moderate test–retest reliability is seen in most measures of the NACC Telephone-administered Neuropsychological Battery, Craft Story 21 Immediate and Delayed Recall, Number Span Forward, and Category and Verbal Fluency; and on a telephone version of the Montreal Cognitive Assessment (MoCA). This is relatively equivalent to face-to-face testing, and suggests adequate reliability for this format.
3. **Future Directions**: Investigations of this cognitive battery in diverse and more cognitively impaired populations are required to improve generalizability and reliability.

2.3 | Statistical analysis

Descriptive statistics were calculated for the participants' demographics. Mean and standard deviation (SD) were reported for the continuous variables, and frequency counts and percentages were reported for categorical variables.

Reliability of the T-cog Neuropsychological Battery was assessed by examining the extent of agreement between the initial and repeated T-cog scores. For continuous variables, test–retest reliability of the repeated administration was measured by two-way random effects, absolute agreement, and single rater/measurement intraclass correlation coefficient (ICC). Repeated responses within a participant were considered longitudinal data, and mean difference and SD were reported. An ICC <0.5 indicates poor reliability, an ICC between 0.5 and 0.75 indicates moderate reliability, an ICC between 0.75 and 0.9 indicates good reliability, and ICC >0.9 indicates excellent reliability.[9]

For categorical variables, the percentage of observed agreement and the weighted Kappa coefficient were calculated, using weights $1 - |i - j|/(k - 1)$, where i and j index the rows and columns of the ratings by the two repeated measures and k is the maximum number of possible measurements. Weighted Kappa results are interpreted as follows: values ≤0 as no agreement, 0.01 to 0.20 as none to slight, 0.21 to 0.40 as fair, 0.41 to 0.60 as moderate, 0.61 to 0.80 as substantial, and 0.81 to 1.00 as almost perfect agreement.[10]

Paired t tests were conducted to compare if there was a significant difference between the MoCA total raw scores, the Craft Story

TABLE 1 Paired *t* tests for the blind MoCA and Craft Story.

Variable		Mean	Standard deviation	Standard error	95% CI	Min	Max
MoCA total score	1st T-Cog	18.53	2.71	0.34	(17.85, 19.21)	11	22
	2nd T-Cog	18.95	2.79	0.35	(18.26, 19.65)	11	22
	Difference	0.42	1.94	0.24	(−0.06, 0.91)	−5	5
	$t = 1.74$, df = 63, $P = 0.087$						
Craft Story 21 Immediate: Total story units recalled, verbatim scoring	1st T-Cog	19.73	23.58	0.85	(18.03, 21.44)	3	31
	2nd T-Cog	23.58	8.07	1.01	(21.56, 25.60)	2	37
	Difference	3.84	5.23	0.65	(2.54, 5.15)	−13	18
	$t = 5.88$, df = 63, $P < 0.0001$						
Craft Story 21 Immediate: Total story units recalled, paraphrase scoring	1st T-Cog	14.77	4.50	0.56	(13.64, 15.89)	3	22
	2nd T-Cog	17.09	4.98	0.62	(15.85, 18.34)	2	25
	Difference	2.33	2.92	0.36	(1.60, 3.06)	−3	10
	$t = 6.38$, df = 63, $P < 0.0001$						
Craft Story 21 Delayed: Total story units recalled, verbatim scoring	1st T-Cog	16.67	8.16	1.02	(14.63, 18.71)	0	34
	2nd T-Cog	20.40	9.18	1.16	(18.09, 22.71)	0	38
	Difference	3.76	4.94	0.62	(2.51, 5.01)	−7	16
	$t = 6.04$, df = 62, $P < 0.0001$						
Craft Story 21 Delayed: Total story units recalled, paraphrase scoring	1st T-Cog	13.05	5.95	0.74	(11.56, 14.53)	0	22
	2nd T-Cog	15.44	6.02	0.76	(13.93, 16.96)	0	25
	Difference	2.43	2.99	0.38	(1.67, 3.18)	−5	9
	$t = 6.44$, df = 62, $P < 0.0001$						

Abbreviations: CI, confidence interval; MoCA, Montreal Cognitive Assessment; T-cog, Telephone-administered Neuropsychological Battery.

21 Recall–Immediate (verbatim and paraphrase) scores, and the Craft Story 21 Recall–Delayed (verbatim and paraphrase) scores from the initial and repeat T-cogs administered. A sample size of 60 was estimated to allow detection of 0.75-point mean change on the MoCA with 80% power and alpha = 0.05 assuming a SD of 2.0. Statistical analyses were conducted using Stata statistical software 15.1 (StataCorp LLC.) and SAS version 9.4 (SAS Institute Inc.).

3 | RESULTS

3.1 | Participants

We approached 73 ADRC participants who had been tested remotely to gain their agreement to be retested. Three did not respond, one declined, and five did not attend their scheduled telephone appointment. Sixty-four participants underwent repeated testing within 12 to 29 days between September 2020 to March 2021. All but three were retested within 12 to 22 days, who were retested from 27 to 29 days. The majority was female ($n = 47$, 73.4%) and White ($n = 54$, 84.4%), with 3 (4.7%) Black, and 7 (10.9%) Asian participants. Eight (12.5%) identified as Hispanic. Mean age at initial testing was 74.8 (8.3 SD) years, and mean years of education was 16.4 (2.2 SD) years. Most participants were diagnosed as not cognitively impaired ($n = 53$, 82.8%), with seven (10.9%) diagnosed as MCI, three (4.7%) as dementia due to AD, and one (1.6%) participant as cognitive impairment, not MCI, all based on previous yearly visit diagnoses.

3.2 | Test–retest reliability

Maximum total MoCA score achieved was 22 and minimum score was 11 at both time points, with the lowest scores attributable to the same participant (Table 1). Seven (10.9%) participants achieved a maximum score of 22 on first testing and 14 (21.9%) during second testing. The distribution of MoCA total scores for all participants by visit is shown in Figure 1A. Mean difference in MoCA total scores between first and second administration was 0.42 (1.94 SD, range −5 to +5) and is shown in Figure 1B. Paired *t* tests did not show a significant difference between MoCA total scores from first to second testing ($t = 1.74$, $P = 0.09$). A Bland–Altman plot showed no improvement in scores in relationship with severity of scores (Supplement 1 in supporting information).

Weighted Kappa calculation for MoCA items ranged greatly, from −0.016 (Orientation–City) to 0.734 (Orientation–Day). However, relatively good observed agreement was seen across all items (Table 2). The lowest observed agreement was 70.3% for Attention–Letter A. Attention–Digits and measures of orientation showed the highest observed agreement (90.6%–98.4%). Moderate reliability of the MoCA total score was observed with ICC = 0.745.

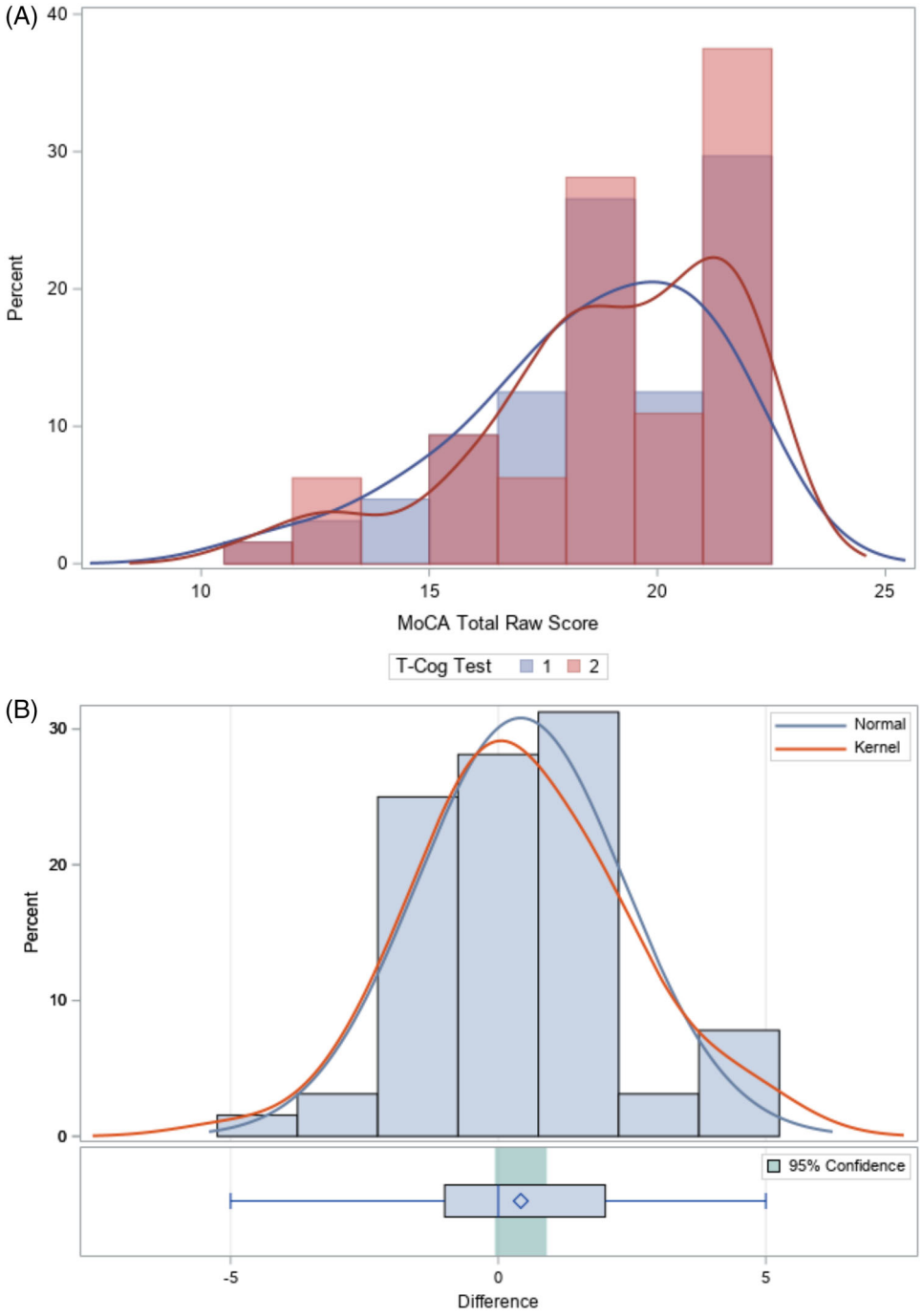

FIGURE 1 A, Distribution of MoCA total scores by first and second administrations. B, The difference in MoCA total scores between first and second administrations follows a relatively normal distribution. With a *t* test statistic = 1.74 and *P*-value = 0.09, the difference in MoCA total raw scores between initial and repeat T-cog is not significantly different from 0. MoCA, Montreal Cognitive Assessment; T-cog, Telephone-administered Neuropsychological Battery.

Craft Story 21 immediate verbatim mean changed by 3.84 points (SD 5.3) and paraphrase mean by 2.33 (SD 2.92;Table 1). The distributions of difference for initial and repeat testing results of Craft Story 21 immediate paraphrase recall are displayed in Figure 2A. Maximum scores achieved increased in both immediate verbatim and paraphrase scoring (31 to 37, and 22 to the maximum of 25 points, respectively) between first and second testing. Paired *t* tests performed for immediate verbatim and paraphrase recall showed significant differences in scores for both between initial and repeat test administrations ($t = 5.88$ verbatim, $t = 6.38$ paraphrase, $P < 0.0001$). ICC calculation for both immediate verbatim and paraphrase scores indicated moderate reliability (Table 2). For immediate paraphrase recall, improvement on retest was approximately 0.5 points better for participants with better immediate recall (Supplement 1).

TABLE 2 Test–retest reliability of the blind MoCA and NACC UDS telephone-administered items. Intraclass correlation coefficients or weighted Kappa coefficients were calculated for each question of the T-cog Neuropsychological Battery scores form.

Question	Weighted Kappa	Observed agreement (%)	ICC	Difference, mean (SD)	Interpretation
Montreal Cognitive Assessment (MoCA) Blind					
Total Score – uncorrected			0.745	0.42 (1.94)	Moderate reliability
Attention – Digits	0.244	93.0			Slight agreement
Attention – Letter A	0.003	70.3			Slight agreement
Attention – Serial 7s	0.541	88.3			Moderate agreement
Language – Repetition	0.362	80.5			Fair agreement
Language – Fluency	0.074	82.8			Slight agreement
Abstraction	0.355	84.4			Fair agreement
Delayed recall – No cue	0.455	79.4			Moderate agreement
Delayed recall – Category cue	0.037	74.2			Slight agreement
Delayed recall – Recognition	0.045	71.2			Slight agreement
Orientation – Date	0.241	90.6			Slight agreement
Orientation – Month	−0.024	93.8			No agreement
Orientation – Year	0.066	98.4			Substantial agreement
Orientation – Day	0.734	96.9			Substantial agreement
Orientation – Place	0.650	96.9			Substantial agreement
Orientation – City	−0.016	96.9			No agreement
Craft Story 21 Recall – Immediate					
Total story units recalled, verbatim scoring			0.669	3.84 (5.23)	Moderate reliability
Total story units recalled, paraphrase scoring			0.726	2.33 (2.92)	Moderate reliability
Number Span Test: Forward					
Number of correct trials			0.614	−0.90 (1.91)	Moderate reliability
Longest span forward			0.529	−0.13 (1.19)	Moderate reliability
Number Span Test: Backward					
Number of correct trials			0.471	−0.44 (2.22)	Poor reliability
Longest span backward			0.464	−0.23 (1.28)	Poor reliability
Craft Story 21 Recall (Delayed)					
Total story units recalled, verbatim scoring			0.769	3.76 (4.94)	Moderate reliability
Total story units recalled, paraphrase scoring			0.811	2.43 (2.99)	Good reliability
Delay time (min)			0.021	−0.61 (2.73)	Poor reliability
Cue ("boy") needed	0.704	95.3			Substantial agreement
Category Fluency					
Animals: Total number of animals named in 60 s			0.792	−0.05 (3.91)	Good reliability
Vegetables: Total number of vegetables named in 60 s			0.714	0.19 (3.06)	Moderate reliability
Verbal Fluency: Phonemic Test					
Number of correct F-words generated in 1 min			0.684	0.53 (3.98)	Moderate reliability
Number of F-words repeated in 1 min			0.825	−0.36 (1.15)	Good reliability
Number of non-F-words and rule violation errors in 1 min			−0.040	−0.08 (1.10)	Poor reliability
Number of correct L-words generated in 1 min			0.763	0.41 (3.04)	Good reliability
Number of L-words repeated in one min			0.562	−0.23 (1.46)	Moderate reliability
Number of non-L-words and rule violation errors in 1 min			0.017	−0.02 (0.72)	Poor reliability
TOTAL number of correct F-words and L-words			0.845	0.94 (4.84)	Good reliability
TOTAL number of F-word and L-word repetition errors			0.825	−0.61 (1.87)	Good reliability
TOTAL number of non-F/L words and rule violation errors			−0.017	−0.06 (1.36)	Poor reliability

(Continues)

TABLE 2 (Continued)

Question	Weighted Kappa	Observed agreement (%)	ICC	Difference, mean (SD)	Interpretation
Overall appraisal					
Per the clinician (e.g., neuropsychologist, behavioral neurologist, or other suitably qualified clinician), based on the UDS neuropsychological examination, the subject's cognitive status is deemed: Tester's assessment of participant cognitive status	0.543	87.5			Moderate agreement

Abbreviations: ICC, intraclass correlation coefficients; MoCA, Montreal Cognitive Assessment; NACC, National Alzheimer's Coordinating Center; T-cog, Telephone-administered Neuropsychological Battery; SD, standard deviation; UDS, Uniform Data Set.

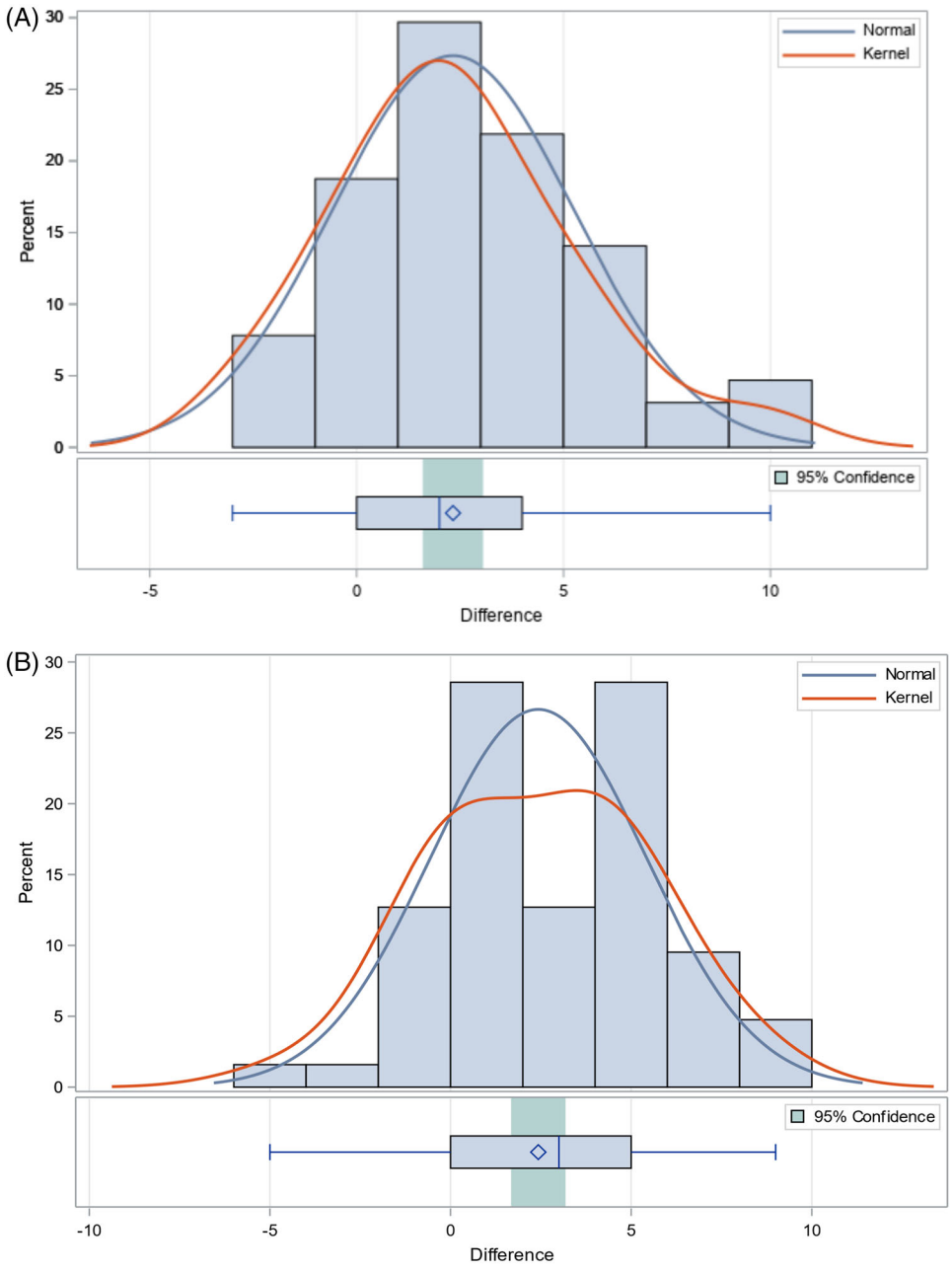

FIGURE 2 The difference in Craft Story 21 Immediate (A) and Delayed (B) total story units recalled, paraphrase scoring, between first and second administrations, follows relatively normal distributions with a t test = 6.38, $P < 0.0001$, the difference in Immediate total story units recalled, and a t test = 6.44, $P < 0.0001$, the difference in Delayed total story units recalled are both significantly different from 0.

Craft Story 21 delayed verbatim mean score changed by 3.76 (SD 4.94) and paraphrase by 2.43 (SD 2.99; Table 1). The distribution of difference for initial and repeat testing results of Craft Story 21 delayed paraphrase is displayed in Figure 2B. Maximum scores achieved for verbatim and paraphrase recall increased from first to second administration (34 to 38, and 22 to 25 points, respectively). Paired *t* tests also showed significant differences between Craft Story delayed scores from initial and repeat test administration ($t = 6.04$ verbatim, $t = 6.44$ paraphrase, $P < 0.0001$). ICC calculation showed good test–retest reliability, with the highest degrees of agreement in delayed paraphrase (ICC = 0.811). For delayed paraphrase recall, improvement on retest was unrelated to the participants' level of severity scores (Supplement 1).

Number Span Forward showed moderate reliability (Table 2); however, Number Span Backward showed poor reliability with ICCs <0.5 for both outcomes of "number of correct trials" (ICC = 0.471) and "longest span backward" (ICC = 0.464).

Category fluency showed moderate to good reliability in both animal and vegetable categories tested (ICC = 0.792 and 0.714, respectively). Tests of verbal fluency had the highest calculated ICCs, including number of F-words generated in 1 minute (ICC = 0.825) and total number of correct F- and L- words (ICC = 0.845), indicating good test–retest reliability.

4 | DISCUSSION

This study demonstrated that test–retest reliability for the overall score on the telephone-administered MoCA as well as subtests had moderate reliability. This compares well to results reported for in-person retesting with the full MoCA, ICC = 0.82 and 0.64 between different MoCA versions (7.1 vs. 7.2/7.3) reported previously.[11] Most other measures also demonstrated moderate to good reliability, suggesting that in remote testing of many verbally presented neurocognitive tests, reliability is adequate for this format. It is worth noting that while very high percentage observed agreement was found between initial and repeat testing of MoCA item scores, calculated weighted Kappa was very low. This is likely due to the high prevalence of one outcome on these item scores. For example, most participants correctly identified the city they were in during measures of orientation, with few giving incorrect responses, giving rise to "Cohen's Kappa Paradox."[12]

Moreover, the tests' reliabilities are relatively equivalent to those found in face-to-face laboratory or clinical administration. Importantly, Craft Story reliability was high even though the performance differences between tests were large, as has been shown previously with its predecessor test in the UDS, Logical Memory.[13] The statistically significant increases in scores for both immediate and delayed recall was most likely due to a practice or learning effect as the same story was repeated for both administrations. However, this also suggests that the magnitude of the practice effect was similar across participants, and they retained their rank order.

The relatively low reliability of Number Span (especially backward) is unsurprising. Of the Wechsler IQ subtests, Digit Span is among the subtests with the lowest reliability in face-to-face testing.[14] Thus, the results here are consistent with earlier observations of reliability for Digit Span Forward and Backward,[15] which was replaced by Number Span in the latest version of the UDS.[5] The low reliability suggests that there is variability in attentional or working memory systems that can produce lapses in performance, especially if the testing is not conducted in a controlled environment.

Telephone administration of limited neuropsychological tests allowed for continuation of research activities despite restrictions on clinic visits imposed during the COVID pandemic. Development and validation of video-, tablet-, and computer display monitor interfaces offer alternatives and advances for remote cognitive testing.

4.1 | Limitations/generalizations

The generalizability of test reliability may be limited by the lack of diversity within the participant sample, with male (26%) and Latino (12.5%) participants being under-represented compared to the overall clinical core sample (36% and 19%, respectively). Requirement to be English speaking, and ability and willingness to participate in the telephone assessments themselves most likely excluded under-represented groups and selected for relatively more educated participants. Predominance of cognitively normal subjects within the sample may also limit generalizability of test reliability to more impaired populations. As participants were recruited from the ADRC at-risk cohort, they had previous, in-person exposure to the full ADRC neuropsychological battery at yearly intervals. This familiarity may have enhanced the test–retest reliability of the telephone-based testing. The use of a short retest interval has been shown to increase reliability coefficients and increases practice effects.[16] As the same tester evaluated a participant on both occasions this can represent a potential bias, resulting in an artificially high agreement.

ACKNOWLEDGMENTS

The authors have nothing to report.

CONFLICTS OF INTEREST

Rebecca Sitra Howard, James Luo, Cynthia Munoz have no disclosures. Terry E. Goldberg reports royalties from VeraSci for use of the BACS cognitive screening instrument in clinical trials. Lon S. Schneider reports personal fees from AC Immune, Alpha-cognition, Athira, Corium, Cortexyme, BioVie, Eli Lilly, GW Research, Lundbeck, Merck, Neurim, Ltd, Novo-Nordisk, Otsuka, Roche/Genentech. Cognition Therapeutics, Takeda; grants from Biohaven, Biogen, Eisai, Eli Lilly, Novartis. Author disclosures are available in the supporting information.

REFERENCES

1. Wright AJ, Mihura, JL, Pade H, McCord DM. *Guidance on psychological tele-assessment during the COVID-19 crisis*. American Psychological Association; 2020. https://www.apaservices.org/practice/reimbursement/health-codes/testing/tele-assessment-covid-19

2. Carlson S, Kim H, Devanand DP, Goldberg TE. Novel approaches to measuring neurocognitive functions in Alzheimer's disease clinical trials. *Curr Opin Neurol*. 2022;35(2):240-248. doi:10.1097/WCO0000000000001041
3. Erber NP. Interaction of audition and vision in the recognition of oral speech stimuli. *J Speech Hearing Res*. 1969;12(2):423-425. doi:10.1044/jshr.1202.423
4. Anonymous. ADC Clinical Task Force & National Alzheimer's Coordinating Center (2020). Instructions for Telephone Neuropsychological Battery, Form C2T. 2020. Accessed April 14, 2022. https://files.alz.washington.edu/documentation/uds3-np-c2t-instructions.pdf
5. Weintraub S, Besser L, Dodge HH, et al. Version 3 of the Alzheimer Disease Centers' neuropsychological test battery in the Uniform Data Set (UDS). *Alzheimer Dis Assoc Disord*. 2018;32(1):10-17. doi: 10.1097/WAD.0000000000000223
6. Beekly DL, Ramos EM, Lee WW, et al. The National Alzheimer's Coordinating Center (NACC) database: the uniform data set. *Alzheimer Dis Assoc Disord*. 2007;21(3):249-258. doi:10.1097/WAD.0b013e318142774e
7. Nasreddine ZS, Phillips NA, Bedirian V, et al. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *J Am Geriatr Soc*. 2005;53(4):695-699. doi:10.1111/j.1532-5415.2005.53221.x
8. Craft S, Newcomer J, Kanne S, et al. Memory improvement following induced hyperinsulinemia in Alzheimer's disease. *Neurobiol Aging*. 1996;17(1):123-130. doi:10.1016/0197-4580(95)02002-0
9. Koo TK, Li MY. A Guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. 2016;15(2):155-163. doi:10.1016/j.jcm.2016.02.012
10. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159-174.
11. Bruijnen CJWH, Dijkstra BAG, Walvoort SJW, et al. Psychometric properties of the Montreal Cognitive Assessment (MoCA) in healthy participants aged 18-70. *Int J Psychiatry Clin Pract*. 2020;24(3):293-300. doi:10.1080/13651501.2020.1746348
12. Zec S, Soriani N, Comoretto R, Baldi I. High agreement and high prevalence: the paradox of Cohen's Kappa. *Open Nurs J*. 2017;11:211-218. doi: 10.2174/1874434601711010211
13. Lo AHY, Humphreys M, Byrne GJ, Pachana NA. Test–retest reliability and practice effects of the Wechsler Memory Scale-III. *J Neuropsychol*. 2012;6(2):212-231. doi: 10.1111/j.1748-6653.2011.02023.x
14. Iverson GL. Interpreting change on the WAIS-III/WMS-III in clinical samples. *Arch Clin Neuropsychol*. 2001;16(2):183-191.
15. Waters GS, Caplan D. The reliability and stability of verbal working memory measures. *Behav Res Methods Instrum Comput*. 2003;35(4):550-564. doi:10.3758/BF03195534
16. Calamia M, Markon K, Tranel D. The robust reliability of neuropsychological measures: meta-analyses of test-retest correlations. *Clin Neuropsychol*. 2013;27(7):1077-1105. doi:10.1080/13854046.2013.809795

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Howard RS, Goldberg TE, Luo J, Munoz C, Schneider LS. Reliability of the NACC Telephone-administered Neuropsychological Battery (T-cog) and Montreal Cognitive Assessment for participants in the USC ADRC. *Alzheimer's Dement*. 2023;15:e12406. https://doi.org/10.1002/dad2.12406