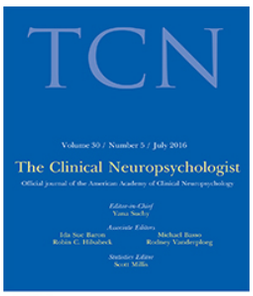

The Clinical Neuropsychologist

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ntcn20

Initial investigation of test-retest reliability of home-to-home teleneuropsychological assessment in healthy, English-speaking adults

Joshua T. Fox-Fuller, Julie Ngo, Celina F. Pluim, Rini I. Kaplan, Dong-Ho Kim, Juliana A. U. Anzai, Defne Yucebas, Soibifaa M. Briggs, Paula A. Aduen, Alice Cronin-Golomb & Yakeel T. Quiroz

To cite this article: Joshua T. Fox-Fuller, Julie Ngo, Celina F. Pluim, Rini I. Kaplan, Dong-Ho Kim, Juliana A. U. Anzai, Defne Yucebas, Soibifaa M. Briggs, Paula A. Aduen, Alice Cronin-Golomb & Yakeel T. Quiroz (2021): Initial investigation of test-retest reliability of home-to-home teleneuropsychological assessment in healthy, English-speaking adults, The Clinical Neuropsychologist, DOI: 10.1080/13854046.2021.1954244

To link to this article: https://doi.org/10.1080/13854046.2021.1954244

Published online: 26 Jul 2021.

THE CLINICAL NEUROPSYCHOLOGIST
https://doi.org/10.1080/13854046.2021.1954244

Initial investigation of test-retest reliability of home-to-home teleneuropsychological assessment in healthy, English-speaking adults

Joshua T. Fox-Fuller[a,b], Julie Ngo[a], Celina F. Pluim[a,b], Rini I. Kaplan[a], Dong-Ho Kim[a], Juliana A. U. Anzai[a], Defne Yucebas[a], Soibifaa M. Briggs[a], Paula A. Aduen[b], Alice Cronin-Golomb[a] and Yakeel T. Quiroz[b,c]

[a]Department of Psychological and Brain Sciences, Boston University, Boston, Massachusetts, USA; [b]Department of Psychiatry, Massachusetts General Hospital, Boston, Massachusetts, USA; [c]Department of Neurology, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA

ABSTRACT

Objective: Prior teleneuropsychological research has assessed the reliability between in-person and remote administration of cognitive assessments. Few, if any, studies have examined the test-retest reliability of cognitive assessments conducted in sequential clinic-to-home or home-to-home teleneuropsychological evaluations – a critical issue given the state of clinical practice during the COVID-19 pandemic. This study examined this key psychometric question for several cognitive tests administered over repeated videoconferencing visits 4-6 months apart in a sample of healthy English-speaking adults. **Methods:** A total of 44 participants (ages 18-75) completed baseline and follow-up cognitive testing 4-6 months apart. Testing was conducted in a home-to-home setting over HIPAA-compliant videoconferencing meetings on participants' audio-visual enabled laptop or desktop computers. The following measures were repeated at both virtual visits: the Controlled Oral Word Association Test (FAS), Category Fluency (Animals), and Digit Span Forward and Backward from the Wechsler Adult Intelligence Scale, Fourth Edition. Intraclass correlation coefficients (ICC), Pearson correlations, root mean square difference (RMSD), and concordance correlation coefficients (CCC) were calculated as test-retest reliability metrics, and practice effects were assessed using paired-samples t-tests. **Results:** Some tests exhibited small practice effects, and test-retest reliability was marginal or worse for all measures except FAS, which had adequate reliability (based on ICC and *r*). Reliability estimates with RMSD suggested that change within +/- 1 SD on these measures may reflect typical test-retest variability. **Conclusions:** The included cognitive measures exhibited questionable reliability over repeated home-to-home videoconferencing evaluations. Future teleneuropsychology test-retest reliability research is needed with larger, more diverse samples and in clinical populations.

ARTICLE HISTORY

Received 17 March 2021
Accepted 7 July 2021
Published online 27 July 2021

KEYWORDS

Telemedicine; telehealth; neuropsychology; reliability; teleneuropsychology; verbal fluency; category fluency; digit span

CONTACT Yakeel T. Quiroz yquiroz@mgh.harvard.edu Departments of Psychiatry and Neurology, Massachusetts General Hospital, Harvard Medical School, Building #39 1st Ave., Suite #101, Charlestown, MA 02129, USA
Supplemental data for this article is available online at https://doi.org/10.1080/13854046.2021.1954244.

Introduction

Research on the remote administration of cognitive tests was of great interest to neuropsychologists even prior to the rapid adoption of teleneuropsychology (tele-NP) during the COVID-19 pandemic (Brearly et al., 2017; Miller & Barr, 2017). Prior tele-NP research has demonstrated that many neuropsychological measures exhibit good reliability between video conferencing and in-person administration (Cullum et al., 2014; Jacobsen et al., 2003; Wadsworth et al., 2018). A recent review (Marra et al., 2020) underscored that the video conferencing design of most previous tele-NP studies followed a clinic-to-clinic approach, wherein the examiner was in their clinic and the participant was at a satellite clinic or in a different room in the same clinic as the examiner (e.g., Cullum et al., 2014; Wadsworth et al., 2018). The clinic-to-clinic approach is the most secure way to minimize potential tele-NP confounds, such as internet connectivity problems in the examiner or examinee's location, as well as an examinee's technological unfamiliarity with the chosen video conferencing technology (i.e., a technician at the satellite clinic can help set up and monitor the visit).

Many clinical and research activities early on in the COVID-19 pandemic, however, consisted of either home-to-home assessments (where the examiner and examinee were in their respective homes) or clinic-to-home assessments (where the examiner is at their clinic and the examinee is at home [e.g., Parks et al., 2021]). During cognitive assessments in which the examinee is in their home, many factors outside of the examiner's control can interfere with cognitive testing, including internet connectivity speed (e.g., audio and/or visual lagging), issues pertaining to the device the examinee is using for the evaluation (e.g., uncertainty about how visual stimuli are being viewed on a participant's screen), and environmental factors (e.g., pets or family members entering the examinee's room, or unexpected distracting noises in the background, such as a telephone ringing). Despite potential confounds relative to traditional in-person neuropsychological assessment, it is likely that many clinical and research tele-NP activities may continue even after the COVID-19 pandemic subsides.

There are clear potential benefits to clinic-to-home, and particularly home-to-home tele-NP, such as an examinee not needing to commute to a clinic for an evaluation and the examiner having a window into the everyday environment of the examinee. These benefits may help expand access to cognitive testing for many populations that are underserved by traditional face-to-face healthcare, and may also benefit neuropsychologists by reducing no-show rates and overall testing time when briefer tele-NP evaluations are clinically appropriate (Caze et al., 2020). Recent surveys of patients who received a tele-NP evaluation reported high patient satisfaction with the virtual evaluations (Appleman et al., 2021; Lacritz et al., 2020).

There are several outstanding questions, however, about the reliability and validity of cognitive tests that are widely used in in-person settings and are now being increasingly used in clinic-to-home or home-to-home tele-NP settings. Regardless of how the implementation of tele-NP continues, test reliability in tele-NP in these settings is an important psychometric issue that the field must address (Bilder et al., 2020; Brearly et al., 2017; Marra et al., 2020).

The current study examined the test-retest reliability of several widely-used cognitive assessments delivered over two home-to-home cognitive tele-NP assessment

sessions conducted 4-6 months apart as part of a larger study of healthy English-speaking adults in the United States. The measures examined in this study were the Controlled Oral Word Association (COWA) Test (FAS) (Lezak et al., 2004); Category Fluency (Animals) (Lezak et al., 2004); and Digit Span Forward and Backward from the Wechsler Adult Intelligence Scale, Fourth Edition (WAIS-IV; Wechsler, 2008). Previous in-person research found moderate test-retest reliability of COWA (FAS) between baseline and 6-month follow up in healthy, English-speaking adults (Ruff et al., 1996). Similarly, a prior study of healthy adults from the United Kingdom found that in-person test-retest reliability of Category Fluency (animals) was in the acceptable range (Harrison et al., 2000). In-person test-retest reliability on WAIS-IV Digit Span (Wechsler, 2008) and similar tasks (e.g., UK Biobank Numeric Memory; Fawns-Ritchie & Deary, 2020) was also found to be in the acceptable-to-good range in healthy, English-speaking adults. We hypothesized that we would find adequate test-retest reliability over the two video conferencing sessions that would be comparable to reliability metrics found in previous in-person research studies in healthy adults.

Methods

Participants

As part of a larger validation study of two computerized cognitive assessments recently developed by our group, 150 English-speaking participants were recruited, including 86 undergraduates from Boston University who participated in the study through their introductory psychology course, as well as 64 members of the broader public who were recruited across the United States using social media and the web (e.g., Facebook, Twitter, Craigslist, and ResearchMatch.com). The inclusion criteria were as follows: participants had to be over 18 years old, speak English as their primary language, have no self-reported vision issues (that were not corrected with glasses or contacts), and have no self-reported neurological or psychiatric disorders that to their knowledge impacted their cognitive functioning. Participants also had to report having access to an audiovisual (AV)-enabled computer, internet access, and a space in their home in which they could engage in cognitive testing privately.

When participants took a Qualtrics survey to determine their eligibility to participate, they attested by typing their name into a box on the survey that they had read the attached consent form and agreed to participate in the study. At baseline, undergraduates were compensated for their participation through course credit, whereas members of the general public received an e-gift card at the conclusion of their study visit. All participants who returned for follow-up testing received another e-gift card as compensation. This study was approved by the Boston University Charles River Campus Institutional Review Board. The data that support the findings of this study are available on reasonable request from the first author (JTFF).

Of the 150 original participants at baseline, 44 (29%) volunteered to return for a follow-up testing session in spring 2021, which was 4-6 months after their baseline visit (mean= 180.9 days; SD = 25.6 days; range= 122-218 days). It should be noted that 13 of the 150 participants were not invited to return for re-testing due to the following reasons: reporting familiarity with the cognitive measures from prior work

settings (n = 4); psychiatric/neurological history (e.g., concussion, obsessive-compulsive disorder) reported during the testing session that was not reported during pre-screening procedures (n = 2); a below-cutoff score on the Telephone Interview for Cognitive Status (n = 1) or Reliable Digit Span (n = 1); a medical event during baseline testing that prevented session competition (n = 1); testing on a tablet instead of computer (n = 1); significant environmental distractions during testing (e.g., screaming in another room, refusing to turn a cellphone off that was ringing repeatedly; n = 2); and significant internet connectivity disruption during the verbal cognitive tests (n = 1).

The resultant 44 participants with data from a baseline and follow-up visit consisted of 7 undergraduates and 37 members of the general public, and comprised the test-retest sample included in the analyses presented here. 34 of these participants self-reported their biological sex as female, and 10 self-reported as male. 25 participants self-identified as non-Hispanic white (56.8%), 14 as Asian American (31.8%), 3 as Latino/a (6.8%), and 2 as African American (4.5%). The demographic and cognitive data of the test-retest sample are provided in Table 1. Data from all 44 participants are included in the analyses outlined in this report.

Measures and procedures

Participants took part in home-to-home study visits via secure, Health Insurance Portability and Accountability Act (HIPAA)-compliant video conferencing meetings on Zoom® with a research assistant trained in the administration of cognitive assessments. The baseline study visit lasted between 60-90 minutes, and the follow-up visit lasted 30-45 minutes. Participants were instructed prior to testing to be in a private space in their home where they would be free of distractions, and research staff were also in private locations in their own home (due to COVID-19 lockdowns and social

Table 1. Demographic and cognitive data of the sample (n = 44).

Characteristic	Mean (SD)	Skewness (SE)	Range
Age (years)	33.98 (15.00)	1.18	18-75
Educational Attainment (years)	16.73 (2.44)	0.25	13-22
TICS Total Score, Baseline Visit	36.00 (1.91)	−0.04	32-40
Reliable Digit Span, Baseline Visit	10.68 (2.23)	0.74	7-17
FAS Total Raw Score, Baseline Visit	45.34 (13.25)	0.37	22-73
FAS (Tombaugh Norms), Baseline Visit	50.80 (11.76)	0.39	30-75
FAS Total Raw Score, Follow-Up Visit	48.45 (12.98)	0.52	22-73
FAS (Tombaugh Norms), Follow-Up Visit	53.50 (11.50)	0.49	32-80
Animals Total Raw Score, Baseline Visit	23.36 (5.32)	0.25	13-37
Animals (Tombaugh Norms), Baseline Visit	53.39 (9.97)	0.39	38-78
Animals Total Raw Score, Follow-Up Visit	23.30 (4.95)	0.42	15-33
Animals (Tombaugh Norms), Follow-Up Visit	53.50 (10.18)	0.86	37-80
WAIS-IV Digit Span Forward Total Raw Score, Baseline Visit	11.00 (2.42)	0.13	6-16
WAIS-IV Digit Span Forward Total Scaled Score, Baseline Visit	10.66 (3.20)	0.42	4-18
WAIS-IV Digit Span Forward Total Raw Score, Follow-Up Visit	11.50 (2.43)	−0.50	6-16
WAIS-IV Digit Span Forward Total Scaled Score, Follow-Up Visit	11.34 (3.16)	−0.30	4-18
WAIS-IV Digit Span Backward Total Raw Score, Baseline Visit	9.52 (2.39)	0.34	5-16
WAIS-IV Digit Span Backward Total Scaled Score, Baseline Visit	11.11 (3.06)	0.59	6-19
WAIS-IV Digit Span Backward Total Raw Score, Follow-Up Visit	10.16 (2.76)	0.47	5-16
WAIS-IV Digit Span Backward Total Scaled Score, Follow-Up Visit	11.89 (3.54)	0.51	6-19

Cognitive data are reported for the baseline visit and the follow-up visit, which was on average 180 days after the initial visit. SD = standard deviation; SE = standard error; TICS = Telephone Interview for Cognitive Status; FAS = letter fluency from the Controlled Oral Word Association Test; WAIS-IV = Wechsler Adult Intelligence Scale, Fourth Edition.

distancing guidelines) for the duration of the testing session. Issues that were encountered during remote cognitive testing over video conferencing (e.g., poor wireless connectivity, environmental distractions in the participant's home, etc.) were documented by the assessor and discussed in depth between the first and last authors to determine if that individual's data were compromised by the interference and should be excluded from analyses. (It should be noted none of the 44 included participants in this report had such issues at their baseline or follow-up virtual evaluations.) Participants and examiners used their personal AV-enabled laptop or desktop computers, reflecting the larger practice of home-to-home tele-NP during the COVID-19 pandemic. All data were stored on a password-protected, encrypted shared drive housed by Boston University using participants' subject identification numbers and the date of the evaluation.

The neuropsychological tests at baseline included the following: the Telephone Interview for Cognitive Status (TICS; Brandt et al., 1988), with some minor adaptations for administration over video conferencing vs. the phone (e.g., tapping fingers together 5 times on camera vs. tapping the phone 5 times); the Controlled Oral Word Association COWA (FAS)(Lezak et al., 2004); Category Fluency (animals) (Lezak et al., 2004); and Digit Span Forward and Backward from the WAIS-IV (Wechsler, 2008). The administration of WAIS-IV Digit Span was abbreviated with only the Forward and Backward conditions (i.e., without Sequencing). These measures were selected *a priori* as the battery for the previously mentioned, unpublished validation study of two newly developed computerized cognitive measures. The TICS suggested cutoff score of <32 was applied to exclude any participants who may have been exhibiting symptoms of cognitive impairment (Knopman et al., 2010). Reliable Digit Span (RDS) was calculated as an embedded measure of participant effort at baseline using the sum of the longest length of Digits Forward and Digits Backwards in which both trials within an item were correct. A score of ≤7 for RDS indicated potentially suboptimal effort (Schroeder et al., 2012). Participants who scored below the TICS or RDS cutoff at baseline or whose scores were removed from analysis after their baseline visit for other reasons (e.g., poor Wi-Fi connectivity, significant environmental distractions) were not invited to return for the follow-up testing session.

At the follow-up visit, study staff re-administered the Controlled Oral Word Association Test (FAS), Category Fluency (animals), and WAIS-IV Digit Span Forward and Backward. The TICS was removed from the follow-up visit, as all participants who were invited to participate in re-testing had a TICS score greater than the suggested cutoff at their baseline visit, 4-6 months prior to the follow-up visit. Participants who returned for re-testing took another eligibility survey in which they self-reported no changes in their neurological or psychiatric history since their baseline visit, as well as no changes with their vision. Participants chose their desired time slot to complete testing based on their schedules at both evaluations.

Statistical analysis

Statistical analyses were performed in SPSS Version 26 (IBM SPSS Statistics, 2019). Chi-square and independent samples t-tests to examine any differences between the

returning group (n = 44) versus the participants who had only baseline testing in the larger study (n = 106). We also examined if there were practice effects on the tests by conducting paired samples t-tests between the mean values on the cognitive tests of the sample at baseline and follow-up, using p = .05 as our alpha value. For all analyses, we used the raw scores of the participants in addition to scaled scores (Tombaugh et al., 1999 for FAS and animal fluency; Wechsler, 2008 for WAIS-IV Digit Span Forward and Backward).

Test-retest reliability was assessed across the two home-to-home tele-NP evaluations in several ways. We first used Pearson correlations as a measure of reliability between the scores on each measure at the two time points. We also calculated intraclass correlation coefficients (ICCs) using a two-way mixed, agreementdesign and the "single measures" output. The interpretations of Pearson correlations and ICCs for reliability have varying specifications for suggested cutoffs. For example, Slick (2006) suggested the following reliability cutoffs: 0.60 – 0.69 = marginal; 0.70 – 0.79 = adequate; 0.80 – 0.89 = high; 0.90+ = very high. These values generally align with Cicchetti (1994), who was slightly more liberal in interpretation (e.g., 0.60 – 0.74 = good reliability, 0.75+ = excellent reliability). Nunnally and Bernstein (1994), however, set a much higher threshold for reliability, suggesting 0.90 as the minimum acceptable value for test-retest reliability. We interpret the reliability metrics in this report cautiously within this context, as tests with low reliability are subject to increased possibility of false-positive and false-negative errors in clinical contexts (Charter & Feldt, 2001). Additionally, Pearson correlations focus on linear agreement between the two time-points, ignoring means and standard deviations, and ICCs assume equal variance (Barchard, 2012). These assumptions of Pearson correlations and ICCs necessitate additional investigation of reliability using methods that consider the means and standard deviations.

Root mean square differences (RMSD) and concordance correlation coefficients (CCC) calculated based on absolute agreement (i.e., Time1 = Time2) may be a more preferable way to examine test-retest reliability (Barchard, 2012). RMSD represents the average amount by which scores tend to differ by calculating the square root of the average squared difference score (Barchard, 2012). The lowest possible RMSD score is 0 (representing perfect agreement), with the maximum value being the difference between the minimum and maximum possible scores on a given measure in the sample (Barchard, 2012). The largest possible value of the CCC is 1, with values closer to 1 representing better agreement. The CCC is a measure of agreement between continuous measures that considers differences in means and variances for the two measures, unlike Pearson correlations and ICCs (Barchard, 2012). We calculated RMSD and CCC values for agreement for the measures using the Microsoft Excel® macro provided in the supplementary materials of Barchard (2012). Agreement of the measures at each timepoint was further assessed with a Bland–Altman plot.

Results

We first examined if there were differences between the participants who returned to take part in re-testing (n = 44) versus those who did not (n = 106). The majority of the 106 participants who did not engage in re-testing were undergraduate students

(79 of the 106 non-returners were undergraduates = 74.5%). The re-testing group has a significantly larger ratio of members of the general public to undergraduates (37:7) than the non-returners (27:79), with a chi-square value of 43.7 ($p<.001$). Relatedly, the returners were on average older (M = 34.0 years, SD =15.0) than the non-returners (M = 23.4 years, SD = 9.2) ($t = 5.28$, $p<.001$) and had more years of formal education (returners: M = 16.7 years, SD = 2.4; non-returners: M = 14.3 years, SD = 2.2; t = 5.97, $p<.001$; but note that the undergraduates had mostly not yet reached their terminal education level). The returners and non-returners did not differ on proportion of biological sex in each group, TICS total score, Digit Span Forward and Backward raw or scaled scores, Reliable Digit Span, FAS total raw score, and animal fluency total raw score ($p > 0.18$ for all). Additionally, among the 44 individuals who completed baseline and follow-up testing, the only cognitive or demographic variable that was significantly skewed was age, wherein many participants in the test-retest group were of younger age (Table 1).

The sample exhibited a practice effect of about 3 words on letter fluency (Visit 1 FAS mean: 45.34 words [SD: 13.25]; Visit 2 FAS mean: 48.45 [SD: 12.98]). This practice effect on letter fluency was small, but was significant (t = 2.38, $p = .02$). There was also a small practice effect on WAIS-IV Digit Span Backward, with participants gaining an average of 0.6 raw points at follow-up (t = 2.04, $p = .05$). Participants did not exhibit a significant practice effect on category fluency (animals) or WAIS-IV Digit Span Forward raw scores (respectively: t = .09, $p = .92$; t = 1.57, $p = .12$).

Table 2 lists the reliability statistics. Based on reliability cutoffs suggested by Slick (2006), which lie between the more conservative cutoffs proposed by Nunnally and Bernstein (1994) and the more liberal cutoffs of Cicchetti (1994), FAS demonstrated adequate repeated videoconferencing test-retest reliability per Pearson and ICC estimates. WAIS-IV Digit Span Forward and Backward raw and scaled scores exhibited marginal repeated videoconferencing test-retest reliability per Pearson and ICC estimates, and animal fluency demonstrated unacceptable repeated videoconferencing test-retest reliability.

As Pearson correlations and ICCs do not consider means and SDs, and ICCs in particular assume equal variance, further examination of reliability using RMSD and CCCs for absolute agreement provided insight about the mean expected changes in this sample on these measures. As shown in Table 3, only WAIS-IV Digit Span Backward change scores were significantly skewed (in a negative direction), with

Table 2. Reliability estimates.

Cognitive measure	Pearson *r*	ICC (95% CI)	RMSD	CCC
FAS, Raw Score	0.78	0.76 (0.59 – 0.87)	9.12	0.76
FAS, Tombaugh Norms	0.78	0.76 (0.60 – 0.87)	8.06	0.76
Animals, Raw Score	0.52	0.52 (0.27 – 0.71)	5.00	0.52
Animals, Tombaugh Norms	0.54	0.55 (0.30 – 0.73)	9.51	0.54
WAIS-IV Digit Span Forward, Raw Score	0.62	0.61 (0.39 – 0.77)	2.14	0.61
WAIS-IV Digit Span Forward, Scaled Score	0.67	0.66 (0.45 – 0.80)	2.66	0.65
WAIS-IV Digit Span Backward, Raw Score	0.69	0.66 (0.45 – 0.80)	2.14	0.66
WAIS-IV Digit Span Backward, Scaled Score	0.70	0.67 (0.47 – 0.81)	2.70	0.67

FAS and animal fluency use demographically-adjusted scores calculated according to Tombaugh et al. (1999). WAIS-IV Digit Span Forward and Backward Scaled Scores are derived according to the WAIS-IV technical manual (Wechsler, 2008). FAS = letter fluency from the Controlled Oral Word Association Test; WAIS-IV = Wechsler Adult Intelligence Scale, Fourth Edition.

Table 3. Change scores of letter fluency, animal fluency, and digit span forward and backward.

Cognitive measure	Mean (SD)	Range	Skewness
FAS, Raw Score	3.11 (8.67)	−14 – 25	0.13
FAS, Tombaugh Norms	2.70 (7.68)	−13 – 22	0.11
Animals, Raw Score	−0.07 (5.06)	−13 – 10	−0.22
Animals, Tombaugh Norms	0.11 (9.62)	−24 – 19	−0.18
WAIS-IV Digit Span Forward, Raw Score	0.50 (2.11)	−7 – 5	−1.16
WAIS-IV Digit Span Forward, Scaled Score	0.68 (2.60)	−8 – 6	−0.96
WAIS-IV Digit Span Backward, Raw Score	0.64 (2.07)	−4 – 6	0.25
WAIS-IV Digit Span Backward, Scaled Score	−0.77 (2.61)	−8 – 5	−0.39

FAS and animal fluency use demographically-adjusted scores calculated according to Tombaugh et al. (1999). WAIS-IV Digit Span Forward and Backward Scaled Scores are derived according to the WAIS-IV technical manual (Wechsler, 2008). SD = standard deviation; FAS = letter fluency from the Controlled Oral Word Association Test; WAIS-IV = Wechsler Adult Intelligence Scale, Fourth Edition.

no other change scores showing significant skewness. For the raw and demographically-adjusted scores of FAS, the RMSD (i.e., mean expected change) was just above two-thirds the SD of the baseline raw and adjusted FAS scores; FAS had a CCC of 0.76 for both the raw and adjusted scores. For animal fluency, the RMSD was nearly equivalent to the SD for the raw and demographically-adjusted scale scores, and animal fluency carried a CCC of 0.52 for the raw score and 0.54 for the adjusted scores. WAIS-IV Digit Span Forward raw and scaled scores had RMSDs that were nearly 90% and 80% for each respective baseline SD; the CCCs for WAIS-IV Digit Span Forward raw and scaled scores were 0.61 and 0.65, respectively. WAIS-IV Digit Span Backward raw and scaled scores had RMSDs that were approximately 90% of each respective baseline SD; the CCCs for WAIS-IV Digit Span Backward raw and scaled scores were 0.66 and 0.67, respectively.

Bland-Altman plots using demographically-adjusted T-scores or scaled scores (Figure 1A–D) provide visualizations that help contextualize RMSD findings. FAS, which had the strongest CCC out of all of the measures and smallest proportional RMSD when compared to its baseline SD, had only 3 individuals who exhibited worse than −1 SD change and 8 individuals who had more than +1 SD improvement on their FAS demographically-adjusted values relative to an assumed change of 0 SD (Figure 1A; Supplementary Materials). On animal fluency (Figure 1B; Supplementary Materials), 5 participants exhibited worse than −1 SD change and 7 demonstrated more than +1SD improvement in their demographically-adjusted scores. For WAIS-IV Digit Span Forward Scaled Scores (Figure 1C; Supplementary Materials), 2 individuals showed worse than −1 SD change, whereas 6 exhibited more than +1SD change. On WAIS-IV Digit Span Backward Scaled Scores (Figure 1D; Supplementary Materials), 5 participants demonstrated worse than −1 SD change, while 3 exhibited more than +1SD improvement.

Discussion

Prior to the COVID-19 pandemic in 2020-21, a large focus of tele-NP research was focused on the reliability and validity of phone and video-based cognitive assessment in clinic-to-clinic settings (Cullum et al., 2014; Wadsworth et al., 2016, 2018) and sought to determine the reliability between face-to-face and remote administration of cognitive assessments (Chapman et al., 2019, 2020; Cullum et al., 2014; Wadsworth

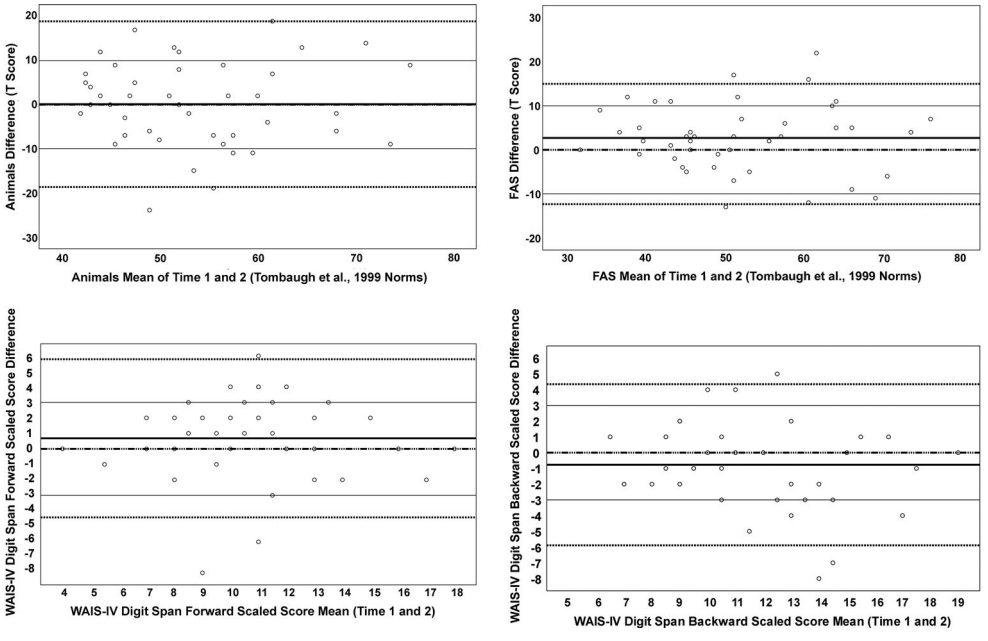

Figure 1. Bland–Altman plot of agreement between repeated administrations of cognitive measures over videoconferencing.
Shown is the agreement (change score relative to mean of Time 1 and 2) for each participant on (A) animal fluency, (B) FAS, (C) WAIS-IV Digit Span Forward, and (D) WAIS-IV Digit Span Backward using demographically-adjusted and scalded scores. The solid black lines represent the sample-derived mean change values, whereas thin black lines represent +/- 1 SD from a mean of 0. The dashed and solid line represents a mean of 0. Lastly, the dashed lines represent the sample-derived 95% confidence interval on the measure. FAS = letter fluency from the Controlled Oral Word Association Test; WAIS-IV = Wechsler Adult Intelligence Scale, Fourth Edition.

et al., 2016, 2018). Balancing the public health demands of the pandemic with the ethical considerations of continuing care for patients and for continuing research studies, the field had to adapt rapidly over the last year to provide clinic-to-home and home-to-home virtual assessments, despite clear need for examination of the reliability and validity of tele-NP in these settings (Bilder et al., 2020; Postal et al., 2021). The current study of test-retest reliability across 4-6 months for some commonly used clinical assessments – COWA (FAS), Category Fluency (animals), and WAIS-IV Digit Span Forward and Backward – expands upon prior tele-NP reliability and validity research. This is the first study, to our knowledge, to examine the stability of these measures over time within individuals assessed via home-to-home video conferencing at both baseline and follow-up examinations and hence provides an important starting point for future investigation on this topic.

Using classical test-retest reliability metrics (Pearson *r* and ICCs) and cutoffs proposed by Slick (2006), we found that FAS demonstrated adequate repeated videoconferencing test-retest reliability at 4-6 months in our sample of healthy English-speaking adults in the United States, whereas WAIS-IV Digit Span Forward and Backward exhibited marginal test-retest reliability in this setting. Animal fluency, according to this cutoff and these metrics, exhibited unacceptable repeated videoconferencing test-retest reliability.

As has been suggested previously (Marra et al., 2020), test-retest reliability may have been better in letter than animal fluency because letter fluency has three trials relative to just one trial for animal fluency. In general, longer tests tend to exhibit more stable reliability metrics, so it is possible that test-retest reliability on a three-trial category fluency task over repeated tele-NP visits may be stronger relative to the sole trial of animal fluency. The marginal test-retest reliabilities found on WAIS-IV Digit Span Forward (0.67 – scaled score) and Digit Span Backward (0.70 – scaled score) in this context of repeated home-to-home tele-NP visits was a slight departure (particularly on Digit Span Forward) from the test-retest reliabilities on these measures reported in the WAIS-IV normative studies. In the entire WAIS-IV normative sample, test-retest reliability was 0.77 for Digit Span Forward and 0.71 for Digit Span Backward (Wechsler, 2008). Notably, the test-retest interval for the in-person WAIS-IV normative studies (mean = 22 days, range = 8-82 days) was much shorter than the current study (mean = 181 days; range = 122-218 days). Also, WAIS-IV normative study participants were older (mean = 52.6 years, SD = 23.6 years) than participants in the current study (mean = 34.0 years, SD = 15.0 years), and the WAIS-IV normative study test-retest sample was substantially larger (n = 298) than the current study (n = 44). These differences in the study characteristics between the two studies could explain the small differences seen on test-retest reliability estimates that were reported.

To account for means, SDs, and variance in scores, RMSD and CCC were also calculated as measurements of test-retest reliability. RMSD and CCC values were generally in agreement with the findings from Pearson correlations and ICCs for the included measures, with FAS exhibiting the best reliability out of all of the measures. Inspection of the Bland-Altman plots (Figure 1A–D) and change values on the demographically-adjusted or scaled scores (Supplementary Material), however, showed that many participants exceeded +/- 1 SD change at their follow-up visit relative to their baseline score. For FAS, 11/44 participants (25%) exhibited more than +/- 1 SD change, and 12/44 participants on animal fluency (27%) demonstrated more than +/- 1 SD change across the two visits. 8/44 participants (18%) had scores on both WAIS-IV Digit Span Forward and Backward that represented a change of +/- 1 SD. Given that this sample was a group of self-reported healthy, English-speaking individuals, it may be inferred that intra-individual change on the measures within approximately +/- 1 SD may reflect within expectation variability on these measures over repeated videoconferencing administrations within a 4-6-month period.

Our work adds to previous in-person reliability studies of these commonly-used measures, demonstrating that these tests exhibit relatively wide estimates of reliability over repeated video conference-based assessments in healthy, English speaking adults. Clinically, change of scores within +/- 1 SD on these measures within a 4-6 month period could reflect normal cognitive variability on these measures when administered over tele-NP with examinees in their home. Aside from letter fluency (FAS), the other measures in this study (animal fluency; WAIS-IV Digit Span Forward and Backward) demonstrated marginal or worse reliability, suggesting that clinicians engaged in tele-NP with these measures should use caution in interpreting scores on these tests. Larger, more representative tele-NP reliability and validity studies are desperately needed in clinic-to-home or home-to-home settings to both replicate (or counter)

the findings of this study and, more broadly, expand the field's psychometric understanding of other tests administered via tele-NP.

Limitations

Though this study still fills a critical gap in the literature given the paucity of such research at this time, one limitation of our study is the small sample size (n = 44). The educational attainment of our sample (at least a 4-year university degree, on average) is also a limitation and necessitates replication in individuals with lower levels of formal educational attainment. We used an abbreviated version on one of the cognitive tests (WAIS-IV Digit Span, Forward and Backward Span only without Sequencing), which is sometimes done in research settings. However, this limits the present study's direct utility for clinical tele-NP settings, as the full measure (with Sequencing) is more commonly used in clinical evaluations (Wechsler, 2008). We recommend replication with the full WAIS-IV Digit Span measure and advise caution in extrapolating our findings about the WAIS-IV Digit Span Forward and Backward to the full measure.

Another limitation of our work is that we required participants to meet via HIPAA-compliant video conferencing meetings on their own AV-enabled laptop or desktop computer. Other technologies, such as the conventional telephone and smartphone, are less expensive for consumers and have broader circulation in the public than personal computers. Additionally, for many people it may be impractical to do cognitive testing over tele-NP in their home if they do not have a private, quiet space (e.g., their own bedroom). Moreover, as discussed in a recent review by Marra and colleagues (2020), racial and ethnic minority groups in the United States are less likely to own a personal computer and have at-home access to the broadband internet speeds requisite for remote cognitive assessment (Perrin & Turner, 2019). Approximately 10% of the United States population does not use the internet at all (Anderson et al., 2019), making a standard telephone call the only way to potentially bridge digital health inequities with these individuals.

We also restricted eligibility criteria for this study to people who identified English as their primary language and reported no history of cognitive or psychiatric comorbidities that would interfere with their cognition. As such, replication of our findings is needed in linguistically diverse populations and in clinical populations, as providing equitable access to neuropsychological care must be a central goal of continued tele-NP research and implementation. Additionally, if participants encountered significant difficulty with the video conferencing platform or had documented connectivity issues and/or significant environmental distractions during their baseline visit that impacted data collection, they were not included in the list of participants who were invited to return for re-testing. By systematically excluding participants with known connectivity problems or significant distractions at baseline from re-testing, we lost an opportunity to examine test-retest reliability of these neuropsychological tests within the context of these ecological threats to tele-NP assessment. Future test-retest research of cognitive tests over repeated video conferencing sessions may wish to include participants with documented technological problems or recorded environmental distractions, as this will more

accurately provide an empirical understanding of test-retest reliability in a tele-NP setting that reflects the routine challenges faced by examiners.

Lastly, we note that the instruments used in this study do not require videoconferencing for administration and could be administered over a telephone audio call. A prior study has investigated this question over repeated telephone evaluations in older adult women, finding generally consistent reliability findings (using Pearson correlations) as the present study (Rapp et al., 2021). The measures we used in our study were selected *a priori* as outcome measures for a larger ongoing study. Nonetheless, in both clinical and research settings, these measures are being used via videoconferencing, warranting the investigation described here. Future studies of the psychometric properties of neuropsychological assessment via videoconferencing should also examine tests that do require visual presentation of stimuli (e.g., the Boston Naming Test; Kaplan et al., 2010).

Conclusion

Prior tele-NP research has largely focused on demonstrating the reliability of cognitive assessments conducted over the telephone and video conferencing relative to in-person administration. This previous research has taken place largely in clinic-to-clinic settings. The present study of healthy English-speaking adults provides evidence that the cognitive tests included in the study – consisting of validated, commonly used measures that examine letter fluency, semantic fluency, and auditory attention/working memory –exhibit variable reliability metrics. Changes of scores on these measures over repeated home-to-home tele-NP visits within +/- 1 SD may reflect within expectation cognitive fluctuation in healthy English-speaking adults. Understanding the wide range of potential test-retest reliability on these measures is essential to reducing false-positive and negative-errors (Binder et al., 2009), and caution is advised in the clinical interpretation of changes on these measures over repeated video conferencing visits. Clinic-to-home and home-to-home tele-NP is a promising way to expand access to neuropsychological services, but continued research is needed on the reliability and validity of testing in these tele-NP settings. In particular, future research should examine test-retest reliability in larger and more diverse samples (e.g., non-English speaking individuals) and should examine test-retest reliability in tele-NP in the context of ecological threats to this modality of assessment (e.g., distractions and internet connectivity problems).

Acknowledgements

The authors would like to thank our participants for their time and collaboration during this study. We also wish to acknowledge other members of our study team for their involvement in data collection and discussions. Specifically, we acknowledge the following undergraduate students from Boston University: Karen Park, Samantha Casey, Kyona Schacht, Lucia Ceron Giraldo, and Sofia Hernandez.

Disclosure statement

YTQ has served as a paid consultant for Biogen (Not related to the content in the manuscript). All co-authors report no conflicts of interest.

Funding

This work was supported by the National Institute on Aging (1F31AG062158-01A1 to JTFF). JN and DHK report support from the Boston University Undergraduate Research Opportunities Program (UROP). AA was supported by a grant from the Massachusetts General Hospital Center for Diversity and Inclusion. YTQ was supported by grants from the NIH National Institute on Aging (R01AG054671, R01AG066823).

References

Anderson, M., Perrin, A., Jiang, J., & Kumar, M. (2019). 10% of Americans don't use the internet. Who are they?Pew Research Center. https://www.pewresearch.org/fact-tank/2019/04/22/some-americans-dont-use-the-internet-who-are-they/

Appleman, E. R., O'Connor, M. K., Boucher, S. J., Rostami, R., Sullivan, S. K., Migliorini, R., & Kraft, M. (2021). Teleneuropsychology clinic development and patient satisfaction. *The Clinical Neuropsychologist*, *35*(4), 819–819. https://doi.org/10.1080/13854046.2020.1871515

Barchard, K. A. (2012). Examining the reliability of interval level data using root mean square differences and concordance correlation coefficients. *Psychological Methods*, *17*(2), 294–308. https://doi.org/10.1037/a0023351

Bilder, R. M., Postal, K. S., Barisa, M., Aase, D. M., Cullum, C. M., Gillaspy, S. R., Harder, L., Kanter, G., Lanca, M., Lechuga, D. M., Morgan, J. M., Most, R., Puente, A. E., Salinas, C. M., & Woodhouse, J. (2020). Inter organizational practice committee recommendations/Guidance for teleneuropsychology in response to the COVID-19 Pandemic†. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists*, *35*(6), 647–659. https://doi.org/10.1093/arclin/acaa046

Binder, L. M., Iverson, G. L., & Brooks, B. L. (2009). To Err is Human: "Abnormal" neuropsychological scores and variability are common in healthy adults. *Archives of Clinical Neuropsychology*, *24*(1), 31–46. https://doi.org/10.1093/arclin/acn001

Brandt, J., Spencer, M., & Folstein, M. (1988). The telephone interview for cognitive status. *Neuropsychiatry, Neuropsychology, and Behavioral Neurology*, *1*(2), 111–117.

Brearly, T. W., Shura, R. D., Martindale, S. L., Lazowski, R. A., Luxton, D. D., Shenal, B. V., & Rowland, J. A. (2017). Neuropsychological test administration by videoconference: A systematic review and meta-analysis. *Neuropsychology Review*, *27*(2), 174–186. https://doi.org/10.1007/s11065-017-9349-1

Caze, T., II, Dorsman, K. A., Carlew, A. R., Diaz, A., & Bailey, K. C. (2020). Can you hear me now? Telephone-based teleneuropsychology improves utilization rates in underserved populations. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists*, *35*(8), 1234–1239. https://doi.org/10.1093/arclin/acaa098

Chapman, J. E., Cadilhac, D. A., Gardner, B., Ponsford, J., Bhalla, R., & Stolwyk, R. J. (2019). Comparing face-to-face and videoconference completion of the Montreal Cognitive Assessment (MoCA) in community-based survivors of stroke. *Journal of Telemedicine and Telecare*, https://doi.org/10.1177/1357633X19890788

Chapman, J. E., Ponsford, J., Bagot, K. L., Cadilhac, D. A., Gardner, B., & Stolwyk, R. J. (2020). The use of videoconferencing in clinical neuropsychology practice: A mixed methods evaluation of neuropsychologists' experiences and views. *Australian Psychologist*, *55*(6), 618–633. https://doi.org/10.1111/ap.12471

Charter, R. A., & Feldt, L. S. (2001). Meaning of reliability in terms of correct and incorrect clinical decisions: The art of decision making is still alive. *Journal of Clinical and Experimental Neuropsychology*, *23*(4), 530–537. https://doi.org/10.1076/jcen.23.4.530.1227

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284. https://doi.org/10.1037/1040-3590.6.4.284

Cullum, C. M., Hynan, L. S., Grosch, M., Parikh, M., & Weiner, M. F. (2014). Teleneuropsychology: Evidence for video teleconference-based neuropsychological assessment. *Journal of the*

International Neuropsychological Society: JINS, *20*(10), 1028–1033. https://doi.org/10.1017/S1355617714000873

Fawns-Ritchie, C., & Deary, I. J. (2020). Reliability and validity of the UK Biobank cognitive tests. *Plos One*, *15*(4), e0231627. https://doi.org/10.1371/journal.pone.0231627

Harrison, J. E., Buxton, P., Husain, M., & Wise, R. (2000). Short test of semantic and phonological fluency: Normal performance, validity and test-retest reliability. *The British Journal of Clinical Psychology*, *39*(2), 181–191. https://doi.org/10.1348/014466500163202

IBM SPSS Statistics (26.0). (2019). [Computer software]. IBM Corp.

Jacobsen, S. E., Sprenger, T., Andersson, S., & Krogstad, J.-M. (2003). Neuropsychological assessment and telemedicine: A preliminary study examining the reliability of neuropsychology services performed via telecommunication. *Journal of the International Neuropsychological Society: JINS*, *9*(3), 472–478. https://doi.org/10.1017/S1355617703930128

Kaplan, E., Goodglass, H., & Weintraub, S., (2000). The Boston Naming Test . (2nd ed.). Philadelphia, PA: Lippincott Williams & Wilkins

Kaplan, E., Goodglass, H., Weintraub, S., Segal, O., & Loon-Vervoorn, A. v. (2001). *Boston naming test*. Pro-ed;/z-wcorg/.

Knopman, D. S., Roberts, R. O., Geda, Y. E., Pankratz, V. S., Christianson, T. J. H., Petersen, R. C., & Rocca, W. A. (2010). Validation of the telephone interview for cognitive status-modified in subjects with normal cognition, mild cognitive impairment, or dementia. *Neuroepidemiology*, *34*(1), 34–42. https://doi.org/10.1159/000255464

Lacritz, L. H., Carlew, A. R., Livingstone, J., Bailey, K. C., Parker, A., & Diaz, A. (2020). Patient satisfaction with telephone neuropsychological assessment. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists*, *35*(8), 1240–1248. https://doi.org/10.1093/arclin/acaa097

Lezak, M. D., Howieson, D. B., Loring, D. W., & Fischer, J. S. (2004). *Neuropsychological assessment*. Oxford University Press.

Marra, D. E., Hamlet, K. M., Bauer, R. M., & Bowers, D. (2020). Validity of teleneuropsychology for older adults in response to COVID-19: A systematic and critical review. *The Clinical Neuropsychologist*, *34*(7-8), 1411–1452. https://doi.org/10.1080/13854046.2020.1769192

Miller, J. B., & Barr, W. B. (2017). The technology crisis in neuropsychology. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists*, *32*(5), 541–554. https://doi.org/10.1093/arclin/acx050

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

Parks, A. C., Davis, J., Spresser, C. D., Stroescu, I., & Ecklund-Johnson, E. (2021). Validity of in-home teleneuropsychological testing in the wake of COVID-19. *Archives of Clinical Neuropsychology*, https://doi.org/10.1093/arclin/acab002

Perrin, A., & Turner, E. (2019). *Smartphones help blacks, Hispanics bridge some – but not all – digital gaps with whites*. Pew Research Center. https://www.pewresearch.org/fact-tank/2019/08/20/smartphones-help-blacks-hispanics-bridge-some-but-not-all-digital-gaps-with-whites/

Postal, K. S., Bilder, R. M., Lanca, M., Aase, D. M., Barisa, M., Holland, A. A., Lacritz, L., Lechuga, D. M., McPherson, S., Morgan, J., & Salinas, C. (2021). Inter organizational practice committee guidance/recommendation for models of care during the novel coronavirus pandemic. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists*, *36*(1), 17–28. https://doi.org/10.1093/arclin/acaa073

Rapp, S. R., Legault, C., Espeland, M. A., Resnick, S. M., Hogan, P. E., Coker, L. H., & CAT Study Group (2012). Validation of a cognitive assessment battery administered over the telephone. *Journal of the American Geriatrics Society*, *60*(9), 1616–1623. https://doi.org/10.1111/j.1532-5415.2012.04111.x

Ruff, R. M., Light, R. H., Parker, S. B., & Levin, H. S. (1996). Benton controlled oral word association test: Reliability and updated norms. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists*, *11*(4), 329–338.

Schroeder, R. W., Twumasi-Ankrah, P., Baade, L. E., & Marshall, P. S. (2012). Reliable digit span: A systematic review and cross-validation study. *Assessment*, *19*(1), 21–30. https://doi.org/10.1177/1073191111428764

Slick, D. J. (2006). Psychometrics in neuropsychological assessment. *A Compendium of Neuropsychological Tests: Administration, Norms, and Commentary*, *3*, 3–43.

Tombaugh, T. N., Kozak, J., & Rees, L. (1999). Normative data stratified by age and education for two measures of verbal fluency: FAS and animal naming. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists*, *14*(2), 167–177. https://doi.org/10.1016/S0887-6177(97)00095-4

Wadsworth, H. E., Dhima, K., Womack, K. B., Hart, J., Weiner, M. F., Hynan, L. S., & Cullum, C. M. (2018). Validity of teleneuropsychological assessment in older patients with cognitive disorders. *Archives of Clinical Neuropsychology : The Official Journal of the National Academy of Neuropsychologists*, *33*(8), 1040–1045. https://doi.org/10.1093/arclin/acx140

Wadsworth, H. E., Galusha-Glasscock, J. M., Womack, K. B., Quiceno, M., Weiner, M. F., Hynan, L. S., Shore, J., & Cullum, C. M. (2016). Remote neuropsychological assessment in rural American Indians with and without cognitive impairment. *Archives of Clinical Neuropsychology*, *31*(5), 420–425. https://doi.org/10.1093/arclin/acw030

Wechsler, D. (2008). Wechsler adult intelligence scale–Fourth Edition (WAIS–IV). *San Antonio, TX: NCS Pearson*, *22*(498), 1.