




Research Article

Evaluation of the Uniform Data Set version 3 teleneuropsychological measures

Theresa F. Gierzynski¹ , Allyson Gregoire¹, Jonathan M. Reader¹, Rebecca Pantis¹, Stephen Campbell¹, Arijit Bhaumik¹, Annalise Rahman-Filipiak², Judith Heidebrink¹, Bruno Giordani^{1,2}, Henry Paulson¹ and Benjamin M. Hampstead^{2,3}

¹Department of Neurology, University of Michigan, Ann Arbor, MI, USA, ²Department of Psychiatry, University of Michigan, Ann Arbor, MI, USA and ³VA Ann Arbor Healthcare System, Mental Health Service, Ann Arbor, MI, USA

Abstract

Objective: Few studies have evaluated in-home teleneuropsychological (teleNP) assessment and none, to our knowledge, has evaluated the National Alzheimer's Coordinating Center's (NACC) Uniform Data Set version 3 tele-adapted test battery (UDS v3.0 t-cog). The current study evaluates the reliability of the in-home UDS v3.0 t-cog with a prior in-person UDS v3.0 evaluation. **Method:** One hundred and eighty-one cognitively unimpaired or cognitively impaired participants from a longitudinal study of memory and aging completed an in-person UDS v3.0 and a subsequent UDS v3.0 t-cog evaluation (~16 months apart) administered either via video conference ($n = 122$) or telephone ($n = 59$). **Results:** We calculated intraclass correlation coefficients (ICCs) between each time point for the entire sample. ICCs ranged widely (0.01–0.79) but were generally indicative of “moderate” (i.e., ICCs ranging from 0.5–0.75) to “good” (i.e., ICCs ranging from 0.75–0.90) agreement. Comparable ICCs were evident when looking only at those with stable diagnoses. However, relatively stronger ICCs (Range: 0.35–0.87) were found between similarly timed in-person UDS v3.0 evaluations. **Conclusions:** Our findings suggest that most tests on the UDS v3.0 t-cog battery may serve as a viable alternative to its in-person counterpart, though reliability may be attenuated relative to the traditional in-person format. More tightly controlled studies are needed to better establish the reliability of these measures.

Keywords: teleneuropsychology; older adults; memory; cognition; aging; mild cognitive impairment; dementia; in-home cognitive assessment

(Received 21 December 2022; final revision 18 May 2023; accepted 29 May 2023; First Published online 27 June 2023)

Introduction

COVID-19 related precautions forced the field of neuropsychology to rapidly embrace telecommunication-based evaluations. At the peak of the COVID-19 crisis, leading public health organizations advised millions to shelter in place and limit person-to-person contact to prevent transmission (World Health Organization, 2020). Consequently, telemedicine became a critical medium for health care delivery among neuropsychologists with many providers pivoting to virtually based assessments (Hammers et al., 2020; Marra, Hoelzle, et al., 2020; Zane et al., 2021). To accommodate this sudden increase in telehealth utilization, insurance billing and reimbursement structures became more flexible (Centers for Medicare and Medicaid Services, 2021) and best practice guidelines emerged to support the responsible and ethical provision of remote neuropsychological services (Bilder et al., 2020). In addition, research that required ongoing in-person participation quickly adapted procedures to promote study continuity and to allow for remote data collection. These efforts were guided by the limited telehealth literature available at that time, little of which evaluated home-based virtual assessment.

Teleneuropsychology (teleNP), which the Inter Organizational Practice Committee defines as the use of any audiovisual technology (e.g., telephone, video conference) to facilitate remote neuropsychological assessment (Bilder et al., 2020), is being increasingly relied upon to bridge gaps in the provision of neuropsychological services, particularly when in-person evaluations are not possible. Although teleNP was used infrequently before the global COVID-19 pandemic (Miller & Barr, 2017), its adoption has grown and many neuropsychologists report an increased use of teleNP for clinical interviewing, test administration, and feedback (Hammers et al., 2020). For the purposes of this study, the term “teleNP” refers to traditional, face-to-face neuropsychological assessments that have been adapted to either a telephone-based or video conference-based format; other forms of audiovisual-aided neuropsychological assessment (e.g., computerized testing via specialized software packages or web-based platforms) were beyond the scope of this report.

Empirical evidence related to telephone-based neuropsychological evaluations

Telephone-based neuropsychological assessments characterize some of the earliest iterations of teleNP (Brandt & Folstein,

Corresponding author: Theresa F. Gierzynski; Email: gierzyns@med.umich.edu

Cite this article: Gierzynski T.F., Gregoire A., Reader J.M., Pantis R., Campbell S., Bhaumik A., Rahman-Filipiak A., Heidebrink J., Giordani B., Paulson H., & Hampstead B.M. (2024) Evaluation of the Uniform Data Set version 3 teleneuropsychological measures. *Journal of the International Neuropsychological Society*, 30: 183–193, <https://doi.org/10.1017/S1355617723000383>

Copyright © INS. Published by Cambridge University Press 2023. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted re-use, distribution, and reproduction in any medium, provided the original work is properly cited.

1988). This format may serve as a particularly useful medium for teleNP administration, given the widespread availability of telephones and the general ease of operating such devices. Cognitive screeners, such as the Telephone Interview for Cognitive Status (TICS) (Brandt & Folstein, 1988) and its modified version (TICS-M) (Welsh et al., 1993), are among the most extensively validated and widely used telephone-based instruments to date (Carlew et al., 2020; Castanho et al., 2014; Hunter et al., 2021). Additionally, the comparability of telephone and in-person assessments is well-supported for most verbally administered tasks (e.g., verbal memory and language measures) (Carlew et al., 2020). For instance, Bunker et al. (2017) administered a battery of neuropsychological tests to a sample of older adults (mean age = 74.9) in-person and then subsequently via telephone 2–4 weeks later; the investigators found that mean scores obtained through in-person and telephone testing were strongly correlated for several measures, including the Hopkins Verbal Learning Test-Revised (HVLT-R) Total Recall ($r = 0.87$), Verbal Fluency ($r = 0.92$), and the Boston Naming Test 15-item (BNT-15) ($r = 0.85$) (note: the BNT-15 was heavily revised by the researchers to be compatible with verbal administration over telephone). However, the evidence supporting the use of visuospatial measures via telephone is exceedingly small (Thompson et al., 2001), and few telephone-based instruments are available which evaluate processing speed and executive functioning (Carlew et al., 2020). Interestingly, whereas early video-based teleNP studies were most frequently conducted onsite at a clinical or research setting (i.e., optimal conditions conducive to high standardization and experimental control), a vast majority of telephone-based investigations involve testing administered to participants directly in their homes (Carlew et al., 2020).

Empirical evidence related to video-based neuropsychological evaluations

Existing research evaluating video-based teleNP relative to traditional, in-person neuropsychological (NP) evaluation has thus far been encouraging, albeit under highly controlled conditions (Brearly et al., 2017; Cullum et al., 2006; Cullum et al., 2014; Grosch et al., 2015; Hildebrand et al., 2004; Marra, Hamlet, et al., 2020a; Wadsworth et al., 2016, 2018). For example, an early meta-analysis of 12 studies revealed that testing modality has a minor, nonstatistically significant influence on performance, with verbally based measures showing the strongest reliability (Brearly et al., 2017). Interestingly, however, this meta-analysis revealed higher (33%), lower (61%), and equivalent (6%) mean test scores for video relative to in-person NP evaluations (Brearly et al., 2017). A relatively recent systematic review comparing video-based and in-person testing in older adults (aged 65 and older) supported teleNP-administered tests, particularly for cognitive screening tools (e.g., Montreal Cognitive Assessment, Mini-Mental State Examination) and tests measuring language, attention, and memory (Marra et al., 2020a). In addition, a large early investigation of video-based teleNP (Cullum et al., 2014) in older adults reported moderate to excellent reliability across the assessed measures (i.e., ICCs ranging from 0.55 to 0.91). Importantly, however, most of these early studies examined a limited number of tests and administered video-based testing in well-controlled environments (e.g., via video in an office next to the examiner) that may not accurately reflect the less predictable nature of in-home teleNP.

There is limited research evaluating in-home, video-administered teleNP when applied to older adult populations

(Abdolahi et al., 2016; Alegret et al., 2021; Fox-Fuller et al., 2022; Lindauer et al., 2017; Parks et al., 2021; Stillerova et al., 2016), with the home environment potentially being more susceptible to confounding factors, such as ambient noise (e.g., visitors or delivery persons ringing the doorbell), lapses in internet connectivity, poor audio or visual quality, people or pets entering the testing area, etc. Only three such studies were available in early 2020 at the start of the COVID-19 pandemic (Abdolahi et al., 2016; Lindauer et al., 2017; Stillerova et al., 2016), all of which focused on the video conference administration of the Montreal Cognitive Assessment (MoCA) and were associated with moderate to strong agreement across modalities.

Thus, there was a paucity of empirical data available when the University of Michigan temporarily halted in-person human subject research to comply with state and federal COVID-19 lock-downs. By necessity, our Michigan Alzheimer's Disease Research Center's (MADRC) longitudinal study of memory and aging was forced to shift to a virtual format. Shortly thereafter, the National Alzheimer's Coordinating Center (NACC) disseminated a revised version of the Uniform Data Set version 3 (UDS v3.0) cognitive test battery to the ADRC network. This UDS v3.0 Telephone Cognitive Battery (known as UDS v3.0 t-cog) preserved many of the core tests from the in-person battery, which we augmented with additional verbal measures (i.e., C Letter Fluency, Hopkins-Verbal Learning Test-Revised) and UDS 3.0 tests that involved the presentation of visual stimuli (i.e., Benson Complex Figure, Multilingual Naming Test). This paper is the first, to our knowledge, to evaluate the reliability of the UDS v3.0 t-cog test battery (and additional tests from our local protocol) in a clinically mixed sample of 181 older adults. In exploratory analyses, we also assessed reliability estimates according to either video-based or telephone-based testing modality. Our primary goal was to describe the reliability of the UDS v3.0 t-cog measures in this real-world teleNP situation through an examination of intraclass correlation coefficients (ICCs).

Method

Participants

All study procedures – which were reviewed and approved by the University of Michigan Medical School Review Board (IRB MED) – adhered to the ethical standards outlined in the Helsinki Declaration. A total of 210 participants provided written informed consent at each time point and completed both an in-person UDS v3.0 assessment before March 12th, 2020 (when COVID-19 restrictions were implemented) and the next subsequent evaluation using the UDS v3.0 t-cog. Given the new remote testing format, a separate virtual meeting was held with the participant prior to virtual testing to obtain informed consent via SignNow, a secure, electronic signature platform supported by the University of Michigan. If participants were unable to navigate the SignNow interface, a physical copy of the informed consent document was mailed to the participant and reviewed during a telephone call. Participants were then instructed to sign and return their consent form via postal mail, using a pre-addressed, pre-stamped envelope that had been provided. Of the 210 total participants, 25 cases were deemed potentially invalid and were not included in our data analysis. Threats to validity were documented for each of these excluded assessments, with several cases (36%) citing multiple potential confounds. Reasons for exclusion included hearing impairment (9/25), technological issues (9/25), distractions/interruptions in the home (5/25), note-taking/“cheating” by participant (3/25), unapproved assistance from others in the

Table 1. Participant demographic characteristics at most recent UDS v3.0 evaluation

| Variable | TeleNP participants (N = 181) | Consecutive in-person participants (N = 276) |
|--------------------------------------|----------------------------------|---|
| Sex, N (%) | | |
| Female | 121 (66.9%) | 190 (68.8%) |
| Male | 60 (33.1%) | 86 (31.2%) |
| Race, N (%) | | |
| White | 98 (54.1%) | 154 (55.8%) |
| Black or African American | 70 (38.7%) | 98 (35.5%) |
| Asian | 0 (0%) | 1 (0.4%) |
| Other | 0 (0%) | 1 (0.4%) |
| N/A | 13 (7.2%) | 22 (8%) |
| Ethnicity, N (%) | | |
| Non-Hispanic | 167 (92.3%) | 254 (92%) |
| Hispanic | 1 (0.6%) | 0 (0%) |
| N/A | 13 (7.2%) | 22 (8%) |
| Diagnostic group, N (%) | | |
| Cognitively unimpaired | 120 (66.3%) | 141 (51.1%) |
| Mild cognitive impairment | 50 (27.6%) | 91 (33.0%) |
| Dementia | 11 (6.1%) | 44 (15.9%) |
| Level of education, years, mean (SD) | 16.3 (2.5) (range 12–20) | 15.9 (2.5) (range 8–20) |
| Age, years, mean (SD) | 71.9 (6.8) (range 52.3–93.9) | 72.1 (7.6) (range 51.1–92.9) |

Abbreviations: SD, Standard Deviation; TeleNP, teleneuropsychology; UDS, Uniform Data Set.

home (2/25), lack of effort or interest (3/25), fatigue (2/25), and emotional issues (2/25). Two cases that used a hybrid testing modality (i.e., a combination of telephone and video) were removed and two other cases with an “impaired, not MCI” research diagnosis (i.e., participants with objectively impaired performance on neuropsychological testing but without subjective cognitive complaints or evidence of functional decline) were also excluded due to the small group size. These steps resulted in a total of 181 participants who were included in our final reported data analysis.

The UDS v3.0 t-cog was administered either via video conference ($n = 122$) or telephone ($n = 59$) with an average of 16 months between evaluations (mean days = 479.2; SD = 122.0 days; range = 320–986 days). All participants were English-speaking adults aged 52 years and older. Exclusionary criteria included history of non-neurodegenerative neurologic injury or disease, such as moderate-severe traumatic brain injury, stroke, or epilepsy, or a history of central nervous system radiation therapy, or developmental delays. Those with significant psychiatric diagnoses (e.g., Bipolar Disorder, Schizophrenia, moderate-severe Major Depressive Disorder) or active substance abuse/dependence were also excluded.

The sample was predominantly female (66.9%) and mostly college educated ($M = 16.3$ years of education; $SD = 2.5$; range = 12–20 years). Mean age was 71.9 ($SD = 6.8$; range = 52.3–93.9). Self-reported race was 54.1% “White” and 38.7% “Black or African American” (see Table 1 for complete demographic characteristics). Participants held a consensus diagnosis of cognitively unimpaired ($n = 120$), mild cognitive impairment (MCI; $n = 50$), or dementia ($n = 11$) following the in-person evaluation and were re-diagnosed following the remote visit. Diagnoses were rendered via consensus conference that included neurologists, neuropsychologists, nurses, social workers, and other relevant specialists according to NACC guidelines (National Alzheimer’s Coordinating Center, 2015).

Table 2. Comparison of in-person, video, and telephone test batteries

| Neuropsychological test | Cognitive domain | IP, N = 181 | Video, N = 122 | Telephone, N = 59 |
|---|-----------------------|----------------|-------------------|----------------------|
| MoCA | Global functioning | × | | |
| Blind/telephone MoCA | Global functioning | | × | × |
| Craft Story 21 recall | Learning and memory | × | × | × |
| Benson complex figure | Visuospatial | × | × | |
| Number span forward | Attention | × | × | × |
| Number span backward | Attention | × | × | × |
| Category fluency (animals & vegetables) | Language | × | × | × |
| Trail making test A ^a | Processing speed | × | | |
| Trail making test B ^a | Executive functioning | × | | |
| Oral trail making test A ^a | Processing speed | | × | × |
| Oral trail making test B ^a | Executive functioning | | × | × |
| MINT ^a | Language | × | × | |
| Letter fluency (C, F, & L) | Language | × | × | × |
| Verbal naming test ^a | Language | | | × |
| HVLT-R | Learning and memory | × | × | × |

Abbreviations: HVLT-R, Hopkins Verbal Learning Test-Revised; IP, In-Person; MoCA, Montreal Cognitive Assessment; MINT, Multilingual Naming Test.

^aProportion correct calculated for MINT and Verbal Naming Test given the different scales. Likewise, TMT B/A ratios were calculated for oral and written trails to ensure comparable metrics.

Procedures

Neuropsychological test battery

Table 2 lists the tests used for each assessment type (i.e., in-person, video, and telephone). No alternate forms were used, as NACC does not provide alternative tests for the UDS v3.0. Measures used in all formats included: the Montreal Cognitive Assessment (MoCA), Craft Story 21, Number Span Forward and Backward, Category Fluency (Animals and Vegetables), Letter Fluency (C, F, and L), the Hopkins Verbal Learning Test-Revised (HVLT-R), and the Trail Making Test A and B (note that C Letter Fluency and the HVLT-R were part of our “local” protocol and are not included in the UDS 3.0). Importantly, the video and telephone evaluations used the oral version of the Trail Making Test A and B as well as the Blind/Telephone MoCA. Blind/Telephone MoCA scores were converted to the traditional MoCA scale using the formula provided on the test publisher’s website (Nasreddine, 2022). We included both the Benson Complex Figure and the Multilingual Naming Test (MINT) during the video visits even though these measures were not included in the UDS v3.0 t-cog battery. For the Benson Complex Figure, examiners shared a digital version of the image via screenshare and asked participants to copy the image following standard (i.e., in-person) instructions. Once completed, participants held their figure in front of the webcam and the examiner saved a screenshot for subsequent scoring. Participants were then instructed to fold the piece of paper in half with the image on the inside, take it in their left hand, and place it on the floor. This three-step command effectively removed the Benson drawing from view while concurrently evaluating the participant’s ability to follow a multi-step command. Following the delay, the participant was asked to draw the figure and another screenshot was captured and later scored. At the end of each session, participants were instructed to dispose of their Benson Figure

drawings in the trash; this was intended to protect test security and to help prevent any unapproved inspection/reproduction of test stimuli. To evaluate confrontation naming, we showed digital images of the MINT stimuli to participants and recorded their responses following standard procedures. Since visually based stimuli could not be administered during the telephone-based sessions, we used the Verbal Naming Test (VNT) instead of the MINT (per NACC guidance) and omitted the Benson Complex Figure.

UDS v3.0 t-cog set-up

Participants completed cognitive testing from their homes using a personally owned telephone (for telephone assessments) or computing device (for video assessments). For video-administered testing, we were unable to standardize the nature of the device or screen size, given pandemic related restrictions; as such, participants were permitted to use any internet-enabled device (e.g., desktop computers, laptops, tablets, and smartphones). Examiners conducted testing from either the MADRC office space or from their homes using a University of Michigan computer and virtual private network (VPN). Technology was consistent across all examiners; the examiner set-up included a desktop computer, dual monitors, a headset with a built-in microphone, and a webcam. The UDS v3.0 t-cog battery was administered by secured video conference ($n = 122$) or telephone ($n = 59$) using either the “BlueJeans” or “Zoom for Health” telecommunication platforms. For video assessments, an identical, nondescript virtual background (i.e., an image of an empty room with a white wall and wooden floor) was used by all examiners. The test examiner asked participants to power down all electronic devices and remain in a quiet place where they would not be disturbed for approximately 90 minutes. Participants were explicitly instructed to complete the testing session by themselves and were reminded that they were not allowed to take notes or receive assistance from others in the home while completing their evaluation. Another person was permitted to set-up the telephone or video call if the participant was unable to do so on their own (National Alzheimer’s Coordinating Center, 2020); the individual providing assistance was then asked to leave the room immediately in order for testing to commence. At the start of each session, examiners performed an initial check of connection quality by ensuring participants could adequately hear and see the examiner and that the audio and visual connections were not “dropping” during conversation. Participants were also reminded to use sensory aids (e.g., hearing aids, eyeglasses), if they normally used such aids. Any factors that may have influenced the validity of a neuropsychological measure (e.g., note-taking, significant disruptions to internet connectivity) were recorded by the examiner and discussed with the larger team before deciding whether to exclude (see above). To enhance comparability of measurement, we converted MINT and VNT scores to percent correct and evaluated both raw (i.e., time to completion in seconds) and ratio (i.e., B/A ratios) for the Trail Making Tests (written for in-person; oral for UDS v3.0 t-cog).

Statistical methods

Except when otherwise noted, all analyses used raw scores. We used ICCs and 95% confidence intervals (CI) to estimate test-retest reliability across neuropsychological measures. ICC figures were interpreted according to established thresholds (Koo & Li, 2016): values ≤ 0.50 indicate “poor” reliability, between 0.50 and 0.75 suggest “moderate” reliability, between 0.75 and 0.90 suggest

“good” reliability, and ≥ 0.90 imply “excellent” reliability. Significance of ICC measurements tested the null hypothesis that $ICC = 0$ and are represented by the 95% CIs.

To frame the primary results more accurately, we calculated comparable ICCs under two control conditions: (1) restricting analyses to only those who remained diagnostically stable across the two assessment points (Table 4) ($n = 158$) and (2) between two consecutive in-person UDS v3.0 evaluations that both occurred prior to the COVID-19 pandemic ($n = 276$; mean time between visits = 398.9 days; $SD = 88.1$; range: 188–880 days) (Table 5). For the repeat in-person control analysis, participants were selected from our longitudinal cohort of older adults who had completed two in-person assessments on or before March 11th, 2020; these analyses included data associated with the participants’ two most recent evaluations. Demographic characteristics associated with the in-person to in-person sample were similar to the primary sample [mean age = 72.1 years ($SD = 7.6$; range = 51.1–92.9); mean years of education = 15.9 ($SD = 2.5$; range = 8–20); 68.8% females; 55.8% White; 35.5% Black or African American]. Of the 276 cases compared in the repeat in-person sample, 141 were cognitively unimpaired; those with cognitive impairment held consensus research diagnoses of Amnesic MCI ($n = 61$), non-Amnesic MCI ($n = 30$), dementia of the Alzheimer’s type ($n = 40$), and mixed dementia ($n = 4$). Notably, this group had a higher proportion of dementia cases (15.9%) relative to the overall sample (6.1%) (Table 1).

Results

Overall sample comparing in-person with UDS v3.0 t-cog

Overall ICCs ranged from 0.01 to 0.79 across the tests of the 181 individuals included in the analysis (Table 3; Figure 1). ICCs fell in poor (15%), moderate (70%), and good (15%) agreement ranges. We found the strongest ICCs (i.e., “good”) for Craft Story Recall – Delayed Verbatim (ICC = 0.79) and Paraphrase (ICC = 0.77), and the Benson Complex Figure – Delayed (ICC = 0.79 during the video assessments). Conversely, the lowest ICCs were observed for the Trail Making Test-A/Oral Trail Making Test-A (TMT-A/OTMT-A) (ICC = 0.01), Trail Making Test-B/Oral Trail Making Test-B (TMT-B/OTMT-B) (ICC = 0.21), and TMT B/A Ratio (ICC = 0.11) (Table 3). This general pattern of results was evident when considering the video-based and telephone-based sessions separately, though it should be noted that four ICCs were relatively lower for telephone than for video-based sessions (Table 3: Number Span Forward, Number Span Backward, Category Fluency – Animals, and TMT B/A Ratio).

Additional analyses

Our additional control analyses revealed two primary findings: (1) results were largely unchanged when limiting our analyses to those who remained diagnostically stable across these two time points (ICC Range: 0–0.78) (Table 4; Figure 1) and (2) ICCs were higher (ICC Range: 0.35–0.87) between consecutive in-person evaluations that occurred on or before March 11th, 2020 (Table 5; Figure 1). Importantly, the mean number of prior evaluations (i.e., those which occurred before the two visits included in the data analysis) was similar for our primary analysis (mean number = 0.9558; $SD = 0.74$; median = 1; range = 0–2) and the in-person to in-person control sample (mean number = 0.38 evaluations; $SD = 0.49$; median = 0; range = 0–2). Of the 181 participants in

Table 3. Average test scores by visit type and test-retest reliability between in-person and remote neuropsychological evaluations both overall ($N = 181$) and stratified by remote visit modality. Approximately 16 months elapsed between evaluations (mean days = 479.2; $SD = 122.0$ days; range = 320–986 days)

| Variable | Number of Obs | In-person, Mean (SD) ^c | Remote, Mean (SD) | Overall, $N = 181$ | | Video only visits, $N = 122$ | | Telephone only visits, $N = 59$ | |
|--|---------------|-----------------------------------|-------------------|--------------------|--------------------|------------------------------|--------------------|---------------------------------|---------------------|
| | | | | Number of Obs | ICC (95% CI) | Number of Obs | ICC (95% CI) | Number of Obs | ICC (95% CI) |
| Craft Story 21 Recall – Immediate Verbatim | 360 | 22.1 (6.2) | 23.5 (7) | 360 | 0.7 (0.6, 0.78) | 242 | 0.71 (0.6, 0.79) | 118 | 0.67 (0.49, 0.8) |
| Craft Story 21 Recall – Immediate Paraphrase | 360 | 15.8 (3.9) | 16.6 (4.3) | 360 | 0.69 (0.59, 0.76) | 242 | 0.7 (0.6, 0.78) | 118 | 0.65 (0.46, 0.78) |
| Craft Story 21 Recall – Delayed Verbatim | 360 | 18.4 (7.4) | 19.9 (8.3) | 360 | 0.79 (0.71, 0.84) | 242 | 0.79 (0.7, 0.85) | 118 | 0.79 (0.64, 0.87) |
| Craft Story 21 Recall – Delayed Paraphrase | 360 | 13.9 (5.1) | 15.1 (5.6) | 360 | 0.77 (0.68, 0.84) | 242 | 0.77 (0.68, 0.84) | 118 | 0.78 (0.58, 0.88) |
| Number Span Forward | 360 | 8.1 (2.2) | 7.8 (2.5) | 360 | 0.63 (0.53, 0.71) | 242 | 0.68 (0.58, 0.77) | 118 | 0.49 (0.27, 0.66) |
| Number Span Backward | 360 | 6.9 (2) | 7.3 (2.3) | 360 | 0.53 (0.41, 0.63) | 242 | 0.61 (0.48, 0.71) | 118 | 0.35 (0.11, 0.55) |
| MoCA/Blind MoCA | 348 | 25.7 (3.1) | 24.3 (4.4) | 348 | 0.64 (0.48, 0.75) | 234 | 0.66 (0.46, 0.78) | 114 | 0.6 (0.39, 0.75) |
| Category Fluency – Animals | 360 | 20.3 (5.3) | 20 (5.6) | 360 | 0.58 (0.47, 0.67) | 242 | 0.62 (0.5, 0.72) | 118 | 0.45 (0.22, 0.63) |
| Category Fluency – Vegetables | 360 | 14.3 (4.2) | 14 (4.4) | 360 | 0.71 (0.62, 0.77) | 242 | 0.7 (0.6, 0.78) | 118 | 0.7 (0.54, 0.81) |
| HVLT-R Total Recall | 346 | 24.8 (6) | 26.4 (6.2) | 346 | 0.65 (0.53, 0.74) | 236 | 0.66 (0.52, 0.76) | 110 | 0.62 (0.42, 0.76) |
| HVLT-R Delayed Recall | 346 | 8.5 (3.5) | 8 (4.1) | 346 | 0.65 (0.56, 0.73) | 236 | 0.66 (0.54, 0.75) | 110 | 0.63 (0.45, 0.77) |
| HVLT-R Retention | 346 | 83.3 (30) | 75.7 (36.1) | 346 | 0.56 (0.44, 0.65) | 236 | 0.54 (0.4, 0.66) | 110 | 0.59 (0.38, 0.74) |
| HVLT-R Recognition | 344 | 9.8 (2.3) | 10 (2.3) | 344 | 0.64 (0.54, 0.72) | 234 | 0.6 (0.47, 0.7) | 110 | 0.71 (0.55, 0.82) |
| Letter Fluency (C, F, & L) | 360 | 45.8 (10.7) | 43.3 (10.7) | 360 | 0.74 (0.64, 0.81) | 242 | 0.73 (0.61, 0.81) | 118 | 0.73 (0.57, 0.83) |
| TMT-A/OTMT-A | 358 | 33.9 (14.7) | 10.1 (3.7) | 358 | 0.01 (−0.03, 0.05) | 242 | 0.01 (−0.05, 0.08) | 116 | −0.01 (−0.05, 0.05) |
| TMT-B/OTMT-B | 342 | 88.2 (44.1) | 43.1 (28) | 342 | 0.21 (−0.06, 0.44) | 234 | 0.21 (−0.06, 0.44) | 108 | 0.19 (−0.08, 0.45) |
| TMT B/A Ratio ^a | 342 | 2.8 (1.2) | 4.6 (3.4) | 342 | 0.11 (−0.02, 0.24) | 234 | 0.19 (0.01, 0.36) | 108 | 0.02 (−0.16, 0.23) |
| MINT/VNT ^a (Proportion correct) | 354 | 0.9 (0.1) | 0.9 (0.1) | 354 | 0.71 (0.63, 0.78) | – | – | 118 | 0.66 (0.48, 0.78) |
| MINT ^b | 236 | 30.4 (1.8) | 30.1 (2.2) | – | – | 236 | 0.73 (0.63, 0.8) | – | – |
| Benson Complex Figure – Copy ^b | 234 | 15.2 (1.7) | 15 (1.6) | – | – | 234 | 0.53 (0.39, 0.65) | – | – |
| Benson Complex Figure – Delayed ^b | 234 | 10.9 (3.8) | 11.1 (3.7) | – | – | 234 | 0.79 (0.71, 0.85) | – | – |

Abbreviations: HVLT-R, Hopkins Verbal Learning Test-Revised; ICC, Intraclass Correlation Coefficient; MINT, Multilingual Naming Test; MoCA, Montreal Cognitive Assessment; Obs, Observations; OTMT, Oral Trail Making Test; SD, Standard Deviation; TMT, Trail Making Test; VNT, Verbal Naming Test.

^aProportion correct calculated for MINT and Verbal Naming Test given the different scales. Likewise, TMT B/A ratios were calculated for oral and written trails to ensure comparable metrics.

^bVideo-only analyses.

^cMeans and SD for the combined (video and telephone remote visits) overall sample except where indicated as video only analyses.

the primary in-person/virtual cohort, 125 were also in the in-person to in-person sample (69.1% overlap across groups).

ICCs by diagnostic group

Exploratory diagnosis-specific results were limited by relatively small sample sizes for those with cognitive impairment (i.e., MCI and dementia) but revealed notable differences across diagnostic groups (Supplemental Tables 1 and 2). Specifically, cognitively unimpaired participants showed poor ICCs for HVLT-R Delayed Recall (ICC = 0.2) and HVLT-R Retention (ICC = 0.14), as well as TMT-A/OTMT-A (ICC = −0.01), TMT-B/OTMT-B (ICC = 0.19), and TMT B/A Ratio (ICC = 0.13). Symptomatic participants showed poor ICCs for Number Span Forward (ICC = 0.31), Number Span Backward (ICC = 0.39), TMT-A/OTMT-A (ICC = 0), TMT-B/OTMT-B (ICC = 0.15), and TMT B/A Ratio (ICC = 0.03).

Discussion

The COVID-19 pandemic necessitated accessible neuropsychological assessments capable of reaching individuals outside of traditional research and clinical settings. Growing evidence suggests that both telephone and video administered teleNP may serve as a viable alternative to traditional, in-person assessment (Brearly et al., 2017; Carlew et al., 2020; Marra,

Hamlet, et al., 2020); however, the psychometric properties associated with teleNP when administered directly to the home remains an understudied area of research, particularly for video-based neuropsychological evaluations. This investigation is the first, to our knowledge, to evaluate the reliability of the UDS v3.0 t-cog test battery, as well as additional measures from our local study protocol. In general, our results are encouraging and suggest mostly moderate to good agreement between in-person and teleNP testing conditions (overall ICC Range = 0.01–0.79; ICC Range = 0.53–0.79, if excluding TMT/OTMT) (Table 3). Although our reliability estimates are, in some cases, less robust than prior teleNP investigations (see Cullum et al., 2014 as an example), this may be partially explained by a lengthier testing interval than has typically been reported in other studies (Brearly et al., 2017; Hunter et al., 2021) – a factor that was outside of our control given the pandemic. Additionally, the variability in scores observed across assessments might be reasonably attributed to a certain degree of expected change within aging populations. For perspective, Webb et al. (2022) conducted test-retest analyses in a large sample ($n = 16,956$) of older adults (age ≥ 65 years) who completed a series of cognitive tests in-person (i.e., the Modified Mini-Mental State, Symbol Digit Modalities Test, Hopkins Verbal Learning Test-Revised, and Controlled Oral Word Association Test) at baseline and at one-year follow-up; results were associated with ICCs in the moderate to good range (ICC Range = 0.53–0.77) and

provide a useful point of comparison with our study. Our findings were not driven by clinical conversion/reversion given our first control analyses that revealed comparable ICCs in a diagnostically stable subgroup (ICC Range = 0–0.78) (Table 4). Our second set of control analyses revealed relatively higher ICCs in comparably timed, repeat in-person evaluations using the same neuropsychological measures (ICC Range = 0.35–0.87) (Table 5). These latter differences cannot be accounted for by prior experience or practice effects since our samples had a comparable number of prior evaluations. Thus, there appears to be some relative loss of reliability when shifting from in person to virtual, though we cannot comment on the clinical or research ramifications of this difference.

Our findings suggest that the general field of neuropsychology can have confidence in several UDS v3.0 measures when administered virtually: specifically, the Craft Story Recall – Delayed Paraphrase and Verbatim, Letter Fluency (C, F, and L), MINT, and the Benson Complex Figure – Delayed. The strong reliability estimates associated with Craft Story and Letter Fluency are consistent with prior research that has supported the cross-modal comparability of verbally mediated tasks across traditional, in-person and remote (i.e., telephone or video-based) testing conditions (Brearly et al., 2017; Carlew et al., 2020; Hunter et al., 2021). The latter two measures (i.e., MINT and Benson Complex Figure) are important to note since they were not included in the NACC UDS 3.0 t-cog test battery. The relatively strong ICCs observed for the Benson Complex Figure – Delayed are important for the field given the relative paucity of teleNP measures that assess visuospatial functioning (Brearly et al., 2017; Carlew et al., 2020). These findings suggest our approach may be viable for other measures of visuospatial functioning and visuoperception and visuoconstruction. The MINT was moderately reliable when comparing video-based and in-person administrations (ICC = 0.73); reliability was also moderate when comparing in-person MINT scores with scores obtained via telephone on the Verbal Naming Test (ICC = 0.71; ICC calculated using the percentage of correct responses to account for different scales on MINT/VNT). Overall, our results are slightly less favorable than past studies that have compared in-person and video-based administrations of the Boston Naming Test-15 item short form (BNT-15) – a confrontation naming task similar to the MINT – which has previously been associated with ICCs of 0.81 (Cullum et al., 2014), 0.87 (Cullum et al., 2006), and 0.93 (Wadsworth et al., 2016). Notably, participants in these prior studies completed both video and in-person evaluations on the same day, whereas our testing interval was far longer (i.e., approximately 16 months). As such, it is unsurprising that our lengthy retest interval resulted in comparatively lower ICCs.

The HVLT-R and its subtests revealed moderately strong ICCs, ranging from 0.56 to 0.65 in our overall remote sample (ICC Range = 0.54–0.66 for video-based evaluations; ICC Range = 0.59–0.71 for telephone-based evaluations). This pattern is consistent with, albeit somewhat weaker than, past studies using video HVLT-R administration (ICCs of 0.77–0.88) (Cullum et al., 2006, 2014; Wadsworth et al., 2016, 2018). Interestingly, Bunker and colleagues (2017) found HVLT-R correlation coefficients of $r = 0.27–0.87$ for in-person versus telephone-based administration. Our data are consonant with those of Bunker et al. (2017) as both studies observed the weakest relationships with the HVLT-R percent retention scores, so some degree of caution may be warranted when interpreting this measure. We again suspect that our longer test–retest interval played a role in these findings but do not believe it fully accounts for them given the stronger

ICCs for the subsequent in-person evaluations (ICCs of 0.56–0.76; Table 5).

Our results warrant caution when using the OTMT-A or B instead of the written versions of these measures based on ICCs that fell in the poor reliability range. This conclusion was somewhat anticipated, as the raw scores for each task are known to vary considerably with one another (i.e., higher raw values are expected on the in-person, written TMT relative to the oral analog of the task) (Ricker & Axelrod, 1994). To account for these differences, we calculated an ICC using the TMT B/A ratio, although this produced a similarly weak correlation between in-person and remote testing conditions (ICC = 0.11). The original OTMT validity study (Ricker & Axelrod, 1994) found strong Pearson's correlation coefficients for OTMT-A/TMT-A ($r = 0.68$) and OTMT-B/TMT-B ($r = 0.72$). However, a more recent study (Mrazik et al., 2010) revealed weaker correlations between OTMT-A/TMT-A ($r = 0.29$) and OTMT-B/TMT-B ($r = 0.62$). Another investigation reported that OTMT-A failed to distinguish between cognitively healthy and cognitively impaired participants due to a compressed range of scores with little variability (Bastug et al., 2013). Thus, there is a consensus across studies (Bastug et al., 2013; Kaemmerer & Riordan, 2016; Mrazik et al., 2010) that OTMT-A may not be an adequate substitute for TMT-A due to fundamental differences in task design: the OTMT-A places fewer cognitive demands on the participant (e.g., does not involve visual scanning or effortful number sequencing) and may elicit an over-learned, rote response. Overall, with respect to the OTMT, our findings align with past studies showing questionable agreement between the OTMT and TMT and suggest that users should be cautious when using OTMT for diagnostic purposes. Future studies should evaluate whether lack of agreement between OTMT and TMT arise from the solely verbal nature and/or the teleNP platform.

Strengths and limitations

As with all studies, several limitations exist that were largely due to the unanticipated COVID-19 pandemic. First, the significant lapse in time between the in-person and UDS v3.0 t-cog testing sessions (i.e., on average 16 months, but in some cases, as great as three years) likely weakened test–retest reliability estimates. However, our control analyses revealed these patterns were not due to diagnostic conversion/reversion (Table 4) and were instead related to the cross-modal assessment since comparably timed in-person evaluations had relatively higher ICCs (Table 5). While our total sample ($n = 181$) rivals the largest pre-COVID-19 investigation of teleNP ($n = 202$) (Cullum et al., 2014), our study was more heavily weighted toward cognitively unimpaired participants, so it was surprising that ICCs on some measures (e.g., HVLT-R) were notably below those for cognitively impaired participants. This is an unexpected finding that warrants replication as we anticipated patient populations showing greater variability. We again emphasize that ICCs were not driven by diagnostic change and that they were relatively stronger for consecutive in-person visits. Another notable limitation relates to our well-educated sample ($M = 16.3$ years of education), which potentially limits generalizability of our findings. Additionally, pandemic-related restrictions rendered us unable to standardize participants' testing equipment and technological set-up (e.g., computing device type, audiovisual quality, internet connection speed). We cannot rule out the possibility that variability in participant technology influenced our results (e.g., perhaps worse performance associated with smaller screen size associated with tablets or smartphones),

Table 4. Average test scores by visit type and test-retest reliability between in-person and remote neuropsychological evaluations both overall ($N = 158$) and stratified by remote visit modality among individuals with static diagnoses from time 1 to time 2

| Variable | Number of Obs | In-Person, Mean (SD) ^c | Remote, Mean (SD) ^c | Overall ($N = 158$) | | Video ($N = 105$) | | Telephone ($N = 53$) | |
|--|------------------|--------------------------------------|-----------------------------------|-----------------------|--------------------|---------------------|-------------------|------------------------|---------------------|
| | | | | Number of Obs | ICC (95% CI) | Number of Obs | ICC (95% CI) | Number of Obs | ICC (95% CI) |
| Craft Story 21 Recall – Immediate Verbatim | 314 | 22.3 (6.4) | 23.9 (7.2) | 314 | 0.71 (0.61, 0.79) | 208 | 0.72 (0.6, 0.8) | 106 | 0.69 (0.49, 0.82) |
| Craft Story 21 Recall – Immediate Paraphrase | 314 | 15.9 (4) | 16.9 (4.3) | 314 | 0.7 (0.59, 0.78) | 208 | 0.71 (0.6, 0.8) | 106 | 0.67 (0.44, 0.81) |
| Craft Story 21 Recall – Delayed Verbatim | 314 | 18.9 (7.1) | 20.5 (8.1) | 314 | 0.78 (0.69, 0.84) | 208 | 0.79 (0.69, 0.85) | 106 | 0.75 (0.56, 0.85) |
| Craft Story 21 Recall – Delayed Paraphrase | 314 | 14.3 (4.8) | 15.5 (5.3) | 314 | 0.76 (0.66, 0.83) | 208 | 0.78 (0.68, 0.85) | 106 | 0.73 (0.46, 0.86) |
| Number Span Forward | 314 | 8 (2.2) | 7.8 (2.5) | 314 | 0.63 (0.52, 0.71) | 208 | 0.66 (0.54, 0.76) | 106 | 0.53 (0.3, 0.7) |
| Number Span Backward | 314 | 6.8 (2.1) | 7.3 (2.3) | 314 | 0.54 (0.41, 0.64) | 208 | 0.61 (0.48, 0.72) | 106 | 0.35 (0.1, 0.56) |
| MoCA/Blind MoCA | 306 | 25.8 (3.2) | 24.6 (4.4) | 306 | 0.66 (0.51, 0.77) | 202 | 0.67 (0.48, 0.79) | 104 | 0.64 (0.44, 0.78) |
| Category Fluency – Animals | 314 | 20.3 (5.3) | 20.2 (5.5) | 314 | 0.63 (0.52, 0.71) | 208 | 0.65 (0.52, 0.74) | 106 | 0.51 (0.28, 0.68) |
| Category Fluency – Vegetables | 314 | 14.5 (4.3) | 14.4 (4.5) | 314 | 0.71 (0.62, 0.78) | 208 | 0.69 (0.58, 0.78) | 106 | 0.72 (0.57, 0.83) |
| HVLT-R Total Recall | 302 | 25.1 (6.2) | 26.8 (6.3) | 302 | 0.68 (0.56, 0.77) | 202 | 0.69 (0.54, 0.79) | 100 | 0.65 (0.45, 0.79) |
| HVLT-R Delayed Recall | 302 | 8.8 (3.3) | 8.6 (3.7) | 302 | 0.69 (0.59, 0.76) | 202 | 0.74 (0.64, 0.82) | 100 | 0.57 (0.35, 0.73) |
| HVLT-R Retention | 302 | 85.6 (27.9) | 80.8 (32.3) | 302 | 0.56 (0.44, 0.66) | 202 | 0.61 (0.48, 0.72) | 100 | 0.5 (0.25, 0.68) |
| HVLT-R Recognition | 300 | 10 (2.1) | 10.1 (2.2) | 300 | 0.65 (0.55, 0.73) | 200 | 0.65 (0.52, 0.75) | 100 | 0.65 (0.45, 0.78) |
| Letter Fluency (C, F, & L) | 314 | 45.4 (10.3) | 43.2 (10.7) | 314 | 0.77 (0.67, 0.83) | 208 | 0.75 (0.64, 0.83) | 104 | 0.75 (0.6, 0.85) |
| TMT-A/OTMT-A | 312 | 33.7 (15.1) | 9.9 (3.5) | 312 | 0 (–0.05, 0.05) | 208 | 0 (–0.06, 0.08) | 106 | –0.01 (–0.05, 0.05) |
| TMT-B/OTMT-B | 296 | 87.9 (45.6) | 42.8 (28.5) | 296 | 0.22 (–0.06, 0.45) | 200 | 0.2 (–0.06, 0.43) | 104 | 0.2 (–0.08, 0.47) |
| TMT B/A Ratio ^a | 296 | 2.8 (1.3) | 4.7 (3.6) | 296 | 0.1 (–0.04, 0.24) | 200 | 0.19 (0.01, 0.37) | 96 | 0.01 (–0.19, 0.24) |
| MINT/VNT ^a (Proportion Correct) | 308 | 0.9 (0.1) | 0.9 (0.1) | 308 | 0.73 (0.64, 0.79) | – | – | 106 | 0.66 (0.48, 0.79) |
| MINT ^b | 202 | 30.4 (1.8) | 30.2 (2.2) | – | – | 202 | 0.75 (0.65, 0.82) | – | – |
| Benson Complex Figure Copy ^b | 200 | 15.2 (1.7) | 15.1 (1.5) | – | – | 200 | 0.61 (0.47, 0.72) | – | – |
| Benson Complex Figure – Delayed ^b | 200 | 11.1 (3.7) | 11.5 (3.6) | – | – | 200 | 0.8 (0.71, 0.86) | – | – |

Abbreviations: HVLT-R, Hopkins Verbal Learning Test-Revised; ICC, Intraclass Correlation Coefficient; MINT, Multilingual Naming Test; MoCA, Montreal Cognitive Assessment; Obs, Observations; OTMT, Oral Trail Making Test; SD, Standard Deviation; TMT, Trail Making Test; VNT, Verbal Naming Test.

^aProportion correct calculated for MINT and Verbal Naming Test given the different scales. Likewise, TMT B/A ratios were calculated for oral and written trails to ensure comparable metrics.

^bVideo only analyses.

^cMeans and SD for the combined (video and telephone remote visits) overall sample except where indicated as video only analyses.

Table 5. Average test scores by visit type and test-retest reliability between two consecutive in-person neuropsychological evaluations on or before March 11th, 2020, both overall ($N = 276$) and stratified by diagnostic group among those with a stable diagnosis at both timepoints ($n = 244$). Approximately 13 months elapsed between evaluations (mean days = 398.9; $SD = 88.1$ days; range = 188–880 days)

| Variable | Overall ($N = 276$) | | | | CU ($N = 130$) | | MCI ($N = 64$) | | Cog Imp ($N = 114$) | |
|--|-----------------------|-----------------------|-----------------------|-------------------|------------------|-------------------|------------------|-------------------|-----------------------|--------------------|
| | Number of Obs | Mean visit 1 (SD) | Mean visit 2 (SD) | ICC | Number of Obs | ICC (95% CI) | Number of Obs | ICC (95% CI) | Number of Obs | ICC (95% CI) |
| Craft Story 21 Recall – Immediate Verbatim | 496 | 20.1 (7.1) | 20.5 (7) | 0.73 (0.66, 0.78) | 256 | 0.61 (0.49, 0.71) | 126 | 0.36 (0.12, 0.56) | 176 | 0.61 (0.46, 0.73) |
| Craft Story 21 Recall – Immediate Paraphrase | 496 | 14.6 (4.6) | 14.7 (4.6) | 0.74 (0.68, 0.79) | 256 | 0.58 (0.46, 0.69) | 126 | 0.48 (0.26, 0.65) | 176 | 0.65 (0.51, 0.76) |
| Craft Story 21 Recall – Delayed Verbatim | 494 | 16.8 (7.6) | 17 (7.9) | 0.76 (0.7, 0.81) | 256 | 0.52 (0.38, 0.64) | 126 | 0.5 (0.29, 0.67) | 174 | 0.74 (0.63, 0.82) |
| Craft Story 21 Recall – Delayed Paraphrase | 494 | 12.9 (5.2) | 12.8 (5.5) | 0.8 (0.75, 0.84) | 256 | 0.58 (0.45, 0.68) | 126 | 0.51 (0.3, 0.67) | 174 | 0.76 (0.66, 0.84) |
| Number Span Forward | 496 | 7.8 (2.3) | 7.8 (2.2) | 0.74 (0.68, 0.79) | 256 | 0.77 (0.7, 0.84) | 126 | 0.69 (0.54, 0.8) | 176 | 0.62 (0.47, 0.73) |
| Number Span Backward | 496 | 6.3 (2.2) | 6.4 (2.1) | 0.68 (0.6, 0.74) | 256 | 0.67 (0.56, 0.75) | 126 | 0.45 (0.24, 0.63) | 176 | 0.52 (0.35, 0.66) |
| MoCA | 498 | 24.3 (4.4) | 24.4 (5) | 0.87 (0.84, 0.9) | 252 | 0.61 (0.49, 0.71) | 126 | 0.55 (0.35, 0.7) | 182 | 0.81 (0.73, 0.87) |
| Category Fluency – Animals | 496 | 19.3 (5.6) | 19.1 (5.9) | 0.72 (0.65, 0.77) | 256 | 0.58 (0.45, 0.69) | 126 | 0.58 (0.39, 0.72) | 176 | 0.65 (0.51, 0.76) |
| Category Fluency – Vegetables | 496 | 14 (4.3) | 13.5 (4.6) | 0.71 (0.64, 0.76) | 256 | 0.52 (0.38, 0.64) | 126 | 0.56 (0.36, 0.7) | 176 | 0.7 (0.57, 0.79) |
| HVLT-R Total Recall | 484 | 22.9 (5.8) | 24 (6.7) | 0.76 (0.7, 0.81) | 254 | 0.44 (0.27, 0.58) | 124 | 0.69 (0.54, 0.8) | 166 | 0.73 (0.62, 0.82) |
| HVLT-R Delayed Recall | 480 | 7.5 (3.5) | 8 (3.8) | 0.72 (0.66, 0.78) | 254 | 0.22 (0.06, 0.38) | 124 | 0.7 (0.55, 0.81) | 164 | 0.75 (0.64, 0.83) |
| HVLT-R Retention | 490 | 74.9 (30.5) | 79.2 (32) | 0.6 (0.51, 0.67) | 254 | 0.17 (0.01, 0.33) | 124 | 0.57 (0.37, 0.72) | 174 | 0.56 (0.4, 0.69) |
| HVLT-R Recognition | 482 | 9.5 (2.6) | 9.5 (2.7) | 0.67 (0.6, 0.74) | 254 | 0.25 (0.08, 0.41) | 124 | 0.36 (0.12, 0.56) | 164 | 0.63 (0.48, 0.75) |
| Letter Fluency (C, F, & L) | 490 | 43.6 (11.5) | 43.6 (11.4) | 0.76 (0.71, 0.81) | 254 | 0.7 (0.6, 0.78) | 126 | 0.8 (0.69, 0.87) | 172 | 0.73 (0.61, 0.82) |
| TMT-A | 492 | 36 (17.7) | 37.1 (21.4) | 0.76 (0.7, 0.81) | 254 | 0.68 (0.58, 0.77) | 126 | 0.55 (0.36, 0.7) | 174 | 0.76 (0.66, 0.84) |
| TMT-B | 464 | 100.3 (58.9) | 101.3 (57.9) | 0.71 (0.64, 0.77) | 252 | 0.76 (0.68, 0.83) | 116 | 0.57 (0.36, 0.72) | 148 | 0.57 (0.39, 0.71) |
| TMT B/A Ratio | 464 | 3 (1.4) | 3 (1.2) | 0.37 (0.26, 0.48) | 252 | 0.5 (0.35, 0.62) | 116 | 0.1 (–0.16, 0.35) | 148 | 0.17 (–0.06, 0.39) |
| Benson Complex Figure – Copy | 494 | 15.4 (1.8) | 14.9 (2) | 0.63 (0.51, 0.72) | 254 | 0.28 (0.11, 0.43) | 126 | 0.56 (0.36, 0.71) | 176 | 0.75 (0.61, 0.84) |
| Benson Complex Figure – Delayed | 490 | 10.6 (3.5) | 10 (4) | 0.79 (0.73, 0.84) | 254 | 0.53 (0.39, 0.64) | 126 | 0.68 (0.52, 0.8) | 172 | 0.81 (0.69, 0.88) |
| MINT | 488 | 29.3 (2.7) | 29.4 (3.1) | 0.85 (0.82, 0.88) | 254 | 0.75 (0.66, 0.82) | 124 | 0.87 (0.79, 0.92) | 170 | 0.85 (0.77, 0.9) |

Abbreviations: CU, Cognitively Unimpaired; Cog Imp, Cognitively Impaired (MCI + dementia); HVLT-R, Hopkins Verbal Learning Test-Revised; ICC, Intraclass Correlation Coefficient; MCI, Mild Cognitive Impairment; MINT, Multilingual Naming Test; MoCA, Montreal Cognitive Assessment; Obs, Observations; SD , Standard Deviation; TMT, Trail Making Test.

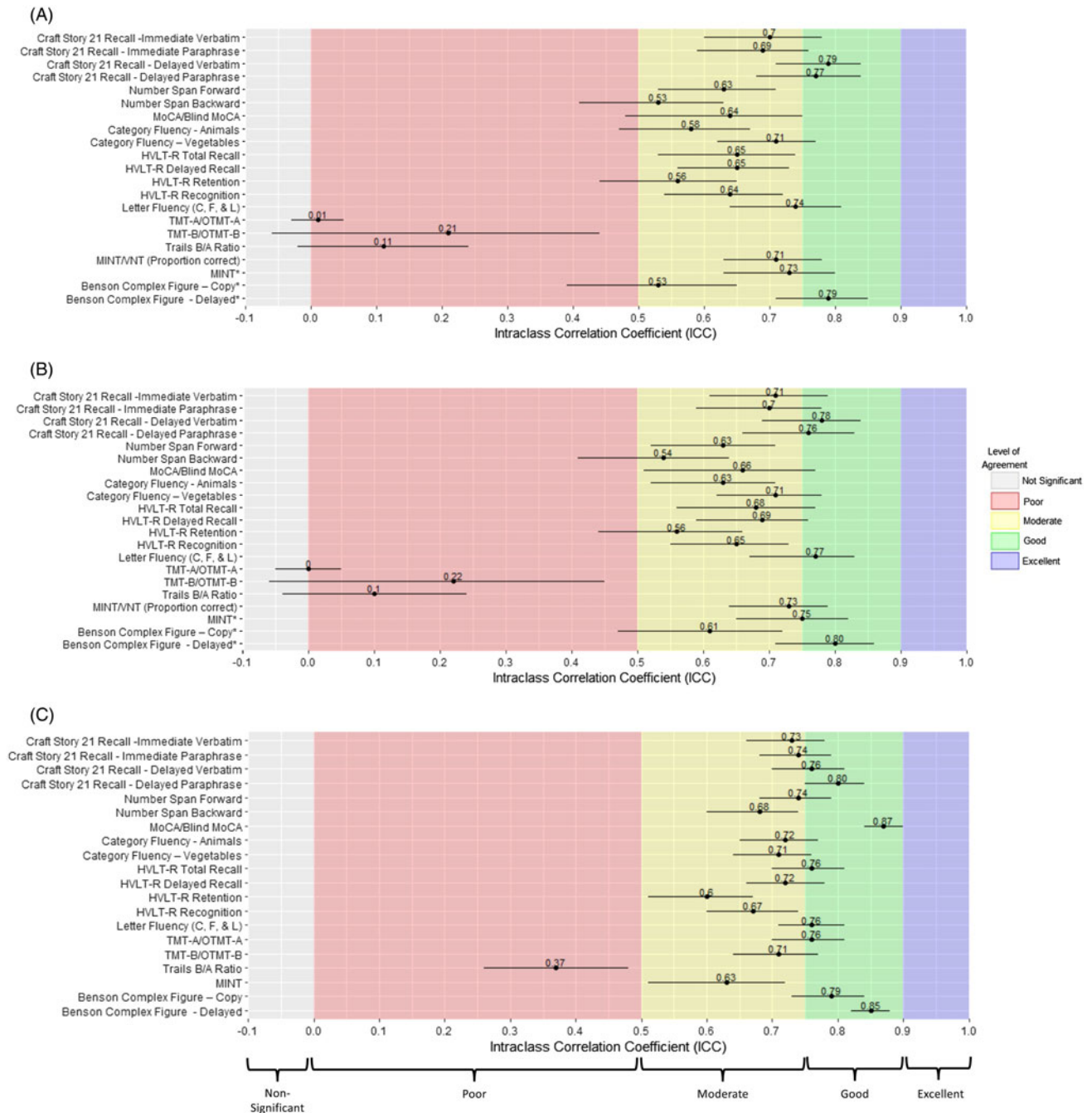


Figure 1. Graphical representation of ICCs for (A) the primary analyses for in-person to remote UDSv3.0 evaluations, (B) diagnostically stable subgroup from (A), and (C) control analyses involving consecutive in-person evaluations. Abbreviations: ICC, Intraclass Correlation Coefficient; UDS, Uniform Data Set.

and we encourage future investigations to control for these factors more systematically. Furthermore, as some methods (e.g., shredding Benson Complex Figure renderings) were simply not feasible given the context of the current study, future efforts should ensure more rigorous test security methods according to published guidelines (Boone et al., 2022). Finally, the number of sessions performed for each modality (i.e., video vs. telephone) was relatively modest, so we encourage replication of all findings. While each limitation is notable, the overall study may provide a more ecologically valid reflection of real-world teleNP when

compared to prior studies conducted under tightly controlled (i.e., ideal) test parameters.

A strength of our study lies in its racial diversity (i.e., 38.7% Black or African American), which addresses a critical gap in the literature (Marra, Hamlet, et al., 2020). Although it was beyond the scope of this investigation to understand whether reliability estimates were differentially influenced by race, our mostly moderate to good agreement across in-person and remote testing modalities lends general support for the adoption of teleNP in a racially diverse sample. We encourage other researchers to explore

teleNP more thoroughly within diverse populations to ensure appropriate inclusion and generalizability of empirical findings.

Conclusion and future directions

Within the context of the naturalistic “experiment” created by the COVID-19 pandemic, our findings revealed primarily moderate to good relationships between the UDS v3.0 t-cog test battery and its in-person counterpart. For certain measures, reliability was somewhat stronger when delivered via video as opposed to telephone, possibly owing to additional visual facial cues available in this format (e.g., participants might more effectively register verbal information when able to see the examiner’s facial expressions and lip movements). In summary, this report is an important initial step in evaluating the reliability of the UDS v3.0 t-cog test battery, and other in-home teleNP testing more broadly. Future work should clarify how diagnostic group and retest duration timeframe affect reliability and should consider the ecological validity of in-home versus traditional, tightly controlled settings.

Supplementary material. The supplementary material for this article can be found at <https://doi.org/10.1017/S1355617723000383>

Acknowledgments. The authors acknowledge the National Institute on Aging and the Michigan Alzheimer’s Disease Research Center, University of Michigan (P30AG053760 and P30AG072931) for making this work possible.

Funding statement. The authors also acknowledge funding from the National Institute on Aging to BMH (R35AG072262).

Competing interests. None.

References

- Abdollahi, A., Bull, M. T., Darwin, K. C., Venkataraman, V., Grana, M. J., Dorsey, E. R., & Biglan, K. M. (2016). A feasibility study of conducting the Montreal Cognitive Assessment remotely in individuals with movement disorders. *Health Informatics Journal*, 22(2), 304–311. <https://doi.org/10.1177/1460458214556373>
- Alegret, M., Espinosa, A., Ortega, G., Pérez-Cordón, A., Sanabria, Á., Hernández, I., Marquí, M., Rosende-Roca, M., Mauleón, A., Abdelnour, C., Vargas, L., de Antonio, E. E., López-Cuevas, R., Tartari, J. P., Alarcón-Martín, E., Tárraga, L. D., Ruiz, A. D., Boada, M., & Valero, S. (2021). From face-to-face to home-to-home: Validity of a teleneuropsychological battery. *Journal of Alzheimer’s Disease*, 81(4), 1541–1553. <https://doi.org/10.3233/JAD-201389>
- Bastug, G., Ozel-Kizil, E. T., Sakarya, A., Altintas, O., Kirici, S., & Altunoz, U. (2013). Oral trail making task as a discriminative tool for different levels of cognitive impairment and normal aging. *Archives of Clinical Neuropsychology*, 28(5), 411–417. <https://doi.org/10.1093/arclin/act035>
- Bilder, R. M., Postal, K. S., Barisa, M., Aase, D. M., Cullum, C. M., Gillaspay, 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). InterOrganizational practice committee recommendations/guidance for teleneuropsychology (TeleNP) in response to the COVID-19 pandemic. *The Clinical Neuropsychologist*, 34(7-8), 1314–1334. <https://doi.org/10.1080/13854046.2020.1767214>
- Boone, K. B., Sweet, J. J., Byrd, D. A., Denney, R. L., Hanks, R. A., Kaufmann, P. M., Kirkwood, M. W., Larrabee, G. J., Marcopulos, B. A., Morgan, J. E., Paltzer, J. Y., Rivera Mindt, M., Schroeder, R. W., Sim, A. H., & Suhr, J. A. (2022). Official position of the American Academy of Clinical Neuropsychology on test security. *Clinical Neuropsychologist*, 36(3), 523–545. <https://doi.org/10.1080/13854046.2021.2022214>
- Brandt, J., & Folstein, M. F. (1988). The telephone interview for cognitive status. *Neuropsychiatry, Neuropsychology, and Behavioral Neurology*, 1, 111–118.
- 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>
- Bunker, L., Hshieh, T. T., Wong, B., Schmitt, E. M., Trivison, T., Yee, J., Palihnich, K., Metzger, E., Fong, T. G., & Inouye, S. K. (2017). The SAGES telephone neuropsychological battery: Correlation with in-person measures. *International Journal of Geriatric Psychiatry*, 32(9), 991–999. <https://doi.org/10.1002/gps.4558>
- Carlew, A. R., Fatima, H., Livingstone, J. R., Reese, C., Lacritz, L., Pendergrass, C., Bailey, K. C., Presley, C., Mokhtari, B., & Cullum, C. M. (2020). Cognitive assessment via telephone: A scoping review of instruments. *Archives of Clinical Neuropsychology*, 35(8), 1215–1233. <https://doi.org/10.1093/arclin/acia096>
- Castanho, T. C., Amorim, L., Zihl, J., Palha, J. A., Sousa, N., & Santos, N. C. (2014). Telephone-based screening tools for mild cognitive impairment and dementia in aging studies: A review of validated instruments. *Frontiers in Aging Neuroscience*, 6, 16. <https://doi.org/10.3389/fnagi.2014.00016>
- Centers for Medicare and Medicaid Services (2021). *COVID-19 emergency declaration blanket waivers for health care providers*. <https://www.cms.gov/files/document/summary-covid-19-emergency-declaration-waivers.pdf>
- 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*, 20(10), 1028–1033. <https://doi.org/10.1017/S1355617714000873>
- Cullum, C. M., Weiner, M. F., Gehrman, H. R., & Hynan, L. S. (2006). Feasibility of telecognitive assessment in dementia. *Assessment*, 13(4), 385–390. <https://doi.org/10.1177/1073191106289065>
- Fox-Fuller, J. T., Ngo, J., Pluim, C. F., Kaplan, R. I., Kim, D.-H., Anzai, J. A. U., Yucebas, D., Briggs, S. M., Aduen, P. A., Cronin-Golomb, A., & Quiroz, Y. T. (2022). Initial investigation of test-retest reliability of home-to-home teleneuropsychological assessment in healthy, English-speaking adults. *Clinical Neuropsychologist*, 36(8), 2153–2167. <https://doi.org/10.1080/13854046.2021.1954244>
- Grosch, M. C., Weiner, M. F., Hynan, L. S., Shore, J., & Cullum, C. M. (2015). Video teleconference-based neurocognitive screening in geropsychiatry. *Psychiatry Research*, 225(3), 734–735. <https://doi.org/10.1016/j.psychres.2014.12.040>
- Hammers, D. B., Stolwyk, R., Harder, L., & Cullum, C. M. (2020). A survey of international clinical teleneuropsychology service provision prior to and in the context of COVID-19. *The Clinical Neuropsychologist*, 34(7-8), 1267–1283. <https://doi.org/10.1080/13854046.2020.1810323>
- Hildebrand, R., Chow, H., Williams, C., Nelson, M., & Wass, P. (2004). Feasibility of neuropsychological testing of older adults via videoconference: Implications for assessing the capacity for independent living. *Journal of Telemedicine and Telecare*, 10(3), 130–134.
- Hunter, M. B., Jenkins, N., Dolan, C., Pullen, H., Ritchie, C., & Muniz-Terrera, G. (2021). Reliability of telephone and videoconference methods of cognitive assessment in older adults with and without dementia. *Journal of Alzheimer’s Disease*, 81(4), 1625–1647. <https://doi.org/10.3233/JAD-210088>
- Kaemmerer, T., & Riordan, P. (2016). Oral adaptation of the Trail Making Test: A practical review. *Applied Neuropsychology: Adult*, 23(5), 384–389. <https://doi.org/10.1080/23279095.2016.1178645>
- Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155–163. <https://doi.org/10.1016/j.jcm.2016.02.012>
- Lindauer, A., Seelye, A., Lyons, B., Dodge, H. H., Mattek, N., Mincks, K., Kaye, J., & Erten-Lyons, D. (2017). Dementia care comes home: Patient and caregiver assessment via telemedicine. *The Gerontologist*, 57(5), e85–e93. <https://doi.org/10.1093/geront/gnw206>
- 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>
- Marra, D. E., Hoelzle, J. B., Davis, J. J., & Schwartz, E. S. (2020). Initial changes in neuropsychologists clinical practice during the COVID-19 pandemic:

- A survey study. *The Clinical Neuropsychologist*, 34(7-8), 1251–1266. <https://doi.org/10.1080/13854046.2020.1800098>
- Miller, J. B., & Barr, W. B. (2017). The technology crisis in neuropsychology. *Archives of Clinical Neuropsychology*, 32(5), 541–554. <https://doi.org/10.1093/arclin/acx050>
- Mrazik, M., Millis, S., & Drane, D. L. (2010). The oral trail making test: Effects of age and concurrent validity. *Archives of Clinical Neuropsychology*, 25(3), 236–243. <https://doi.org/10.1093/arclin/acq006>
- Nasreddine, Z. (2022). *How do I score the MoCA-Blind?* MoCA Cognitive Assessment. <https://www.mocatest.org/faq/>
- National Alzheimer's Coordinating Center (2015). *National Alzheimer's Coordinating Center uniform data set coding guidebook for initial visit packet*. <https://files.alz.washington.edu/documentation/uds3-ivp-guidebook.pdf>
- National Alzheimer's Coordinating Center (2020). *NACC uniform data set instructions for the T-cog neuropsychological battery (form C2T)*. <https://files.alz.washington.edu/documentation/uds3-np-c2t-instructions.pdf>
- Parks, A. C., Davis, J., Spreser, C. D., Stroescu, I., & Ecklund-Johnson, E. (2021). Validity of in-home teleneuropsychological testing in the wake of COVID-19. *Archives of Clinical Neuropsychology*, 36(6), 887–896. <https://doi.org/10.1093/arclin/acab002>
- Ricker, J. H., & Axelrod, B. N. (1994). Analysis of an oral paradigm for the Trail Making Test. *Assessment*, 1(1), 47–51. <https://doi.org/10.1177/1073191194001001007>
- Stillerova, T., Liddle, J., Gustafsson, L., Lamont, R., & Silburn, P. (2016). Could everyday technology improve access to assessments? A pilot study on the feasibility of screening cognition in people with Parkinson's disease using the Montreal Cognitive Assessment via Internet videoconferencing. *Australian Occupational Therapy Journal*, 63(6), 373–380. <https://doi.org/10.1111/1440-1630.12288>
- Thompson, N. R., Prince, M. J., Macdonald, A., & Sham, P. C. (2001). Reliability of a telephone-administered cognitive test battery (TACT) between telephone and face-to-face administration. *International Journal of Methods in Psychiatric Research*, 10(1), 22–28. <https://doi.org/10.1002/mpr.97>
- 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*, 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>
- Webb, K. L., Ryan, J., Wolfe, R., Woods, R. L., Shah, R. C., Murray, A. M., Orchard, S. G., & Storey, E. (2022). Test-Retest reliability and minimal detectable change of four cognitive tests in community-dwelling older adults. *Journal of Alzheimer's Disease*, 87(4), 1683–1693. <https://doi.org/10.3233/JAD-215564>
- Welsh, K. A., Breitner, J. C. S., & Magruder-Habib, K. M. (1993). Detection of dementia in the elderly using Telephone Screening of Cognitive Status. *Neuropsychiatry, Neuropsychology, and Behavioral Neurology*, 6(2), 103–110.
- World Health Organization (2020). *COVID-19 strategy update*. <https://www.who.int/publications/m/item/covid-19-strategy-update>
- Zane, K. L., Thaler, N. S., Reilly, S. E., Mahoney, J. J., & Scarisbrick, D. M. (2021). Neuropsychologists' practice adjustments: The impact of COVID-19. *The Clinical Neuropsychologist*, 35(3), 490–517. <https://doi.org/10.1080/13854046.2020.1863473>