Journal of the International Neuropsychological Society (2022), **28**, 835–844

doi:10.1017/S1355617721001053

Equivalency of In-Person Versus Remote Assessment: WISC-V and KTEA-3 Performance in Clinically Referred Children and Adolescents

Taralee Hamner[1,2], Cynthia F. Salorio[1,3,*], Luther Kalb[1,4] and Lisa A. Jacobson[1,2]

[1]Department of Neuropsychology, Kennedy Krieger Institute, Baltimore, MD, USA
[2]Department of Psychiatry and Behavioral Sciences, Johns Hopkins School of Medicine, Baltimore, MD, USA
[3]Department of Physical Medicine and Rehabilitation, Johns Hopkins School of Medicine, Baltimore, MD, USA
[4]Department of Mental Health, Johns Hopkins School of Public Health, Baltimore, MD, USA

(Received January 18, 2021; Final revision June 15, 2021; Accepted June 23, 2021; First Published Online September 27, 2021)

Abstract

Objective: Teletesting has the potential to reduce numerous barriers to patient care which have only become exacerbated during the COVID-19 pandemic. Although telehealth is commonly utilized throughout medicine and mental health practices, teletesting has remained limited within cognitive and academic evaluations. This may be largely due to concern for the validity of test administration via remote assessment. This cross-sectional study examined the equivalency of cognitive [Wechsler Intelligence Scales for Children – Fifth Edition (WISC-V)] and academic [Kaufman Test of Educational Achievement – Third Edition (KTEA-3)] subtests administered via either teletesting or traditional in-person testing within clinically referred youth. **Method:** Chart review using a retrospective, cross-sectional design included a total of 893 children and adolescents, ranging from 4 to 17 years (Mean age = 10.2 years, $SD = 2.9$ years) who were administered at least one subtest from the aforementioned cognitive or academic assessments. Of these, 285 received teletesting, with the remaining ($n = 608$) receiving in-person assessment. A total of seven subtests (five from the WISC-V and two from the KTEA-3) were examined. A series of inverse probability of exposure weighted (IPEW) linear regression models examined differences between groups for each of the seven subtests after adjustment for numerous demographic, diagnostic, and parent-reported symptom variables. **Results:** Only two significant differences were found, such that WISC-V Visual Puzzles ($p < .01$) and KTEA-3 Math Concepts ($p = .03$) scores were slightly higher in the teletesting versus in-person groups. However, these differences were quite small in magnitude (WISC-V Visual Puzzles, $d = .33$, KTEA-3 Math Concepts, $d = .18$). **Conclusions:** Findings indicate equivalency across methods of service delivery without clinically meaningful differences in scores among referred pediatric patients.

Keywords: Telemedicine, Telehealth, Cognitive, Academic, Neuropsychology, Pediatric

INTRODUCTION

Telehealth is the remote delivery of healthcare services from one site to another (e.g., office, home, clinic). Telehealth has evolved over the past two decades as an important mode of healthcare delivery. From medicine to behavioral health, telehealth reaches individuals beyond the treatment room to deliver assessment, intervention, and follow-up services remotely. Within the field of psychological therapy, teletherapy has grown in popularity. This is due, in part, to the increasing number of studies demonstrating evidence-based treatments can be delivered efficaciously through telehealth (Bashshur, Shannon, Bashur, & Yellowlees, 2016; Berryhill et al., 2019; Egede et al., 2015). On the other hand, research on cognitive and other standardized evaluations delivered remotely (i.e., teletesting), particularly for children and adolescents, remains limited.

Telehealth has the potential to increase access for populations who experience social, economic, geographical, and/or health-related barriers to care (Marcin, Shaikh, & Steinhorn, 2016; Weinstein et al., 2014). Telehealth may also inadvertently exacerbate economic disparities based on access and technological literacy. However, access is rapidly growing; the American Community Survey documents 90.3% of US households have a computer and 82.7% of households have broadband Internet subscriptions (see United States Census Bureau, 2019). In addition, the American Academy of Pediatrics is actively working to improve access to telehealth for groups impacted by these factors (see Jenco, 2020), and

*Correspondence and reprint requests to: Cynthia Salorio, PhD, ABPP, Department of Neuropsychology, Kennedy Krieger Institute, 1750 E. Fairmount Avenue, Baltimore, MD 21231, USA. Phone: 443-923-9440. Email: Salorio@KennedyKrieger.org

teleassessment should evolve with these considerations in mind. Remote cognitive assessment via teletesting requires additional consideration given the standardized nature of administration procedures that are inherently changed when presented over a screen (Hewitt, Rodgin, Loring, Pritchard, & Jacobson, 2020). Presentation of stimuli has been demonstrated to be equivalent when shown on an iPad versus a traditional printed booklet (Daniel, Wahlstrom, & Zhang, 2014). Some test publishing companies had previously released digital assessments tools (e.g., Q-Interactive iPad assessments via Pearson) designed to replace physical "pen and paper" testing materials, but these still require face-to-face administration. In response to the COVID-19 pause in clinic-based services, Pearson Assessments released a "Letter of No Objection," dated March 20, 2020, permitting the use of copyrighted materials to assist in remote assessments. Given the pivot to "on screen" presentation of stimulus books and stimuli that have not yet been normed for remote administration, there is a critical need to assess the equivalence of remote assessments delivered via telehealth.

The majority of studies examining equivalency of performance between in-person and teletesting have focused on adult populations with few studies among children and adolescents. Overall, the literature suggests no effect of testing for remote versus in-person administration for neuropsychological testing (for metanalysis, see Brearly et al., 2017), though this was limited by few available studies, selection bias related to participant age, and mixed designs. This comparability has been documented across a range of populations, such as for those with cognitive impairment (Wadsworth et al., 2018), within culturally diverse groups (Vahia et al., 2015), and for individuals with intellectual disability (Temple, Drummond, Valiquette, & Jozsvai, 2010). Support for equivalence also exists for differing referral concerns, such as dementia (Cullum, Weiner, Gehrmann, & Hynan, 2006; Cullum, Hynan, Grosch, Parikh, & Weiner, 2014), speech–language (Waite, Theodoros, Russell, & Cahill, 2010), academic (Wright, 2016), learning disabilities (Hodge et al, 2019), demyelinating disorders (Harder et al., 2020), neurodegenerative diseases (Ragbeer, Augustine, Mink, et al., 2016), and broader neuropsychological (Galusha-Glasscock, Horton, Weiner, & Cullum, 2016) evaluations. However, given that the majority of these studies included adult measures or those specific to neuropsychological assessment [e.g., Boston Naming Test, Clock Drawing, Mini-Mental Status Exam, Rey Auditory Verbal Learning Test, Repeatable Battery for the Assessment of Neurological Status, Weschler Adult Intelligence Scale (WAIS); see Brearly et al., 2017 for metanalysis], the translation to pediatric care remains unclear.

The COVID-19 pandemic abruptly halted all nonessential services, centering the importance of teletesting before researchers could establish an evidence base. This left many stakeholders without a clear path for assessment services for referred patients as well as students. Several conflicting position papers were released citing validity concerns and lack of evidence regarding teleassessment for school-based evaluations. Given the need to adhere to timelines, the National Association of School Psychologists released updated guidelines specific to the school setting (National Association of School Psychologists, 2020), whereas others encouraged waiting for the return to in-person assessment.

As the situation continues to evolve, psychologists grapple with balancing safety, validity, and ethical responsibility. Farmer et al. (2020a) offered several considerations for the delivery of teleassessment with a lens toward implications for policy and practice. The authors argue that although evidence exists within the adult literature, the child and adolescent literature remains limited and requires unique considerations, particularly in relation to special education services. In the second paper, the authors provide a commentary on the limitations of validity for local educational agencies to consider (Farmer et al., 2020b). Others have outlined solutions for teleassessment and emphasize the importance of moving forward with remote administration, with appropriate caution, for both pediatric and adult groups (Hewitt et al., 2020).

More recently, the feasibility of teletesting has been demonstrated in pediatric patients using a wide variety of measures (specific cognitive measures included selected subtests from the Wechsler Abbreviated Scales of Intelligence – Second Edition and Differential Ability Scales – Second Edition and specific academic measures included selected subtests from Bracken expressive form, Comprehensive Test of Phonological Processing, and Wechsler Individual Achievement Test – Third Edition; Ransom et al., 2020) yet comparison to in-person administration remains limited. Harder et al. (2020) examined teletesting versus in-person assessment via a test–retest design among pediatric patients recruited within a demyelinating clinic. Findings did not reveal any differences in test scores between the two conditions on selected subtests of the California Verbal Learning Test (Children's Version and Second Edition), Symbol Digit Modalities Test, or selected subtests of the Wechsler Intelligence Scale for Children – Fifth Edition (WISC-V), WAIS-IV, Beery-Buktenica Developmental Test of Visual-Motor Integration – Visual Perception, Delis–Kaplan Executive Function System, or Woodcock–Johnson Tests of Academic Achievement – Third Edition.

As discussed by Wright (2020) and Hewitt et al. (2020), remote assessment has the potential to alleviate many of the preexisting structural and systematic challenges to educational evaluations that have only become compounded by the COVID-19 crisis. With the uncertainty of timelines for school services resuming fully in person, teletesting also has the potential to address challenges related to social distancing and long waitlists.

In light of these concerns, Wright (2020) examined performance on the WISC-V (Wechsler, 2014) administered either in person or remotely utilizing a proctor, among a sample of 256 school children. Results did not reveal a method effect for most subtests (except for Letter–Number Sequencing), Index, or Full-Scale IQ scores. Although encouraging, this study did

https://doi.org/10.1017/S1355617721001053 Published online by Cambridge University Press

not include clinically referred children who may show differences in performance by administration type. This study also required the use of a proctor to manage materials in the remote condition, which is not consistent with clinical practices, and the authors only examined the WISC-V. As such, there is a need to examine remote testing strategies among children who are clinically referred in a real-world setting using both cognitive and academic assessments that are crucial for educational evaluations.

Evaluations are often coupled with unique circumstances, such as time-sensitive referral questions or high-stakes eligibility determinations. Additionally, clinically referred children may perform differently due to suspected cognitive or learning needs. Given that some clinics have either been instructed (by institutional guidelines) or opted to convert to telehealth rather than, or in addition to, in-person visits, it is critical that research examine the equivalence of teletesting within this group. Thus, the goal of the present study was to expand upon the nascent pediatric teletesting literature by examining the equivalence of subtests of cognitive (WISC-V) and academic [Kaufman Test of Educational Achievement – Third Edition (KTEA-3); Kaufman & Kaufman, 2014] batteries administered via teletesting versus face to face within a clinically referred sample.

METHODS

Study Design, Inclusion Criteria, and Population

Participants in this retrospective cross-sectional study were referred for psychological/neuropsychological assessment at an urban outpatient testing service of a pediatric hospital in the Mid-Atlantic region of the US. To be included in this study, the participant: (a) must have been between 4 and 18 years of age; and (b) have received a psychological/neuropsychological assessment using any of the measures of interest between November 2019 and March 2020 (for in-person assessment) or April 2020 and August 2020 (for telehealth-based assessment). Evaluations were conducted by clinical psychologists and clinical neuropsychologists, utilizing the Q-Global platform for remote subtest administration. Data from clinical evaluations are routinely entered into the electronic medical record and de-identified records can be retrieved for analysis following appropriate approvals. The hospital's Institutional Review Board approved this retrospective review.

Participants were included if they received at least one subtest on either measure. Thus, some received KTEA-3 but not WISC-V, or vice versa. A total of 893 youth were included (Mean age = 10.1 years, *SD* = 2.9 years); 61% were male and 35% of families were receiving Medical Assistance/Public Insurance. ADHD (61%) and Anxiety/Depression (22%) disorders were the most common billing diagnoses. Slightly more than half were White (54%), with the remaining identifying as Black/African-American (31%) or "Other" races (15%). Within the "Other racial group" ($n = 129$), 41% were listed as "Other" race in the electronic health records system, 29% were Asian, 27% were Multiracial, and the remaining were Hispanic (7%), Native American (2%), and Asian Indian (2%); race was missing for 3% of the sample. See Table 1 for details.

Dependent Variables

Academic achievement

Educational screening was conducted using the KTEA-3 (Kaufman & Kaufman, 2014). The KTEA-3 is a psychometrically sound, academic assessment designed for individuals aged 4–26 or grades prekindergarten through 12. The Letter and Word Recognition and Math Concepts and Applications subtests were chosen as these subtests provide a brief screening of core academic skills. Most importantly for this study, these tests are amenable to telehealth and were available online for remote administration since the beginning of the COVID-19 stay-at-home order in our state. Standard scores for Letter and Word Recognition as well as Math Concepts were used in the analyses.

Intelligence

Core reasoning and brief attention were measured using the WISC-V (Wechsler, 2014). The WISC-V is a well-validated, psychometrically sound, cognitive assessment for use in children aged 6–16. As teletesting utilized subtests most amenable to remote administration by not requiring physical manipulation or written responses, subtests examined included Similarities, Matrix Reasoning, Digit Span, Vocabulary, and Visual Puzzles.

Independent and Control Variables

Demographics

Date of appointment, mode of assessment (in-person vs. telehealth), age (in years), sex (male and female), race (White, Black/African American, Other), insurance type (private and public), and billing diagnosis were captured from the electronic health records system. Billing diagnoses, based on International Classification of Diseases codes, 10th edition, were classified as anxiety/depression (F41, F32, F33, F34, F39), adjustment disorders (F43), attention-based disorders (e.g., F90, R41), epilepsy (G40), oncology (C and D), encephalopathy (G93, G94, G95, G96), genetic conditions (Q), and other (less commonly billed medical and mental health diagnoses). Diagnoses were coded if they were billed as either primary or secondary. Parental education was also captured from the online pre-visit questionnaire (see below).

Online pre-visit parent ratings.

Parents of children scheduled for psychological or neuropsychological assessment were sent a letter providing information about their upcoming appointment. The letter

https://doi.org/10.1017/S1355617721001053 Published online by Cambridge University Press

Table 1. Demographic, clinical, and testing differences between in-person versus telehealth appointments

	In-person assessment	Telehealth assessment	Total
N (%)	608 (68.1)	285 (31.9)	893 (100)
Age (M, *SD*)	10.1 (2.9)	10.2 (3.0)	10.1 (2.9)
Sex (N, %)			
Female	232 (38.6)	113 (39.6)	345 (38.6)
Male	376 (61.8)	172 (60.3)	548 (61.4)
Race (N, %)			
White	318 (53.5)	152 (55.4)	470 (54.1)
Black/AA	189 (31.8)	80 (29.2)	269 (31.0)
Other	87 (14.6)	42 (15.3)	129 (14.9)
Insurance Type (N, %)			
Medical Assistance	224 (36.8)	88 (31.0)	312 (34.9)
Commercial	384 (63.2)	197 (69.1)	581 (65.1)
Billing Diagnoses (N, %)†			
ADHD	393 (64.0)	157 (55.1) *	546 (61.1)
Anxiety/Depression	166 (27.3)	35 (12.3) *	201 (22.5)
Encephalopathy	41 (6.7)	32 (11.2)*	73 (8.2)
Adjustment	40 (6.6)	20 (7.0)	60 (6.7)
Genetic	37 (6.0)	15 (5.2)	52 (5.8)
Oncology	28 (4.6)	17 (6.0)	45 (5.0)
Epilepsy	20 (3.3)	18 (6.3)*	38 (4.3)
Other	50 (8.2)	43 (15.1)*	93 (10.4)
Completed KTEA-3 (N, %)			
Letters and Words	338 (55.9)	184 (64.5) *	522 (58.4)
Math Concepts	210 (34.5)	135 (47.4) *	345 (38.6)
Completed WISC-V (N, %)			
Similarities	493 (81.1)	119 (41.7) *	612 (68.5)
Matrix Reasoning	498 (81.9)	132 (46.3) *	630 (70.5)
Digit Span	509 (83.7)	200 (70.2) *	709 (79.4)
Vocabulary	468 (77.0)	112 (39.3) *	580 (64.9)
Visual Puzzles	413 (67.9)	77 (27.2) *	490 (54.9)
Number of completed KTEA-3 and WISC-V subtests *			
1	53 (8.7)	70 (24.6)	123 (13.8)
2	62 (10.2)	58 (20.3)	120 (13.4)
3	21 (3.4)	42 (14.7)	63 (7.0)
4	29 (4.8)	24 (8.4)	53 (6.0)
5	201 (33.1)	29 (10.0)	230 (25.8)
6	126 (21.1)	28 (9.8)	154 (17.2)
7	116 (19.0)	34 (11.9)	150 (16.9)
Completed parent forms (N, %)	500 (82.2)	205 (71.9) *	705 (79.0)
Parent-reported symptoms (M, *SD*)			
Depression	6.1 (5.0)	5.3 (4.7)	5.9 (5.0)
Anxiety	5.7 (5.0)	5.1 (4.6)	5.5 (4.5)
Oppositional/Conduct	7.0 (6.2)	6.2 (6.1)	6.8 (6.2)
ADHD – Hyperactivity	11.1 (7.4)	10.3 (7.5)	10.9 (7.5)
ADHD – Impulsivity	17.2 (6.2)	16.6 (6.4)	17.0 (6.3)
Reading Problems	17.7 (7.9)	17.4 (7.8)	17.7 (7.8)
Math Problems	14.8 (6.1)	13.8 (6.1)	14.5 (6.1)
Impairment	14.2 (6.8)	13.3 (7.0)	14.0 (6.9)
Sluggish Cognitive Tempo	15.3 (8.2)	13.5 (7.4) *	14.8 (8.1)
KTEA-3 (M, *SD*)			
Letters and Words	88.2 (16.4)	89.6 (16.3)	88.7 (16.2)
Math Concepts	83.9 (15.5)	88.1 (18.0) *	85.6 (16.7)
WISC-V (M, *SD*)			
Similarities	9.2 (3.0)	9.1 (3.2)	9.2 (3.1)
Matrix Reasoning	8.8 (3.3)	8.5 (3.4)	8.7 (3.4)
Digit Span	7.9 (3.0)	8.4 (3.2)	8.1 (3.1)
Vocabulary	9.1 (3.4)	9.4 (3.5)	9.2 (3.5)
Visual Puzzles	9.2 (3.2)	10.1 (3.1) *	9.4 (3.2)

* $p < .05$; †first and second billing diagnoses were included in the coding, thus patients could have multiple diagnoses billed.

https://doi.org/10.1017/S1355617721001053 Published online by Cambridge University Press

included a weblink to an online pre-visit custom developmental history questionnaire hosted via a secure third-party data collection platform. The questionnaire included a series of embedded parent-reported rating scales (described below).

Most (79%) of parent ratings were completed prior to the assessment (median days between questionnaire completion and assessment was 66), although parents were given the option to complete the questionnaire on the day of evaluations, as needed. Parent ratings were more often available for children who completed in-person assessment (82%) compared to telehealth (71%).

In total, eight parent-rated measures were captured from the online pre-visit questionnaire. As with the demographic variables, scores on these measures were employed as control variables to account for any potential differences between children receiving in-person versus telehealth assessments. Internalizing problems were assessed via the Generalized Anxiety (6 items) and Major Depression subscales (10 items) from the Revised Children's Anxiety and Depression Scale – Parent Version (RCADS: Ebesutani et al., 2010). Externalizing problems were evaluated using a subset of eight items tapping Oppositional Defiance and Conduct Disorder (VAN-Conduct) from the Vanderbilt ADHD Diagnostic Parent Rating Scale (Wolraich et al., 2003). The Colorado Learning Difficulties Questionnaire (CLDQ; Willcutt et al., 2011) was employed to identify potential academic problems in the areas of math (CLDQ – Math; five items) and reading (CLDQ – Reading; six items). The Attention Deficit Hyperactivity Disorder (ADHD) Rating Scale-5 (DuPaul et al., 2016), Home Version, was employed to evaluate ADHD symptoms based upon criteria from the Diagnostic and Statistical Manual of Mental Disorders – Fifth Edition (DSM-5; American Psychiatric Association, 2013). The ADHD Rating Scale-5 includes a total of 18 items assessing the hyperactivity (ADHD-HY) and inattention (ADHD-IN) symptom criteria. The Impairment Rating Scale (IRS; Fabiano et al., 2006) was used to measure the impairment of patients across several domains of functioning covering social/peer relationships, relationship with caregivers, academic progress, home life, and self-esteem domains of functioning. The 14-item Sluggish Cognitive Tempo (SCT; Penny, Waschbusch, Klein, Corkum, & Eskes, 2009) scale was used as a measure of cognitive processing speed. All parent-reported measures discussed above have demonstrated strong psychometric properties.

STATISTICAL ANALYSIS PLAN

The goal of this retrospective, cross-sectional study was to evaluate differences between in-person versus virtual administration of select cognitive and academic tests. The primary methodologic concern with this design, and thus this study, is confounding. In the present study, confounding is an important concern because we assume that, after adjusting for numerous sociodemographic and clinical variables, the samples of children in our study who received in-person versus virtual tests were exchangeable (i.e., similar on all observed and unobserved variables).

The first step in the analysis was to examine any differences between the teletesting and in-person groups on demographic and clinical characteristics as well as subtests administered, using descriptive statistics and bivariate (*t*-test, χ^2) analyses. The next step was to address missingness. For demographics and billing diagnoses, there was little to no missing data (<2%). However, there was substantial missingness for the parent-reported ratings (see Table 1). To address this, parent-reported mean raw summary scores were imputed using multiple linear regression methods. Included in the model were patient demographics (age, sex, insurance type) and billing diagnoses (see Table 1). This allowed for full sample inclusion, which has been shown to be less biased than complete case analysis (Wang & Rao, 2001). The imputed data were only employed in the regression analyses.

The third and final step was to examine the association between assessment type and scores on academic and cognitive subtests after adjusting for all demographic, parent-reported, and diagnostic variables. We took a test-wise approach for the analysis, such that each analysis was conducted by subtest, rather than by participant. This allowed for preserving as much data as possible, rather than requiring each participant to have all subtests of interest. For instance, only 15% of the sample received all seven subtests (See Table 1).

To identify differences in subtest scores between in-person versus teletesting methods, a doubly robust inverse probability of exposure weighted (IPEW) linear regression model was employed. This model has two parts. It includes IPEW, based on a propensity score or a single, numerical summary of information representing the probability of exposure (or assessment type) conditional on a set of baseline covariates (i.e., demographic and clinical differences). Weighting by the inverse of exposure, in effect, creates a synthetic sample in which assessment is independent of, and thus balanced across, covariates (Joffe, Ten Have, Feldman, & Kimmel, 2004). The second part of the model, which reflects the term "doubly robust," means the regression model also includes all the variables as covariates, along with the IPEW, in the multivariate model to ensure any residual differences, not captured by the weights, are addressed. Robust standard errors were employed in all models, to address any unobserved clustering or misspecification. When a significant difference was found (i.e., $p < .05$), effect sizes were calculated using Cohen's *d* (Lakens, 2013).

A major benefit of the IPEW model, compared to least squares regression, is that this approach is not subject to multicollinearity. As such, a total of 21 variables were included in the model (i.e., demographics, billing diagnoses, parent-reported symptoms, and an indicator of missingness of the parent-report forms). Using all available information is the recommended approach in IPEW since it maximizes exchangeability between groups (Austin & Stuart, 2015). The final step in the IPEW analysis is to ensure that the weighting procedure was effective in addressing differences

https://doi.org/10.1017/S1355617721001053 Published online by Cambridge University Press

between the groups. This is achieved by: (1) a chi-square test of any remaining differences between groups in the final model and (2) ensuring the standardized differences in means (interpreted the same as effect size) are relatively small (e.g., <.1, or a 10% difference; Austin & Stuart, 2015) after the weighting procedure is employed. All analyses were performed in STATA 15.0 (College Station). The *t*-effects package in STATA was employed to calculate the IPEW. Alpha was set at $p < .05$ for determining statistical significance.

RESULTS

Demographic and diagnostic differences between groups

There were very few demographic differences between the in-person versus teletesting groups. No differences were found in age, sex, race, or insurance type (all $p > .05$; see Table 1). However, there were differences in billing diagnoses, such that those in the teletesting group had lower proportions of ADHD ($\chi^2 = 6.45$, $p = .01$) and anxiety/depression ($\chi^2 = 25.10$, $p < .001$) and greater proportions of encephalopathy ($\chi^2 = 5.20$, $p = .02$) and epilepsy ($\chi^2 = 4.40$, $p = .04$).

Completed subtests and parent-report ratings between groups

The in-person group was less likely to receive the KTEA-3 Letter and Word Recognition ($\chi^2 = 6.49$, $p = .01$) or the Math Concepts and Applications ($\chi^2 = 13.47$, $p < .001$) subtests. Those in the in-person group, however, were more likely to receive at least one of the five WISC-V subtests (Similarities, $\chi^2 = 139.19$, $p < .001$; Matrix Reasoning, $\chi^2 = 118.31$, $p < .001$; Digit Span, $\chi^2 = 21.75$, $p < .001$; Vocabulary, $\chi^2 = 120.99$, $p < .001$; Visual Puzzles, 131.15, $p < .001$) than the telehealth group ($p > .05$). Overall, there was a greater number of KTEA-3/WISC-V subtests completed among those in the in-person versus telehealth group ($\chi^2 = 148.35$, $p < .001$; see Table 1 for details). Finally, there was a greater proportion of completed parent-report ratings among those in the in-person versus teletesting groups, likely related to a suspension of the requirement for completion prior to scheduling during the telehealth period ($\chi^2 = 12.40$, $p < .001$).

Unadjusted subtest and parent-reported symptom differences between groups

Tables 1 and 2 display the unadjusted differences in KTEA-3 and WISC-V scores as well as symptom ratings between groups. For the KTEA-3, no differences were found in Letter and Word Recognition scores; however, there were slightly higher Math Concepts scores in the teletesting group. The only difference in WISC-V scores was a slightly higher Visual Puzzles score in the teletesting group (see Table 2). For parent-reported symptoms, those in the teletesting group had lower SCT scores when compared to the in-person group ($t = 2.57$, $p = .01$); no other differences were found.

Adjusted subtest differences between groups

After employing the doubly robust, IPEW regression model, no differences were found in KTEA-3 Letter and Word Recognition scores ($\beta = 1.12$, 95% CI: −1.14, 3.37, $p = .33$); however, there remained a small difference in Math Concepts scores ($\beta = 2.95$, 95% CI: .24, 5.67, $p = .03$). For cognitive scores, no differences were found for the following WISC-V subtests: Similarities ($\beta = .18$, 95% CI: −.33, .69, $p = .47$), Matrix Reasoning ($\beta = -.24$, 95% CI: −.83, .35, $p = .42$), Digit Span ($\beta = .42$, 95% CI: −.20, 1.05, $p = .18$), and Vocabulary ($\beta = .44$, 95% CI: −.11, 1.00, $p = .11$). However, Visual Puzzles was slightly higher for the teletesting group ($\beta = .96$, 95% CI: .29, 1.63, $p = .005$). Effect sizes for both of the significant findings were small (WISC-V Visual Puzzles, $d = .33$; KTEA-3 Math Concepts, $d = .18$)

After the weighting procedure, the chi-square test for covariate balance was nonsignificant for all tests (all $p > .50$). This demonstrates there were no statistically significant residual differences between the groups after the weights were applied. However, a few imbalances (>.10 or a 10% difference; Austin & Stuart, 2015) remained, albeit nonsignificantly, for parent-reported symptoms. No covariate imbalances were observed for the KTEA-3 tests. For the WISC-V tests, a few imbalances remained for the Similarities (Depression, Math, and SCT), Matrix Reasoning (Depression), Digit Span (Depression and SCT), Vocabulary (Depression and Math), Visual Puzzles (Depression, Math, and Reading). A total of 21 variables were included as covariates, across 7 tests (equaling 147 total adjustments). Given only 7% were above the threshold for imbalance, and the chi-square test was highly nonsignificant for each test, these findings suggest overall covariate balance was well achieved for this study.

DISCUSSION

This study examined the equivalence of in-person versus teletesting within a referred pediatric sample. This question is of critical importance given the current COVID-19 pandemic and related changes in assessment methodology across both healthcare and school settings. Results from the multivariate analyses found equivalency of performance on four of the five WISC-V subtests and one of the two KTEA-3 subtests across administration methods. Of the two subtests that differed statistically, the effect sizes were both small in magnitude. In both instances, the teletesting group scored slightly higher than the in-person group. However, given there was less than a 1-point difference in Visual Puzzles scaled scores and less than a 3-point difference in KTEA-3 Math Concepts standard scores, these differences are not clinically meaningful. These findings, along with the equivalency of five additional subtests, provide support for the use of these subtests

https://doi.org/10.1017/S1355617721001053 Published online by Cambridge University Press

Table 2. Changes in mean differences

	Unadjusted			Adjusted		
	mean differences	*t*-value	*p*	Mean differences	*z*-value	*p*
KTEA-3						
Letter and Word	1.37	$t = -.91$,	$p = .36$	1.11	z = .97,	$p = .33$
Math Concepts	4.49	$t = -2.47$,	$p = .01$	2.95	z = 2.13	$p = .03$
WISC-V						
Similarities	−.06	$t = .21$	$p = .84$	.18	z = .71	$p = .48$
Matrix Reasoning	−.27	$t = .83$	$p = .41$	−.24	z = −.81	$p = .42$
Digit Span	.09	$t = -.29$	$p = .77$	.43	z = 1.34	$p = .18$
Vocabulary	.30	$t = -.80$	$p = .42$	.44	z = 1.58	$p = .11$
Visual Puzzles	.85	$t = 2.18$	$p = .03$	.96	z = 2.81	$p < .01$

Positive mean differences reflect greater teleassessment scores compared to in-person assessment.

via teletesting. Within the present study, the adaptation of teletesting was left to clinician discretion. As such, there were some diagnostic group differences that resulted from greater uptake of telehealth by neuropsychologists in certain clinics.

Although test publishers have released digital assessments that are ideal for use in teletesting, these have been critiqued as inherently different from remote administration (e.g., iPad display replicates testing booklets by lying flat on the testing table, whereas remote assessment often involves upright screens; Farmer et al., 2020a). Despite these concerns, findings suggest that standardized materials are robust to these variations and teletesting does not substantively impact scores in either direction.

Findings add to the nascent literature on validity for remote administration in children and adolescents. Specifically, although prior studies have demonstrated the feasibility of teletesting, this study expands the literature by demonstrating the equivalence of teletesting relative to in-person assessment. We also replicate prior work investigating the WISC-V and expand prior work including academic measures by examining selected subtests of the KTEA-3. In addition, whereas prior work has included the use of a "proctor" or "assistant" for administration of the WISC-V on the participant's end (Wright, 2016), this study documents that children and adolescents are generally capable of navigating the testing environment without that level of assistance. This greatly reduces the risk of viral exposure and also provides support for improving broader access to services. Further, this study demonstrates equivalency within a diverse sample, as our sample is representative of our city (Baltimore, Maryland) as well as the US overall (United States Census Bureau, 2019).

With due consideration to the technical, ethical, and legal factors related to service delivery (Farmer et al., 2020a, b; Hewitt et al., 2020), providers now have additional evidence for comparability of telehealth and traditional in-person assessment methods. With accumulating evidence for comparability, psychologists can have confidence in the validity of these measures and can now turn their focus to considering whether telehealth methods are appropriate for specific patients or students, based upon factors unique to each referral. Further, several recent papers have detailed models for clinical decision-making related to teletesting, such as through tiered triage (Koterba et al., 2020; Peterson, Ludwig, & Jashar, 2020; Pritchard et al., 2020).

Beyond the unique circumstances that the COVID-19 pandemic has imposed on families, educators, schools, psychologists, and test publishers alike, teletesting has critical implications for reducing longstanding barriers related to distance/transportation and subsequent disparities in care. Evidence of equivalency provides a strong foundation from which providers can actively and confidently serve students and patients from underserved populations moving forward.

LIMITATIONS AND STRENGTHS

There are several methodological strengths and limitations that should be considered when interpreting the findings. First, this cross-sectional study employed a sequential cohort design in which the in-person and teletesting groups included different samples. One benefit of this design is that it avoids practice effects or fatigue of repeat administration. However, this design raises concerns about confounding between groups. The analysis paid close attention to this issue in two ways. First, this study employed robust measurement of demographic, clinical, and parent-reported characteristics of the child. Second, modern statistical methods were used to account for these differences, including testing for any residual differences after adjustment.

Unfortunately, the sample size could not support the assessment of subgroup differences due to lack of power. This was particularly true for race, where the sample underrepresented minorities other than Black/African Americans. This will be an important area of future research. Cognitive assessments were only obtained for children 6 years or older (i.e., WISC-V) and thus such assessments of younger children should be explored. Another limitation is that there may be additional factors to consider which were not available for analysis, such as technological (e.g., Internet speed, screen size, setting/environment) and clinical (e.g., referral

https://doi.org/10.1017/S1355617721001053 Published online by Cambridge University Press

reasons may impact whether testing was completed and which subtests were administered) factors which warrant further investigation.

Teletesting may be limited by economic barriers and findings should be interpreted with attention to technological access and literacy. Our department's care coordination center assessed interest/comfort in telehealth visits as well as access to appropriate technology at the start of the pandemic. Data related to this are described in Pritchard et al., 2020 and detail that 94% of respondents were interested in telehealth appointments, whereas 74% had access to the technology needed. National rates of computer and Internet access exceed that of our study (United States Census Bureau, 2019). For those without adequate technology, hotspots or tablets were provided as part of a grant. Our department saw an increase in the proportion of visits for those with medical assistance following the transition to telehealth.

Furthermore, not all subtests were administered via both assessment modalities and thus current comparisons are limited to the subtests amenable to telehealth administration (i.e., do not require manipulatives/motor response). For a review of novel telehealth triage models with considerations for in-person versus teletesting, see Koterba et al., 2020 and Peterson et al., 2020. In addition, results do not include composite and Full-Scale IQ scores (although selected subtests do allow for calculation of several composite scores). Pearson offers guidance on calculating non-motor Full-Scale IQ and GAI scores as well as non-motor processing speed and visual-spatial indexes through *Essentials of WISC-V Integrated Assessment* (Raiford, 2017). Given demonstrated equivalency, however, it is unlikely that differences would emerge when composites are computed. Future studies should explicitly explore subtests that are motor dependent to inform teletesting practices. Of note, the field is beginning to adopt technology that may provide avenues through which motor tasks are amenable to telehealth (e.g., Coding and Symbol Search administered through Q-Interactive).

Finally, this study employed a sample, albeit with missing data, gathered from a single site and thus may not be entirely generalizable across the US. Nevertheless, the sample was large and conducted in a real-world setting among a clinically and demographically heterogeneous group.

CONCLUSION

There is limited evidence for the validity of cognitive and academic teletesting with children and adolescents. The present study fills this gap by offering timely results demonstrating equivalence, between tele- and in-person assessment, across select WISC-V and KTEA-3 subtests in a large heterogeneous sample of referred children using robust measurement and analytic procedures. The findings hold important implications for reducing disparities through expanding teleassessment in the era of COVID-19 and beyond.

ACKNOWLEDGMENTS

None.

FINANCIAL SUPPORT

None.

CONFLICT OF INTEREST

The authors have nothing to disclose.

ETHICAL STANDARDS

Data were collected as part of routine clinical care. The Johns Hopkins Medicine Institutional Review Board granted approval to extract the data from the electronic health record and to create a separate de-identified research database for this study.

References

American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). https://doi.org/10.1176/appi.books.9780890425596

Austin, P.C. & Stuart, E.A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in medicine*, *34*(28), 3661–3679.

Bashshur, R.L., Shannon, G.W., Bashshur, N., & Yellowlees, P.M. (2016). The empirical evidence for telemedicine interventions in mental disorders. *Telemedicine and e-Health*, *22*(2), 87–113.

Berryhill, M.B., Culmer, N., Williams, N., Halli-Tierney, A., Betancourt, A., Roberts, H., & King, M. (2019). Videoconferencing psychotherapy and depression: a systematic review. *Telemedicine and e-Health*, *25*(6), 435–446.

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.

Cullum, C.M., Weiner, M.F., Gehrmann, H.R., & Hynan, L.S. (2006). Feasibility of telecognitive assessment in dementia. *Assessment*, *13*(4), 385–390.

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.

Daniel, M.H., Wahlstrom, D., & Zhang, O. (2014). Equivalence of Q-interactive® and Paper Administrations of Cognitive Tasks: WISC®–V. *Q-Interactive Technical Report, 8*.

DuPaul, G.J., Reid, R., Anastopoulos, A.D., Lambert, M.C., Watkins, M.W., & Power, T.J. (2016). Parent and teacher ratings of attention-deficit/hyperactivity disorder symptoms: Factor structure and normative data. *Psychological Assessment*, *28*(2), 214.

Ebesutani, C., Bernstein, A., Nakamura, B.J., Chorpita, B.F., & Weisz, J.R. (2010). A psychometric analysis of the revised child

https://doi.org/10.1017/S1355617721001053 Published online by Cambridge University Press

anxiety and depression scale—parent version in a clinical sample. *Journal of Abnormal Child Psychology*, *38*, 249–260.

Egede, L.E., Acierno, R., Knapp, R.G., Lejuez, C., Hernandez-Tejada, M., Payne, E.H., & Frueh, B.C. (2015). Psychotherapy for depression in older veterans via telemedicine: a randomised, open-label, non-inferiority trial. *The Lancet Psychiatry*, *2*(8), 693–701.

Fabiano, G.A., Pelham, J., William E, Waschbusch, D.A., Gnagy, E.M., Lahey, B.B., … Burrows-MacLean, L. (2006). A practical measure of impairment: Psychometric properties of the impairment rating scale in samples of children with attention deficit hyperactivity disorder and two school-based samples. *Journal of Clinical Child and Adolescent Psychology*, *35*(3), 369–385.

Farmer, R.L., McGill, R.J., Dombrowski, S.C., McClain, M.B., Harris, B., Lockwood, A.B., … Loethen, E. (2020a). Teleassessment with children and adolescents during the coronavirus (COVID-19) pandemic and beyond: Practice and policy implications. *Professional Psychology: Research and Practice*, *51*(5), 477–487.

Farmer, R.L., McGill, R.J., Dombrowski, S.C., Benson, N.F., Smith-Kellen, S., Lockwood, A.B., … Stinnett, T.A. (2020b). Conducting Psychoeducational Assessments During the COVID-19 Crisis: the Danger of Good Intentions. *Contemporary School Psychology*, *25*(1), 27–32.

Galusha-Glasscock, J.M., Horton, D.K., Weiner, M.F., & Cullum, C.M. (2016). Video teleconference administration of the repeatable battery for the assessment of neuropsychological status. *Archives of Clinical Neuropsychology*, *31*(1), 8–11.

Harder, L., Hernandez, A., Hague, C., Neumann, J., McCreary, M., Cullum, C.M., & Greenberg, B. (2020). Home-based pediatric teleneuropsychology: A validation study. *Archives of Clinical Neuropsychology*, *35*, 1266–1275.

Hewitt, K.C., Rodgin, S., Loring, D.W., Pritchard, A.E., & Jacobson, L.A. (2020). Transitioning to telehealth neuropsychology service: considerations across adult and pediatric care settings. *The Clinical Neuropsychologist*, *34*(7–8), 1335–1351.

Hodge, M.A., Sutherland, R., Jeng, K., Bale, G., Batta, P., Cambridge, A., … Silove, N. (2019). Agreement between telehealth and face-to-face assessment of intellectual ability in children with specific learning disorder. *Journal of Telemedicine and Telecare*, *25*(7), 431–437.

Jenco, M. (2020). *AAP awarded $6 million to improve access to care via telehealth models in response to the COVID-19 pandemic*. Retrieved 23 March 2021. American Academy of Pediatrics. https://www.aappublications.org/news/2020/04/30/covid19grant043020.

Joffe, M.M., Ten Have, T.R., Feldman, H.I., & Kimmel, S.E. (2004). Model selection, confounder control, and marginal structural models: review and new applications. *The American Statistician*, *58*(4), 272–279.

Kaufman, A.S. & Kaufman, N.L. (2014). *Kaufman Test of Educational Achievement* (3rd ed.). Bloomington, MN: Pearson.

Koterba, C.H., Baum, K.T., Hamner, T., Busch, T.A., Davis, K.C., Tlustos-Carter, S., … Slomine, B.S. (2020). COVID-19 issues related to pediatric neuropsychology and inpatient rehabilitation–challenges to usual care and solutions during the pandemic. *The Clinical Neuropsychologist*, *34*(7–8), 1380–1394.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in psychology*, *4*, 863.

Marcin, J.P., Shaikh, U., & Steinhorn, R.H. (2016). Addressing health disparities in rural communities using telehealth. *Pediatric Research*, *79*(1), 169–176.

Penny, A.M., Waschbusch, D.A., Klein, R.M., Corkum, P., & Eskes, G. (2009). Developing a measure of sluggish cognitive tempo for children: Content validity, factor structure, and reliability. *Psychological assessment*, *21*(3), 380.

Peterson, R.K., Ludwig, N.N., & Jashar, D.T. (2020). A case series illustrating the implementation of a novel tele-neuropsychology service model during COVID-19 for children with complex medical and neurodevelopmental conditions: A companion to Pritchard et al., 2020. *The Clinical Neuropsychologist*, *35*(1), 99–114.

Pritchard, A.E., Sweeney, K., Salorio, C.F., & Jacobson, L.A. (2020). Pediatric neuropsychological evaluation via telehealth: Novel models of care. *The Clinical Neuropsychologist*, *34*(7–8), 1367–1379.

National Association of School Psychologists (2020). *Virtual Service Delivery in Response to COVID-19 Disruptions*.

Ragbeer, S.N., Augustine, E.F., Mink, J.W., Thatcher, A.R., Vierhile, A.E., & Adams, H.R. (2016). Remote assessment of cognitive function in juvenile neuronal ceroid lipofuscinosis (Batten disease) a pilot study of feasibility and reliability. *Journal of child neurology*, *31*(4), 481–487.

Raiford, S.E. (2017). *Essentials of WISC-V integrated assessment*. Hoboken, NJ, USA: John Wiley & Sons.

Ransom, D.M., Butt, S.M., DiVirgilio, E.K., Cederberg, C.D., Srnka, K.D., Hess, C.T., … Katzenstein, J.M. (2020). Pediatric Teleneuropsychology: Feasibility and Recommendations. *Archives of Clinical Neuropsychology*. https://doi.org/10.1093/arclin/acaa103

Temple, V., Drummond, C., Valiquette, S., & Jozsvai, E. (2010). A comparison of intellectual assessments over video conferencing and in-person for individuals with ID: preliminary data. *Journal of Intellectual Disability Research*, *54*(6), 573–577.

United States Census Bureau (2019). *Quick Facts Population Estimates*. Retrieved 11/05/2020 from https://www.census.gov/quickfacts/fact/table/US/PST045219

Vahia, I.V., Ng, B., Camacho, A., Cardenas, V., Cherner, M., Depp, C.A., … Agha, Z. (2015). Telepsychiatry for neurocognitive testing in older rural Latino adults. *The American Journal of Geriatric Psychiatry*, *23*(7), 666–670.

Wadsworth, H.E., Dhima, K., Womack, K.B., Hart Jr, 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.

Waite, M.C., Theodoros, D.G., Russell, T.G., & Cahill, L.M. (2010). Internet-based telehealth assessment of language using the CELF–4. *Language, Speech, and Hearing Services in Schools*.

Wang, Q. & Rao, J. (2001). Empirical likelihood for linear regression models under imputation for missing responses. *Canadian Journal of Statistics*, *29*(4), 597–608.

Wechsler, D. (2014). *Wechsler Intelligence Scale for Children* (5th ed.). Bloomington, MN: NCS Pearson, Incorporated.

Weinstein, R.S., Lopez, A.M., Joseph, B.A., Erps, K.A., Holcomb, M., Barker, G.P., & Krupinski, E.A. (2014). Telemedicine, telehealth, and mobile health applications that work: opportunities and barriers. *The American journal of medicine*, *127*(3), 183–187.

Willcutt, E.G., Boada, R., Riddle, M.W., Chhabildas, N., DeFries, J.C., & Pennington, B.F. (2011). Colorado Learning Difficulties Questionnaire: validation of a parent-report screening measure. *Psychological assessment*, *23*(3), 778.

Wolraich, M.L., Lambert, W., Doffing, M.A., Bickman, L., Simmons, T., & Worley, K. (2003). Psychometric properties of the Vanderbilt ADHD diagnostic parent rating scale in a referred population. *Journal of pediatric psychology*, *28*(8), 559–568.

Wright, A.J. (2016). Equivalence of remote, online administration and traditional, face-to-face administration of the Woodcock–Johnson IV cognitive and achievement tests (online white paper).

Wright, A.J. (2020). Equivalence of remote, digital administration and traditional, InPerson administration of the Wechsler Intelligence Scale for Children, Fifth Edition (WISC-V). *Psychological Assessment*. Advance online publication. http://dx.doi.org/10.1037/pas0000939