ASHA American Speech-Language-Hearing Association

AJSLP

Research Article

Reliability of Scoring Telehealth Speech Sound Assessments Administered in Real-World Scenarios

Deborah R. Campbell[a] and **Howard Goldstein**[a]

[a] Department of Communication Sciences and Disorders, University of South Florida, Tampa

ARTICLE INFO

Article History:
Received July 15, 2021
Revision received October 3, 2021
Accepted January 20, 2022

Editor-in-Chief: Katherine C. Hustad
Editor: Ignatius Nip

https://doi.org/10.1044/2022_AJSLP-21-00219

ABSTRACT

Purpose: COVID-19 caused a worldwide conversion from in-person therapy to telehealth; however, limited evidence to support the efficacy of remotely delivering standardized tests puts the future of widespread telehealth use at risk. The aim of this study is to investigate the reliability of scoring a speech sound assessment administered in real-world scenarios including two examples of telehealth technology.
Method: A total of thirty-nine 3- to 8-year-olds were administered the Goldman-Fristoe Test of Articulation–Third Edition. Licensed speech-language pathologists (SLPs) concurrently scored children's responses in person and in two telehealth conditions considered typical and enhanced. Mean standard scores and interrater reliability results were compared among the three conditions. Descriptive statistics were used to summarize the frequency of technology and behavior disruptions during administration and the results of an SLP telehealth perception survey.
Results: All scoring conditions were found to be highly correlated, with mean differences revealing no systematic differences of one condition over- or underestimating another. Although response agreement was high (85%–87%), final sounds in words or sounds that are difficult to observe tended to attenuate reliability. Neither child nor technology disruptions affected SLPs' ability to score responses. Despite no significant differences between conditions on scoring reliability, SLP participants reported they continued to prefer in-person over a telehealth speech sound assessment.
Conclusions: This study supports the provision of a pediatric speech sound assessment using consumer-grade equipment, as in-person, typical telehealth, and enhanced telehealth scoring conditions produced similar results. However, SLP participants' skeptical attitudes toward remote delivery of standardized tests reveal an ongoing barrier to widespread telehealth use.
Supplemental Material: https://doi.org/10.23641/asha.19593367

In March 2020, the COVID-19 pandemic caused a massive, worldwide conversion from in-person care to synchronous video conferencing or telehealth (also known as telemedicine, telepractice, teletherapy, telespeech, teleassessment, and telerehabilitation; American Speech-Language-Hearing Association [ASHA], 2016a, 2016c; Bashshur et al., 2020; Cason & Cohn, 2014; Freckmann et al., 2017; Hao et al., 2021; Keck & Doarn, 2014; Kichloo et al., 2020; Smith et al., 2020; Tohidast et al., 2020). For many speech-language pathologists (SLPs), this was their first experience with remote service delivery. Despite its novelty, most SLPs learned and then quickly adopted telehealth, a testament to their resiliency (Campbell & Goldstein, 2021; Tohidast et al., 2020). Furthermore, SLPs' perceptions of telehealth evolved during the pandemic. Initially, telehealth was the only option many SLPs had if they were going to continue providing direct therapy services. However, over time, many SLPs discovered the benefits of offering services remotely. Speech-language pathology clinicians self-reported that they planned to continue to use telehealth well into the future, thus creating a whole new generation of telepractitioners (Campbell & Goldstein, 2022; Tohidast et al., 2020).

Correspondence to Deborah R. Campbell: drcampbell1@usf.edu. ***Disclosure:** The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

During COVID-19, telehealth therapy providers discovered that children were using a range of equipment and software variations, such as a smartphone connected to a mobile broadband connection or a tablet with earbuds connected to a public hotspot (ASHA, n.d.a). SLPs realized that they rarely had an option other than using the technology a child's family had readily available to provide direct patient care during the pandemic (Hao et al., 2021; Tohidast et al., 2020; Wang et al., 2020). Additionally, SLPs employed varying types of technology to provide their services remotely, including laptops and tablets, headphones, and external microphones (Campbell & Goldstein, 2022; Tohidast et al., 2020). Providers, responding to unprecedented circumstances acknowledged the prevalent disparities in resources that often fell short of conditions recommended for ideal telehealth delivery models (Tohidast et al., 2020).

ASHA also recognized that the therapy offered via telehealth during the pandemic may not be ideal (ASHA, 2020c). ASHA continued to encourage SLPs to adhere to their guidelines and Code of Ethics to ensure that SLPs were providing services of the highest quality when delivering therapy remotely (ASHA, 2016a, 2016c, 2020a, 2020b). ASHA's position remained that telepractice must be consistent with the quality of care offered in person (ASHA, 2020c). In particular, ASHA expressed hesitancy about broadly supporting remote delivery of evaluative or diagnostic services (ASHA, 2020c). ASHA noted that the validity and reliability of most pediatric assessments have not been evaluated for remote administration (ASHA, 2020c; Freckmann et al., 2017; McGill et al., 2021; Werfel et al., 2021). ASHA advised that standardized evaluations conducted with deviations, such as prompting or modifications to delivery, may impact interpretation of scores or require children to be reassessed in the future through in-person administration to acquire valid results (ASHA, 2020c; Freckmann et al., 2017). Moreover, an assessment's scoring could be affected if judging responses could be degraded through remote technology.

Telehealth Research Literature

Taylor et al. (2014) undertook a systematic review of articles from January 2004 through July 2014 that investigated the reliability and validity of scoring speech and language assessments administered to children through synchronous videoconferencing. Out of 180 articles identified initially, only five met the inclusion criteria: peer-reviewed comparison studies of speech and language assessments administered remotely and in person. Overall, the authors concluded that evidence suggested valid clinical use of assessments for articulation screening, language, oral–motor function, and overall speech intelligibility agreement. However, the findings revealed inadequate evidence to support overall remote administration of standardized diagnostic testing (Taylor et al., 2014). Similarly, Sutherland et al. (2017) noted that there were few studies examining teleassessments for diagnosing and monitoring speech and language deficits in children. Sutherland et al. also suggested that teleassessment investigations need to be performed using consumer-grade equipment to reflect real-world applications. The need for research on telehealth assessments for children with communication disorders remains acute.

Because of the limited evidence to support assessments administered remotely, many school- and clinic-based pediatric SLPs initially opted to defer evaluating children with standardized assessments in lieu of offering them remotely during the pandemic (ASHA, n.d.a; Hall-Mills et al., 2021). Campbell and Goldstein (2021) reported that only half of speech-language pathology survey participants administered speech sound production assessments via telehealth during the coronavirus pandemic. Moreover, speech-language clinicians reported that standardized assessments for speech disorders were one of the most difficult to administer and score remotely, second only to swallowing evaluations (Campbell & Goldstein, 2021; Hall-Mills et al., 2021). The ability to obtain reliable assessment results are an integral part of interpreting test results and making informed diagnostic decisions (Daub et al., 2021).

During the coronavirus pandemic, SLPs were not the only ones expressing concerns about the difficulty with remote delivery of assessments. Werfel et al. (2021) described parents' reservations about participating in teleassessments due to concerns about a child's ability to meaningfully participate in a virtual testing environment. Parents worried about a child's ability to stay engaged, a child's distractibility, disruptive behaviors, and challenges with technology potentially affecting the outcomes of evaluations (Werfel et al., 2021). It is unknown the extent to which behavior and technology problems affect SLPs' ability to remotely hear or see a child to score test items.

Despite some evidence of comparable remote versus in-person testing outcomes, most current pediatric assessments have not been evaluated (McGill et al., 2021; Sutherland et al., 2017). Some standardized tests that were previously investigated are now outdated (i.e., Clinical Evaluation of Language Fundamentals–Fourth Edition; Goldman-Fristoe Test of Articulation–Second Edition [GFTA-2]). Others were conducted with inadequate samples or performed under ideal laboratory conditions, that is, high-end, custom-built computers (Eriks-Brophy et al., 2008; Sutherland et al., 2017; Taylor et al., 2014; Waite et al., 2010). Moreover, studies did not assess SLPs' satisfaction and impressions. These are factors that can help us understand SLPs' willingness and confidence in conducting teleassessments (Raatz et al., 2021; Taylor et al., 2014). However, investigators have found that SLPs remain especially wary about their ability to score articulation tests

(ASHA, 2020c; Campbell & Goldstein, 2022; Hall-Mills et al., 2021; McGill et al., 2021).

Despite the low incidence of telehealth use prior to COVID-19, innovative pediatric and adult telehealth researchers advanced our knowledge about this delivery method (ASHA, 2002, 2020b; Fong et al., 2021; Hill & Miller, 2012; Mohan et al., 2017; Taylor et al., 2014; Tucker, 2012). In pediatric telehealth research, Coufal et al. (2018) used the ASHA Functional Communication Measure and National Outcome Measurement System to investigate differences between therapy provided for speech sound disorders delivered remotely in contrast to the traditional, in-person setting. The lack of significant difference between treatment outcomes supported the use of telehealth for children with speech sound deficits (Coufal et al., 2018). Hodge et al. (2019) and Wright (2020) provided evidence for the reliability and feasibility for delivering and scoring the Wechsler Intelligence Scale for Children–Fifth Edition, a pediatric cognitive assessment remotely. Moreover, they also reported that psychologists and parents rated telehealth positively. In adult telehealth research, Dekhtyar et al. (2020) recognized the increased growth in telehealth in both research and clinical settings. Therefore, they set out to validate the synchronous videoconference administration of a widely used assessment for aphasia, Western Aphasia Battery–Revised (Kertesz, 2007). Dekhtyar et al. found that in-person and remote delivery methods were highly correlated, with no differences in domain scores. Importantly, they encouraged participants to use their own technology to maintain ecological validity and provided guidelines for modifying telehealth administration to enhance feasibility.

A telehealth infrastructure that uses consumer-grade equipment is a critical component for replication of telehealth studies (Dekhtyar et al., 2020). Taylor et al. (2014) noted that studying telehealth under ideal research conditions in contrast to real-world scenarios is a limitation to this line of research. Research investigating the effectiveness of telehealth services must mirror current, real-world conditions to apply the findings to everyday use (Benda et al., 2020; Dekhtyar et al., 2020; Sutherland et al., 2017, 2018; Taylor et al., 2014). The importance of telehealth infrastructure and technology choice has become apparent in telehealth studies performed in the past 20 years. On the basis of a systematic review of speech and language therapy services for school-aged children, Wales et al. (2017) reported the widespread use of custom-built hardware and software for studies in the early 2000s (e.g., Jessiman, 2003). Because of recent growth in the use of tablets and smartphones by parents during telehealth services, Snodgrass et al. (2017) proposed a telehealth infrastructure with mobile devices to implement communication interventions via parent coaching telehealth services. Investigations are needed to determine the contexts where widely available telehealth technology both improves access and is sufficient to provide effective services.

Speech Sound Assessment Scoring

Given that speech sound production disorders are prevalent on SLPs' pediatric caseloads, it is important to determine whether reliability of scoring articulation tests for diagnostic purposes is jeopardized in a variety of telehealth conditions (ASHA, 2016b; Grogan-Johnson et al., 2013; Hall-Mills et al., 2021). ASHA's 2010 school survey reported 92% of a typical school-based SLP's students consist of children with articulation impairments, and ASHA's 2020 school survey noted that 89% of SLPs were providing interventions to children with speech sound disorders (ASHA, 2018, 2020d, n.d.b.). Thus, effective assessments of children's speech sound production are critical for SLPs to reliably identify children with speech disorders. However, the task of scoring speech sound production depends on an SLP's ability to use visual and auditory input to judge how a child articulates individual speech sounds (Hall-Mills et al., 2021). This is perhaps more challenging for SLPs when using synchronous videoconferencing than in-person administration of picture stimuli. SLPs must be able to ensure factors such as audio and video quality do not negatively affect their ability to accurately record children's responses. For example, barriers such as inconsistent broadband connections and technology failures could potentially call into question the feasibility of conducting speech sound assessments with valid results (Freckmann et al., 2017; Hall-Mills et al., 2021; Lincoln et al., 2015). Behaviors such as a child turning away from the device's camera and microphone could impede an SLP from accurately scoring a child's response. Thus, it is important to investigate whether widely used articulation assessments can be scored reliably via telehealth (ASHA, 2020c; Taylor et al., 2014).

Although SLPs commonly use picture-based tests to evaluate children's articulation skills (Madison et al., 1982), there is inadequate evidence demonstrating the reliability of scoring a standardized speech assessment administered remotely (Taylor et al., 2014). For example, Waite et al. (2006) compared the scoring of the Single Word Articulation Test in videoconferencing and face-to-face scoring conditions. Even though this study reported a high level of agreement, they only had two assessors who used custom-built telehealth platforms to evaluate just six children. Eriks-Brophy et al. (2008) compared remote to on-site scoring agreement on the GFTA-2. In contrast to Waite et al. (2006), their findings revealed high levels of scoring disagreement. However, in the Eriks-Brophy et al. (2008) study, they lacked a description of the telehealth equipment used and had a small sample size (n = 5). Moreover, the authors noted that difficulty with acoustic transmission and the absence of headphones and microphones could have negatively affected their scoring.

Because of the high prevalence of speech sound disorders on SLPs' pediatric caseloads, it is important to be able to identify and diagnose speech sound disorders accurately to avoid jeopardizing the long-term viability of offering diagnostic and treatment telehealth services for speech sound disorders. The results of standardized tests are part of the diagnostic procedure that SLPs use to make evidence-based decisions about the identification of speech and language disorders. These decisions inform the development of interventions and progress monitoring of the services a child receives (Daub et al., 2021). Even though there is research to support therapy interventions delivered through synchronous videoconferencing, the lack of evidence to support remote scoring of standardized speech and language assessments represents a significant deterrent to implementing diagnostic services, putting the future of widespread speech-language telehealth use at risk.

Purpose

Speech-language assessments are composed of two dimensions: administration and scoring. Investigating variations in both dimensions simultaneously presents considerable challenges, and how to prioritize the dimensions will vary for different types of assessments. To be able to apply findings in the real world, the telehealth infrastructure must be accessible to both SLPs and clients. The purpose of this research is to investigate scoring reliability under live versus two telehealth scenarios. To evaluate scoring conditions that mimic the parameters that are feasible for practicing SLPs, we devised two conditions that might be considered typical or enhanced scenarios based on the technology used to deliver the acoustic signal to remote SLPs (McGill et al., 2021; Rauwerdink et al., 2019; Sutherland et al., 2016, 2017, 2018; Taylor et al., 2014).

SLPs' opinions and perceptions toward using telehealth represent potential barriers to care (Orlando et al., 2019; Sutherland et al., 2016, 2017). Negative attitudes have prevented SLPs in the past from accepting remote delivery of speech and language therapy services as a viable option (Fong et al., 2021; McClellan et al., 2020). Although the coronavirus pandemic resulted in a widespread transition from in-person care to therapy services being delivered via synchronous videoconferencing, long-term sustainability of telehealth services depends on the attitudes of SLPs toward remote delivery of their services. Thus, we sought feedback from the clinicians who were blinded to the two remote scoring conditions while scoring but had the opportunity to score an articulation test under all three conditions.

In addition to a traditional, in-person delivery and scoring of the Goldman-Fristoe Test of Articulation–Third Edition (GFTA-3; Goldman & Fristoe, 2015), concurrent scoring was compared with one SLP using Wi-Fi connections and standard tablet transmission and to another SLP using an enhanced telehealth assessment condition (i.e., receiving client speech sound production from an external mic input). Thus, traditional in-person assessment, typical telehealth assessment, and telehealth assessment with mic enhancement were compared to address the following research questions:

1. Do mean interrater agreement percentages reveal differences in scoring in the typical and the enhanced telehealth assessment conditions versus the in-person condition?
2. Do standard scores of a speech sound assessment differ when scored in person, in a typical telehealth assessment condition, and in an enhanced telehealth assessment condition?
3. To what extent do child behavior or technical disruptions interfere with administration and scoring of a speech sound assessment?
4. Do SLPs' opinions or perceptions about the three conditions differ?

Method

Participants

SLPs

Six SLPs were recruited. All SLPs were licensed providers in the state of Florida and currently working in outpatient clinics. SLP participants' experience ranged from 10 to 43 years, inclusive of extensive experience evaluating children with speech sound disorders in school and clinical settings. All SLP participants had a minimum of 1 year experience providing therapy using telehealth technology. However, all six SLPs had no more than 1-year experience administering a standardized speech sound assessment via synchronous videoconferencing. SLP participants signed an informed consent prior to participation. Each SLP was required to be fully vaccinated for COVID-19 at least 3 weeks prior to participation.

Children

Children between ages 3 and 8 years were recruited via fliers distributed in West Central Florida. Child participants included typically developing children and children with speech sound deficits, inclusive of childhood apraxia of speech, dysarthria, developmental phonological disorder, and delayed articulation. None of the children had reported hearing loss. Both male and female children from different socioeconomic backgrounds were included. Children were included even if they were receiving speech-language therapy and were previously diagnosed with a speech disorder. Child participants were excluded if they were under the age of 3 years or were age 9 years or older. Children were excluded if they had a limited lexicon as the assessment required the child to spontaneously name presented pictures.

A total of 39 children, 14 females and 25 males, participated in the speech sound assessments. Their mean age was 5 years 10 months (*SD* = 1 year, 7 months). Parents completed a demographic questionnaire. Parents self-reported their child's race/ethnicity as White/Caucasian (84.6%), Black/African American (10%), Hispanic (3%), or more than one race (3%). There were three children who were dual language learners. Parents classified their income level as low socioeconomic (79%) or middle/high (21%) by indicating their income range on an optional demographic question. Of the 39 children in the study, 92% had been formally diagnosed with a speech sound disorder, with 90% of the children currently receiving speech-language services. Comorbid conditions included autism spectrum disorder (18%), childhood apraxia of speech (10%), and cerebral palsy (5%). None of the children had been diagnosed with a hearing loss. Signed institutional review board parental permission and child assent forms were required prior to participation.

Speech Sound Assessment

The GFTA-3 (Goldman & Fristoe, 2015) is a standardized test used for the clinical assessment of speech sound production of individuals ages 2 through 21 years 11 months. The Sounds-in-Words subtest of the GFTA-3 was used. The picture stimuli include 60 target words that contain the initial, medial, and final sound positions in words of 23 consonants as well as 15 consonant blends. The third edition of the test is offered in both digital and print formats. However, the development of the GFTA-3 did not include remote delivery when establishing its validity and reliability, which calls into question the option of digital administration and scoring. In contrast to many assessments conducted by SLPs, the GFTA-3 instructions for in-person administration make it especially conducive to remote delivery. Per the GFTA-3 manual, it is appropriate to administer the test in a home, clinical, or school setting. The testing environment, no matter the setting, is recommended to be in a comfortable, quiet, well-lit area where distractions are minimal (Goldman & Fristoe, 2015). These testing friendly environments are commensurate with what is needed for remote delivery to be successful (Werfel et al., 2021). Administration of the test items also is conducive to remote delivery. As noted in the GFTA-3 manual, the evaluating SLP is afforded multiple opportunities to allow the child to respond to a test item, allowing the SLP to readminister an item if the child is off-task, distracted, missed the target prompt, or if the administration is interrupted. SLPs can even go as far as creating their own prompts or cues to elicit the target word(s) (cf. Goldman & Fristoe, 2015). The flexibility to repeat and prompt test items during a remote administration creates an optimal scenario for scoring speech sound production.

Experimental Design

GFTA-3 was administered in person, whereas scoring was completed simultaneously across three scoring conditions. A within-subject group design was used to compare the SLPs' scoring of children's speech production responses under the three conditions: (a) in person, (b) synchronous videoconferencing with the child using a tablet device's built-in microphone, and (c) synchronous videoconferencing with the child using a tablet device with an external microphone. Thus, one in-person SLP participant and two SLP participants via Wi-Fi connections at remote locations concurrently scored child participants. The SLP evaluators were randomly assigned and responsible for scoring children in each of the three conditions. However, constraints to randomization included dates the SLP participants were available for each of the scoring conditions, dates and times child participants were available, and the child participants' attendance (i.e., cancellations) for scheduled assessments. The resulting distribution across conditions was roughly equal for four of the SLPs, but one SLP who was less available resulted in one SLP who was overrepresented for the in-person condition and one SLP who was overrepresented for telehealth scoring conditions.

Settings and Telehealth Infrastructure

Data from Campbell and Goldstein's (2021) *Telehealth Services: Pediatric Provider Survey* reported computers with a broadband, Wi-Fi connection were used frequently by most clinicians in school and clinic therapy settings, whereas children used computers almost as often as mobile devices (i.e., tablet and smartphone) with a broadband, Wi-Fi connection. Both clients and clinicians reported the typical telehealth setup consists of using the device's built-in speakers and microphone for their audio component. The most common hardware upgrades added to enhance sound quality were headphones with built-in microphones, with clinicians more likely to use headphones than clients. On the basis of the findings of the survey, the typical and current, real-world technological infrastructure for remote data collection was established and used during the telehealth assessment.

Clinic and Test Administration Setting

The speech sound assessments were conducted in a standard (3 × 3 m) clinic room with the client and SLP sitting across from one another at a table. The in-person SLP always wore a standard Level 1 disposable face mask and sat behind a 24-in. trifold plexiglass barrier (see Supplemental Material S1 for a picture of the setup). Except for one child, all children did not wear a mask. The child who started testing wearing a mask opted to take it off early on during the testing. On the child's side of the barrier, two fifth generation Wi-Fi–enabled iPads (side by side) running IOS Version 14.5.1 were situated for simultaneously transmitting audio and video signals to the remote SLPs. The two iPads were placed in front of the child at a 120° angle with the cameras on, allowing an unobstructed view of the child's face for the in-person SLP. At the beginning of each testing session, the children greeted each remote SLP clinician, ensuring each child was aware that there were two additional SLPs observing their assessment. Children wore low-cost gaming headphones (Anivia AH28 Gaming Headset with Mic) during the testing and if requested, were allowed to take breaks from wearing them. The headphones were plugged into one of the two iPads. The GFTA-3 stimulus book was in front of the child at the top of the plexiglass, allowing the in-person SLP to turn the pages.

Each remote SLP used a desktop or laptop computer connected to a Wi-Fi at their remote location (i.e., at home and clinic). Both remote SLPs were wearing the same set of low-cost gaming headphones (Anivia AH28 Gaming Headset with Mic) throughout the testing.

For the two SLPs who scored the GFTA-3 via synchronous teleconferencing, the client's and SLP's devices were connected via Wi-Fi using the *Zoom for Healthcare* platform (Zoom Video Communications, Inc., 2021). Before the child's testing was begun, Wi-Fi speeds were verified and recorded on both the client and SLPs side using an online speed test (https://www.speedtest.net). The minimum requirement for connectivity using the Zoom platform is 600 kbps/1.5 Mbps (up/down), but the minimum Internet speed of 20 Mbps was required for this study to minimize the occurrences of potential connectivity loss (Zoom Video Communications, Inc., 2021). Each iPad's microphone sound enhancements were disabled (i.e., "original sound" setting) before connecting to the platform, and the remote SLPs muted their microphones prior to testing (Zoom Video Communications, Inc., 2020).

The in-person and remote SLPs had smartphones available to communicate with each other during the assessment. This allowed the in-person SLP to verify that remote SLPs were ready to begin scoring, to report technical issues on the remote SLPs' and client's side, and to receive feedback at the end of testing for any items that needed to be readministered.

Procedure

SLP Training

SLP participants received training in the GFTA-3 testing administration and scoring requirements as well as narrow International Phonetic Alphabet transcription for consonant sounds. Additionally, SLPs reviewed all procedures to implement the assessment tool both in-person and via remote delivery.

The first step in the training process required the SLP participants to demonstrate adequate narrow International Phonetic Alphabet (IPA) phonetic transcription skills using a calibration test. The purpose of the calibration test was to determine a baseline level of transcription agreement that was consistent within and across SLP participants using the same recorded samples. While wearing headphones and using their own computers, SLPs listened to a video recording of two children with speech disorders each responding to an assessment tool that targeted 34 consonant sounds. The assessment, created with words following the Moving Across Syllables: Articulatory Sound Movement Sequence (Kirkpatrick et al., 1990), included only one-syllable words. SLPs transcribed the children's speech production responses, and transcriptions were scored as correct or incorrect, based on the exact agreement with the child's in-person score. To be eligible to participate in this study, SLP participants were required to obtain a combined exact agreement score of 90% or higher for the total consonant sounds recorded between both children (Oller & Ramsdell, 2006; Preston et al., 2011; Shriberg et al., 1997). All six SLP participants met the participation criterion, obtaining a score of either 61/68 (90%) or 62/68 (91%).

Second, SLP participants reviewed GFTA-3 test administration procedures and data collection tool. Procedures specified in the GFTA-3 manual's instructions were reviewed, and SLPs were instructed how to clearly document any deviations from the test administrations procedures. In addition to scoring the phonemes in the GFTA-3, they were asked to note any child behavior and technology disruptions for each test item on the data collection tool.

Third, because the SLPs were taking turns administering the GFTA-3 in person, they needed to follow the protocol for setting up the test environment in the clinic setting. This included safety and cleaning procedures, including use of personal protective equipment (PPE) and barrier, and equipment setup. Training sessions detailed procedures for turning on, setting up, and positioning the devices; checking, verifying, and documenting the audio, video, and broadband speed on the client's side; logging into and setting up the teleconferencing platform (Zoom for Healthcare); setting up the picture stimuli book; placing the headphones on the child; and troubleshooting technology difficulties.

GFTA-3 Administration

The in-person scoring condition followed traditional administration. The Sound-in-Words testing for each child was completed in a single visit using standard procedures specified in the GFTA-3 manual (Goldman & Fristoe, 2015). The child was sitting across the table from the in-person SLP. The in-person SLP presented the target pictures to elicit speech production. If the child was unable to name the target item, or if the SLP's prompting did not elicit target word, the in-person SLP provided the verbal stimuli provided in the GFTA-3 picture book. Additionally, the in-person SLP managed any child disruptions and addressed technical disruptions.

Both remote SLPs wore gaming headphones connected to a laptop computer. However, one remote SLP received the child's responses directly from the child's microphone connected to the child's headset and the other remote SLP received the child's responses through the iPad's built-in microphone. They were blind to the condition they were in, as they were not able to see if their audio signal came from the external or built-in microphone.

Even though the child was wearing headphones, each child was able to sufficiently hear the in-person SLP present the test item stimuli. This was aided by the child's microphone picking up the in-person SLP's voice and sending the amplified signal directly to the child's headphones.

GFTA-3 Scoring

The scoring of the child's speech sound production was in person and synchronous; responses were not recorded and scored asynchronously. The child was sitting directly in front of two iPads and, thus, in view of the remote SLPs as well as the in-person SLP. The in-person SLP presented and scored each test item, whereas the remote SLPs only scored the child's responses. Each SLP recorded and, using narrow IPA transcription, phonetically transcribed the attempts at test item responses, even if one was considered unscorable due to technical issues (see Supplemental Material S1). Upon completion of the test, the SLPs scoring in the telehealth condition were able to request from the in-person SLP any test items that needed to be readministered. This could be due to a child behavior disruption, technology disruption, or difficulty scoring the item upon its presentation. Only final responses provided by the child were used to score the test item.

Data Collection

Prior to beginning the administration of the GFTA-3, each SLP recorded the computer that they were using and their Internet speed on their data sheet. SLP participants' computers included three MacBook Airs, a Lenovo desktop, a Lenovo Flex laptop, and a Hewlett Packard laptop. Before the evaluation began, the remote SLPs reported to the in-person SLP if there were any difficulties with setting up the session from the telehealth side, such as difficulty logging in, inadequate connectivity, or audio/visual difficulties. All technology deficiencies were resolved before testing could be initiated. Once the assessment began, the SLP participants recorded each sound production on the GFTA-3 data collection protocol form. Incorrect responses were transcribed phonetically, ensuring each SLP participant identified the error type produced by the child participant.

During the assessment, a data collection protocol form was also used to record child and technology disruptions. Upon completion of a child's testing session, each SLP transferred their phonetic transcription from the data collection protocol form to the Sounds-in-Words section of the GFTA-3 protocol. SLPs hand scored the total target speech sound errors to obtain an overall raw score. The completed GFTA-3 protocols were retallied for verification of calculated results. The first author reviewed any identified discrepancies, and then the SLPs resolved any mistakes in extrapolating the standard score. As an additional procedure to verify the accuracy of testing results, de-identified speech sound error data were manually entered into the Q-global (https://qglobal.pearsonclinical.com) web-based, The Health Insurance Portability and Accountability Act of 1996–compliant scoring program to extrapolate raw score totals and standard scores. Results that yielded disagreements between the hand and electronic scoring were checked for data entry errors in Q-global and calculation errors from hand-scoring. Identified errors were corrected before advancing to the final score validation step. The de-identified speech sound error data were entered into an Excel spreadsheet created to evaluate the individual item agreement between scoring conditions. Any disagreements between tallied scores in Excel and electronic raw scores were checked for data entry errors, and all disagreements were corrected. The final, verified raw, and standard scores were used for data analysis (see Supplemental Material S1 for data collection forms).

Telehealth Perception Questionnaire

At the end of the study, SLP participants completed postassessment questionnaire (see Supplemental Material S1). SLPs reflected on their experiences and general satisfaction with scoring a speech sound assessment via telehealth. Additional questions included SLP participants' awareness of the differences in the audio quality in the telehealth conditions, the degree to which their judgments of speech sounds were affected by telehealth use, and when sources of audio differences were evident. The evaluators rated 10 questions on an analog scale, ranging from *strongly disagree* to *strongly* agree reflected in a 0–100 numerical representation. Additionally, one yes/no question and three open-ended questions were asked.

Data Analysis

GFTA-3 – Interrater Agreement: Individual Speech Sounds

Interrater reliability was calculated for all speech sound data. Reliability was calculated by dividing the number of agreements by the total number of ratings. The average item-level agreement percentages among the three scoring conditions were reported. Additionally, severity ratings (i.e., average, mild, moderate, and severe) were derived from standard scores. Severity ratings were then compared to determine severity rating agreement percentages among scoring conditions.

GFTA-3 – Interrater Agreement: Standard Scores

GFTA-3 scoring results were analyzed using JMP 15.2 statistical software. A Bland–Altman analysis was used to evaluate agreement among standard scores on GFTA-3 for the three scoring conditions. The Bland–Altman (Bland & Altman, 1986, 2010) graphical display is a frequently used and preferred technique when investigating the agreement between two or more different methods on the same measure. The paired *t*-test summary statistic is presented in association with the Bland–Altman difference plot for method comparison, with the difference between the paired measures plotted on the graph's *y* axis. Paired-samples *t* tests were completed to compare (a) in-person versus telehealth with built-in microphone, (b) in-person versus telehealth with external microphone, and (c) telehealth with built-in microphone versus telehealth with external microphone. Because face-to-face scoring is the gold standard, the Bland–Altman (1986, 2010) analysis was used to determine if responses differed reliably among the scoring conditions. The confidence intervals from the GFTA-3 established the a priori confidence interval to determine if the mean differences in the standard scores were within the interval limits. Results were graphed to display any differences that existed. The mean differences between the three scoring conditions were assessed for skewness to verify the assumption of normality.

GFTA-3 – Standard Score Differences

GFTA-3 scoring results were analyzed using JMP 15.2 statistical software. A repeated-measures analysis of variance (ANOVA) was performed to evaluate mean differences between the three scoring conditions. Cohen's *d* was performed to evaluate the effect sizes associated with scoring conditions. Finally, a chi-square analysis with a Fisher's exact test was performed to determine if there was an association between scoring condition and speech sound severity classification.

Disruption Scoring

Descriptive statistics were used to summarize the data collected for child-related disruptions and technology-related disruptions. This included the percentage of test items that were completed without any participant or technology (i.e., hardware or platform) issues or problems with connectivity (i.e., broadband connection or Wi-Fi connection). Results were classified as technology or child-related disruptions and were summarized using descriptive statistics. Child-related disruptions distributed across age and disorder classification (i.e., apraxia) were summarized.

Telehealth Perception Questionnaire

The SLP participants rated the questions on the Telehealth Perception Questionnaire using an analog scale, ranging from strongly disagree to strongly agree reflected in a 0–100 numerical representation. Ratings were converted into percentages and mean scores calculated for each item. Responses to open-ended questions about SLP participants' impressions about telehealth assessments were reported.

Results

Telehealth Scoring Results

Interrater Agreement: Individual Speech Sounds

The GFTA-3 Sounds-in-Words test evaluates 141 sounds in words, with multiple chances for the child to produce each target speech sound (Goldman & Fristoe, 2015). Each item was analyzed for interrater agreement or disagreement. Percentage of agreement among the three scoring conditions was calculated. The mean item agreement for Live versus Typical was 86.3% (*SD* = 5.65), Live versus Enhanced was 86.7% (*SD* = 5.56), and Typical versus Enhanced was 85.2% (*SD* = 5.58).

The results were reviewed to identify the sources of scoring disagreements (Preston et al. 2011). Twenty-one of the 141 total items had less than 80% agreement among all three scoring conditions. Those disagreements occurred for only eight sounds, many of which were assessed multiple times in the same position of a word, such as the final /l/ (5 times). Table 1 presents the speech sounds and the position the sound occurred within the word, as well as the agreement percentages.

Interrater Agreement: Standard Scores

The Bland–Altman (Altman & Bland, 1983; Bland & Altman, 2010) analysis was used to evaluate the continuous variable agreement of GFTA-3 standard scores across the three scoring conditions (see Figure 1). A score of zero indicates perfect agreement between two conditions, and a larger number for each plot point indicates

Table 1. Scoring agreement across all three scoring conditions.

Speech sound	Position	Live/Typ	Live/Enh	Enh/Typ
1	Final	0.64	0.69	0.49
1	Final	0.64	0.62	0.56
1	Final	0.75	0.62	0.62
1	Final	0.49	0.69	0.54
1	Final	0.67	0.74	0.77
1	Medial	0.69	0.80	0.69
p	Initial	0.77	0.72	0.80
d	Final	0.72	0.80	0.67
ŋ	Final	0.56	0.49	0.54
ŋ	Final	0.69	0.69	0.59
ŋ	Final	0.77	0.77	0.69
tr	Initial	0.77	0.80	0.77
fr	Initial	0.77	0.74	0.77
gr	Initial	0.77	0.80	0.77
r	Initial	0.72	0.80	0.67
r	Medial	0.69	0.74	0.74
g	Final	0.72	0.72	0.63
t	Final	0.67	0.54	0.56
z	Final	0.62	0.56	0.74
z	Final	0.56	0.59	0.67
z	Final	0.77	0.51	0.69

Note. Live = in-person condition; Ehn = enhanced condition; Typ = typical condition.

greater disagreement between the scoring conditions. The zero (on the *y* axis), perfect agreement, is represented by the solid black line in the middle of each graph. A calculation of 95% limits of agreement (LoAs; mean difference ± 1.96 *SD* of the difference) were derived for each set of comparisons (represented by the upper and lower dotted green lines) with confidence limits for upper and lower LoAs considered as a pair. These confidence limits have been included as the upper and lower gray shaded areas around the LoAs in each figure panel. LoA confidence intervals indicate that at least 95% of population differences lie inside the limits $\bar{d} \pm c_{t0.025}\ s_{diff}$ and outside the limits $\bar{d} \pm c_{t0.975}\ s_{diff}$ (Zou, 2013). Importantly, data points outside these confidence limits give an indication of the frequency of significant disagreements. The Bland–Altman plot (see Figure 1) did not reveal a trend in the difference ($\bar{d}$) between the scoring conditions, with mean bias ranging from standard scores of only −1.79 to 1.0, as indicated by the red line being close to the solid black line (zero). This red line of equality for each scoring condition fell within the 95% confidence interval, middle gray box, of the mean differences, indicating there is not a significant systematic difference of one condition over- or underestimating the second condition. Likewise, paired-samples *t* tests showed no significant difference between standard scores for Live and Typical, $t(38) = -0.56$, $p = .58$; Live and Enhanced, $t(38) = 0.72$, $p = .47$; and Enhanced and Typical, $t(38) = -1.18$, $p = .24$. Skewness of standard score mean differences ranged from −.57 to .08, indicating a normal distribution (George & Mallery, 2016). When comparing the pairs of conditions, there were only one or two standard score outliers outside the upper and lower gray shaded areas; because those rare instances of standard score differences ranged from 25 to 27 points, they represent differences likely to affect severity ratings but were not systematically associated with a particular condition.

GFTA-3 Standard Scores Differences

The standard score distributions for each scoring condition were normally distributed. The GFTA-3 mean standard scores by scoring condition were 62.64 (*SD* = 19.21) for Live, 61.64 (*SD* = 18.53) for Typical, and 63.44 (*SD* = 18.68) for Enhanced. A repeated-measures ANOVA was performed to evaluate mean differences between the three scoring conditions. Results demonstrated no significant main effect by scoring condition, $F(2, 37) = 0.69$, $p = .51$. Cohen's *d* ranged from 0.09 to 0.19, suggesting that there were minimal effects associated with scoring conditions.

Speech Sound Disorder Standard Score Severity Classification

Standard scores of Live versus Typical and Live versus Enhanced scoring conditions were compared to identify incidences of study participants whose standard scores would result in different severity classifications (Shriberg et al., 2010). The GFTA-3 (Goldman & Fristoe, 2015) classifies severity ratings based on standard scores as average/above average above 85, mild/at-risk −1 *SD* (between 78 and 85), moderate −1.5 *SD* (between 71 and 77), and severe −2 *SD* or lower (70 or below). The distribution of severity of classifications comparing conditions is presented in Table 2. A chi-square analysis with a Fisher's exact test suggested that there was no association between scoring condition and speech sound severity classification ($p = .92$). There was almost perfect agreement for moderate and severe speech sound disorder classifications for the Live versus Typical and Live versus Enhanced scoring conditions. However, there were more discrepancies for mild and average classifications for both sets of comparisons.

To assess if the telehealth scoring conditions could accurately classify a speech sound disorder (i.e., greater than one standard deviation below the mean), agreement with the in-person scoring condition was calculated. Of the 39 child participants, the Live scoring condition classified 33 children with a speech sound disorder, and six with average speech sound production. The Typical scoring condition classified 32 children as having a speech sound disorder and seven children with average speech sound production, that is, 97% agreement with Live scoring. The Enhanced scoring condition classified 36 children as having a speech sound disorder and three children with average speech sound production, that is, 92% agreement with Live scoring.

Figure 1. Bland–Altman plot analysis by scoring condition. Bland Altman scatterplots illustrating interrater reliability and agreement for each pair of scoring conditions. The red lines above or below 0 show the small differences in means between conditions. For each pair of conditions, the y-axis reflects differences in standard scores, and the x-axis reflects the mean standard scores, per participant. Live = in-person condition; Enh = enhanced condition; Typ = typical condition; SS = standard score.

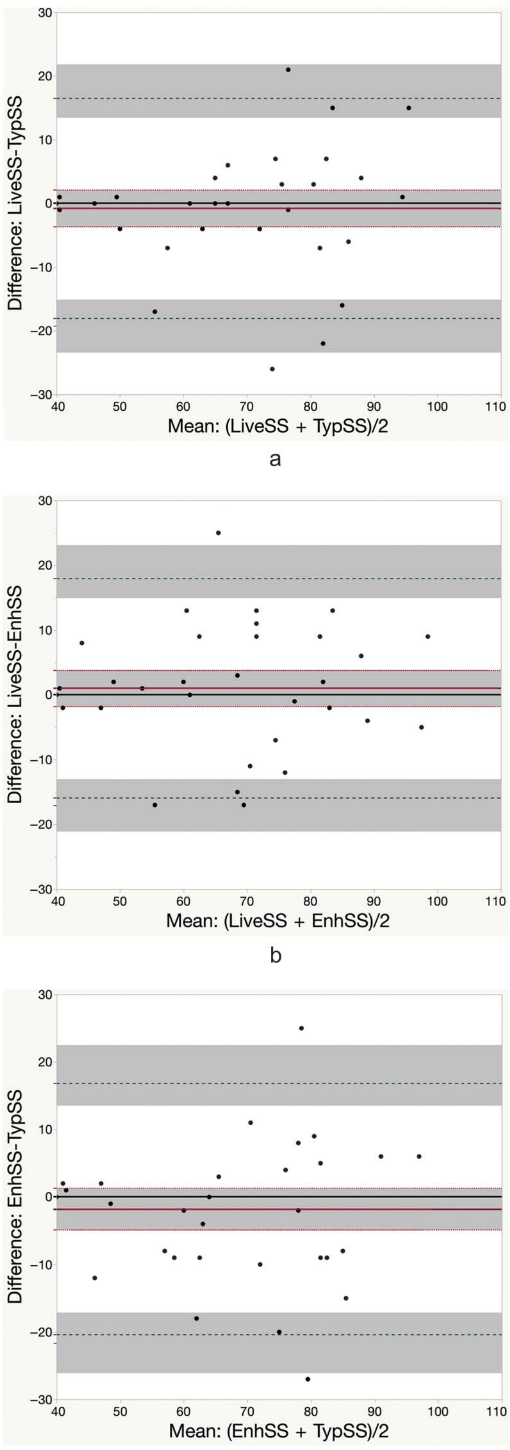

Table 2. Speech sound disorder severity classification by scoring condition.

Scoring condition	Average	Mild	Moderate	Severe	Total
Live	6	4	4	25	39
Typ	7	3	5	24	39
Enh	3	7	4	25	39

Note. Standard score severity classifications—average/above average above 85, mild/at-risk −1 *SD* (between 78 and 85), moderate −1.5 *SD* (between 71 and 77), and severe −2 *SD* or lower (70 or below). Chi-square likelihood ratio = 2.09, p = .91. Live = in-person condition; Enh = enhanced condition; Typ = typical condition.

Technical and Child Behavior Disruptions

Child Disruptions

A parent or caregiver was seated behind the child and available to assist the SLP during test administration for 15% of child participants. The in-person SLPs came out from behind the plexiglass barrier to manage child behavior (i.e., climbing out of their chair, touching the iPad screens, and not attending to picture book) for 21% of the child participants. During the testing procedure, SLPs documented when a child's behavior impeded their ability to score a test item. Children with behaviors impeding sound scoring were distributed across all ages and disorder types. Half of all participants (n = 19) were reported as having at least one incident where the child's behavior impeded the SLPs' ability to initially score one or more test items.

Technology Disruptions

During the testing procedure, the three SLPs documented when the telehealth technology impeded their ability to score a test item. Of the 39 assessments, 18 of them (45%) were impacted, at least minimally, by the technology used to perform them. Technology disruptions included screen freezing for less than 5 s (6 times), audio compromised (e.g., intermittent buzzing sounds; 6 times), and complete loss of transmission (6 times). Six more assessments had their starts delayed due to technology issues at their onset: signal strength was under the study protocol, complete audio/visual loss, or total transmission loss. These technology issues were able to be resolved except for one assessment. For that assessment, the remote SLP had to be changed due to the remote SLP being unable to establish a stable connection to participate.

At the end of each assessment, the remote SLPs were given the opportunity to contact the in-person SLP via text messaging and request test items be readministered. This would allow the remote SLPs the opportunities to score items that may have been previously marked unscorable during the test administration due to child behavior or technology disruptions. On average, six items were asked to be repeated by each remote SLP, with a minimum of two items and a maximum of 15 having been requested. Because of the test's flexibility in allowing test items to be repeated and prompted during administration, final scoring was not affected by child or technology disruptions.

Telehealth Perception Questionnaire

Upon completion of the study, the six SLPs were asked to reflect on their experiences in each of the three scoring conditions. Results of the postassessment survey are in Table 3. The SLPs gave high ratings for the directions for using, scoring, and administering a speech sound disorder assessment remotely, with mean percentages of 92% and 98%, respectively. However, SLP participants' mean ratings were only 41% and 42%, respectively, when asked about their enthusiasm and motivation for administering and scoring a standardized speech assessment remotely. SLP participants reported their preferred method of delivering a speech sound assessment as in person, with a mean percentage of 85%. In response to the open-ended questions, SLP participants reported five reasons why they did not prefer remote administration of a standardized speech assessment: (a) lack of client's telehealth infrastructure, (b) more response repetitions needed for accurate scoring, (c) difficulty managing a child's behaviors remotely, (d) lack of family support, and (e) home environments not conducive to telehealth.

Table 3. Speech-language pathologist (n = 6) postassessment questionnaire.

Survey questions	*M*	
I would be enthusiastic to use telehealth to complete a standardized speech assessment.	41%	
I understand how to administer a speech sound disorder evaluation remotely.	98%	
Remote delivery is a reliable way to administer speech sound disorder evaluations.	50%	
I am motivated to use this delivery method for speech sound evaluations.	42%	
My preferred method of administrating a speech sound assessment is		
a. Typical: in person	85%	
b. Remote delivery	15%	
Telehealth is an effective choice for evaluating children with speech sound disorders.	39%	
The directions for using remote delivery of a standardized speech sound assessment are clear.	92%	
The amount of time required to perform a speech sound assessment remotely is reasonable.	82%	
The amount of time required for record keeping with this evaluation format is reasonable.	76%	
Implementation of an evaluation delivered remotely would require support from family members.	95%	
Were you aware of the differences in the audio quality in the telehealth scoring conditions?	Yes 17%	No 83%
Reasons I would prefer to administer a speech sound assessment remotely: *Illness, distance and if in-person assessments were not an option*		
Reasons I would not prefer to administer a speech sound assessment remotely: *Child's behaviors, lack of client telehealth infrastructure, more response repetitions needed, lack of family support, home environments not conducive to telehealth*		
The future of teleassessments: *Will continue to grow, will be contingent upon SLP attitudes and advancements in research demonstrating the validity and reliability of evaluations delivered remotely*		

Note. Rating percentages reflected in a 0–100 numeric representation, ranging from *strongly disagree* (0) to *strongly agree* (100). SLP = speech-language pathologist.

Discussion

In summary, this study revealed high agreement among the three scoring conditions. This indicates that typical and enhanced telehealth scoring both could be deemed equally reliable to in-person scoring. The mean GFTA-3 standard scores of all three scoring conditions had mean differences of two or less, demonstrating that there was not a significant systematic difference of one scoring condition over- or underestimating the second scoring condition, as noted on the Bland–Altman plot (see Figure 1). In addition, all scoring conditions were found to be highly correlated, demonstrating a strong relation between in-person and telehealth scoring, both in the typical and enhanced scenarios. Both telehealth scoring conditions allowed SLPs to evaluate a child's speech sound production and derive results that were similar to in person, supporting the validity of remotely scoring a speech sound assessment. Although limited, previous research that compared in person scoring of a speech sound assessment to remote delivery also found the two scoring conditions to yield virtually equivalent test results (Taylor et al., 2014; Waite et al., 2010). This study's findings are consistent with prior findings and add to information contrasting in-person scoring with remote scoring but, in this case, using widely available, consumer-grade technology.

The overall SLP participants' scoring agreement for individual speech sounds among scoring conditions was high, with small decrements when compared with baseline calibration scoring agreement, which ranged from 90% to 91% (Preston et al., 2011). In contrast to the baseline scoring, the scoring agreement comparing the three conditions ranged from 85% to 87%, indicating that the variation associated with the different acoustic conditions amounted to a 3%–6% attenuation in scoring agreement. Moreover, telehealth scoring conditions yielded reliable results (> 80% agreement between in-person and telehealth scoring conditions) for all but eight speech sounds (i.e., l, r, ŋ, z, d, g, t, and p). As the GFTA-3 provides multiple chances for the child to produce each target item, these eight sounds resulted in 21 out of the 141 target items with less than 80% agreement. It is worth noting that final sounds and liquids are especially difficult to score reliably. As a result, SLPs administering a telehealth assessment will need to pay particular attention when scoring these individual speech sounds to verify correct production and, most likely, require the child to repeat production of words containing these sounds to score the items accurately. The disagreement among scoring specific, individual sounds via telehealth found in this study are consistent with previous telehealth literature reviews, noting SLPs have difficulties identifying correct speech production for sounds involving articulators that are difficult to see (e.g., r and g) and cognate pairs (e.g., t and d; Eriks-Brophy et al., 2008; Taylor et al., 2014).

Furthermore, a closer examination of standard score severity classifications was virtually identical across conditions for severe and moderate speech sound disorders. The discrepancies seen were mainly in distinguishing average versus mild disorders. However, the bias for both sets of scoring conditions was small (less than two points) and not clinically significant, indicating that the in-person and telehealth scoring conditions produced virtually identical results. Importantly, SLPs do not rely on severity classifications alone or a single test score when determining if a child has a speech sound disorder. An SLP's clinical judgment must be considered; SLPs should use tests as one tool within a diagnostic framework (Daub et al., 2021). For example, a difference in severity classification (e.g., mild vs. moderate) or GFTA-3 standard score above 85 does not imply that a child would not be identified as having speech sound disorder. To the contrary, when a remote SLP uses the GFTA-3's test administration guidelines to derive a standard score, whether in-person or remotely, the additional consideration of an SLP's own clinical judgement would influence clinical decision making when determining the presence of a speech sound disorder. For example, if a child's standard score falls in the average range on the GFTA-3 but the student is lateralizing /s/ or /ʃ/, an SLP recognizes these types of errors are not part of typical speech sound development. As these articulation errors typically do not self-correct, an SLP could recommend speech therapy despite the child's standard score severity rating.

SLPs have expressed concerns that telehealth assessments were difficult to administer, as children with off-task behaviors were difficult to manage remotely, and that the technology needed to implement assessments often caused disruptions that could impede the scoring of test items (Campbell & Goldstein, 2021, Werfel et al., 2021). Therefore, this study investigated the frequency of child and technology disruptions and how they related to an SLP's ability to score a child's speech sound production. The SLPs needed to come out from behind the plexiglass barrier to manage child behavior or required parent assistance during the test administration for about one third of child participants. Even with the in-person SLP and parent to address any behaviors that arose during testing, half of all child participants had at least one incident where the child's behavior impeded the SLPs' ability to initially score one or more test items. However, these incidences of child behavior during a speech sound assessment did not mean that the overall ability for an SLP to accurately score speech sound items was compromised. It did, however, mean that the compromised test items needed to be readministered, often more than 3 times, for the SLP(s) to reliably score the prompted item. Thus, a speech sound assessment may require a longer testing session to administer remotely; however, scoring would not be affected. It should be noted, although, that having children repeat a test item, even 3 times or more, is not exclusive to a telehealth delivery method. The GFTA-3 scoring manual notes that children may not pay attention intermittently during in-person test administrations. Therefore, creating positive testing environments is of utmost importance no matter the speech assessment delivery method (i.e., quiet room that is well-lit with minimal distractions, adult physically present for the duration of the test administration). Moreover, the GFTA-3 manual's administration procedure, allowing a test prompt or verbal stimuli to be repeated multiple times, allowed the in-person and remote SLPs to score all target items, even with child and technology disruptions (Goldman & Fristoe, 2015). However, future research is needed to compare different telehealth assessment scenarios versus in-person assessment administration and not just scoring.

This study adhered to the telehealth infrastructure consistent with currently typical telehealth technology (Campbell & Goldstein, 2021). Therefore, it was not surprising to the SLP participants that almost half of the 39 assessments were impacted, at least minimally, by technology. Technological disruptions, such as a screen freezing or complete loss of transmission, are not uncommon during speech-language telehealth sessions (Campbell & Goldstein, 2021). The ability to resolve technology barriers can be frustrating and can even prevent a session from occurring at all. Six of the 39 assessments were at risk of cancelation due to technology disruptions. Fortunately, five of six were able to be resolved at the onset of the telehealth assessments, and the sixth one proceeded with a remote SLP replacement at a different location that had connectivity conducive to the study's bandwidth parameters. The technology disruptions experienced during this study demonstrated an ongoing barrier experienced with telehealth use (Campbell & Goldstein, 2021). Stable telehealth infrastructure is imperative for telehealth sessions, including assessments, to be successful.

Among the areas investigated in this study, the one that cannot be overlooked is the SLP participants' attitudes toward telehealth assessments, a critical component of future telehealth use. As SLPs' confidence in telehealth assessments have been associated with use, an SLP's general satisfaction with scoring a speech sound assessment via telehealth could affect the SLP's overall view of the telehealth assessment's administration (Taylor et al., 2014). The SLPs' responses on the postassessment questionnaire positively rated (76% or higher) clarity of the directions and ability to remotely score a speech sound assessment as well as the time needed to administer and score telehealth assessments. However, despite these positive ratings and the positive findings of this study, SLP participants did not perceive telehealth assessments to be a replacement for face-to-face administration. They rated their enthusiasm and motivation for teleassessments poorly (42% or lower). SLP participants' responses to the open-ended questions revealed how their previous and current experiences influenced their answers to the questionnaire. Concerns about client telehealth infrastructure, lack of family support, and difficulty managing client's behaviors as well as the current need for more advancements in telehealth research illustrated the ongoing barrier to widespread telehealth use—provider acceptance. The SLP participants acknowledged the value of telehealth assessments, even stating the use of telehealth will grow; however, their perceptions remained that telehealth is only for certain populations and prefer face-to-face administration of assessments. Even though in-person administration may be their preference, the findings of this study indicate that accurate scoring of a speech sound assessment remotely is not a barrier to an alternative method of service delivery for children.

Limitations

The study was performed with seasoned SLPs who had no less than 10 years of experience administering face-to-face speech sound assessments but only 1 year of experience with telehealth assessments. The postassessment survey with opinions about remote delivery of a speech sound assessment may have been different if SLP study participants included early career professionals.

All children were administered the test within a clinical setting while the remote SLP participants were scoring the assessment while at home. The outcomes of this study may not truly reflect scoring outcomes for children evaluated within their natural setting or SLPs providing services in changing work environments (e.g., working remotely from a school; bandwidth when using a cellular phone connection in contrast to Wi-Fi); this study does not reflect remote administration of an assessment where the test would be administered while the child is at a location other than a clinic setting (i.e., home, daycare, and library). Moreover, children were tested using the Zoom for Healthcare (Zoom Video Communications, Inc., 2021) platform on Apple iPads and computers; however, the findings of this study may not generalize to other devices, such as cellphones, or platforms, such as doxy.me. A future study could investigate in what remote delivery scenarios, if any, does the administration of the GFTA-3 have to be modified for telehealth, considering the current in-person administration is already favorable for remote delivery (e.g., ability to repeat testing items).

Last, even though children were tested in person, the clinical setting was still different than what was considered typical prior to the COVID-19 pandemic. The in-person SLP was wearing a mask throughout the test administration, something not typical prior to the pandemic. SLP participants had a plexiglass barrier between them and the child participant. They also were sitting a distance that was further back from the child than would be a typical distance between child and SLP prepandemic. It is unknown how the presence of the plexiglass barrier or the distance between SLP and child participant affected the in-person SLP's ability to score the child's speech sound production.

Conclusions and Future Research

The COVID-19 pandemic created a whole new generation of pediatric telehealth speech-language pathology providers (Campbell & Goldstein, 2021). These SLPs discovered the benefits of telehealth, clinically managing speech and language disorders without compromising their health or the health of their clients. Even though there was broad support for remote delivery of services during the pandemic, there remains significant skepticism among SLPs and policymakers about the use of telehealth to deliver evaluative or diagnostic services. Given the limited research on this topic, this study sought to add to the body of research on the reliability of telehealth assessments.

Given the scenarios examined, the overall results were not affected in any systematic fashion. In fact, exact agreement percentages were negligibly worse than the calibration exact agreement percentages, 85%–87% versus 90%–91%. The findings of this study demonstrate that the in-person, typical telehealth, and enhanced telehealth scoring conditions produce nearly identical classification results. The disagreements tend to be between normal and mild impairments. This speaks to the need for clinical judgment to factor into final decisions about whether a child has a speech sound disorder. Other attributes, such as errors that are not developmental in nature or speech intelligibility that affect a child's education or socialization must be taken into account.

This study evaluated children as young as age 3 years and children who demonstrated behaviors that required adult intervention. In addition, the telehealth assessment procedure used commercial-grade equipment to provide the speech sound assessment remotely. As such, both child behaviors and technology disruptions were reported as to their effect on item scoring. Even though half of all assessments reported at least one incidence of child behavior and/or technology disruption initially impeding the SLPs' ability to score one or more items, individual agreement among test items was high, with only eight sounds demonstrating scoring agreements below 80% across all three conditions. Not surprisingly, these were sounds mostly in the final position of words (e.g., /l/), that were difficult to see (e.g., /g/) or required voicing (e.g., /z/) or absence of (e.g., /p/) for scoring. However, the GFTA-3 (Goldman & Fristoe, 2015) Sounds-in-Words test administration has allowances for test items to be repeated. With the ability to repeat test items and provide additional verbal stimuli to elicit a child's response, an SLP is afforded ample opportunities to score a response item accurately, no matter if impeded by a child's behavior, technology disruption, or difficulty scoring specific speech sounds remotely. Future telehealth assessments should investigate diagnostic tests whose administration are not as accommodating with their assessment items, that is, questions cannot be repeated, or item prompts cannot deviate in their presentation.

ASHA expressed hesitancy about conducting standardized evaluations remotely, as modifications to test delivery could impact interpretation of scores, potentially requiring children to be reassessed through in-person administration to acquire valid results (ASHA, 2020c; Campbell & Goldstein, 2021, 2022; Freckmann et al., 2017). Even though SLPs have been leery about the accuracy of scoring speech sound assessments via telehealth, our results indicated no advantage among the three scoring conditions and did not affect the overall judgement of SLPs' ability to score children's speech sound production. As the telehealth scoring conditions did not require modifications for delivery, there would be no reason to repeat a speech sound assessment live. Rather, SLPs should feel confident accurately scoring a child's speech sound production remotely, using consumer-grade equipment.

Future studies are needed to evaluate current pediatric telehealth assessments for a variety of conditions, such as pediatric language disorders and deficits in phonological awareness. The SLP participants, despite their involvement in this study, reported attitudes that continued to question the use of telehealth for evaluative and diagnostic services. The SLP participants' telehealth views demonstrate an ongoing barrier of widespread telehealth use. Until more studies are completed demonstrating the reliability of telehealth assessments comparable to in-person administration, the attitudes and views of SLPs may remain unchanged. For this reason, it would behoove current and future developers of pediatric speech-language assessments to include both delivery modalities—in person and remote—when establishing test validity, reliability, and effectiveness.

References

Altman, D. G., & Bland, J. M. (1983). Measurement in medicine: The analysis of method comparison studies. *Journal of the Royal Statistical Society: Series D (The Statistician), 32*(3), 307–317. https://doi.org/10.2307/2987937

American Speech-Language-Hearing Association. (2002). *Survey of telepractice use among audiologists and speech-language pathologists.*

American Speech-Language-Hearing Association. (2018). *Schools survey report: SLP caseload characteristics trends 2000–2018.* https://www.asha.org/siteassets/surveys/2018-schools-survey-caseload-trends.pdf

American Speech-Language-Hearing Association. (2016a). Code of ethics [Ethics]. https://www.asha.org/code-of-ethics/

American Speech-Language-Hearing Association. (2016b). *2016 Schools Survey: SLP caseload and workload characteristics.* https://www2.asha.org/uploadedFiles/2016-Schools-Survey-SLP-Caseload-Characteristics.pdf

American Speech-Language-Hearing Association. (2016c). *Scope of practice in speech-language pathology* [Scope of Practice]. https://www.asha.org/policy/sp2016-00343/

American Speech-Language-Hearing Association. (2020a). *ASHA COVID-19 survey results - March 2020.* https://www.asha.org/siteassets/uploadedfiles/COVID-19-Tracker-Survey-March-2020.pdf

American Speech-Language-Hearing Association. (2020b). *ASHA COVID-19 Survey results - May 2020.* https://www.asha.org/siteassets/uploadedfiles/COVID-19-Tracker-Survey-May-2020.pdf

American Speech-Language-Hearing Association. (2020c). *Considerations for speech, language, and cognitive assessment via telepractice.* https://www.asha.org/slp/clinical/considerations-for-speech-language-and-cognitive-assessment-via-telepractice/

American Speech-Language-Hearing Association. (2020d). *2020 Schools Survey: SLP caseload and workload characteristics report.* https://www.asha.org/siteassets/surveys/2020-schools-survey-slp-caseload.pdf

American Speech-Language-Hearing Association. (n.d.a). *School services, interrupted: What parents of students receiving speech and language treatment in schools should know during COVID-19 closures.* https://www.asha.org/public/school-services-interrupted-what-parents-of-students-receiving-speech-and-language-treatment-in-schools-should-know-during-covid-19-closures/

American Speech-Language-Hearing Association. (n.d.b). *Speech sound disorders: Articulation and phonology* [Practice Portal]. Retrieved November 24, 2020, from www.asha.org/Practice-Portal/Clinical-Topics/Articulation-and-Phonology/

Bashshur, R., Doarn, C. R., Frenk, J. M., Kvedar, J. C., & Woolliscroft, J. O. (2020). Telemedicine and the COVID-19 pandemic, lessons for the future. *Telemedicine and E- Health, 26*(5), 571–573. https://doi.org/10.1089/tmj.2020.29040.rb

Benda, N. C., Veinot, T. C., Sieck, C. J., & Ancker, J. S. (2020). Broadband internet access is a social determinant of health! *American Journal of Public Health, 110*(8), 1123–1125. https://doi.org/10.2105/AJPH.2020.305784

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet, 1*(8476), 307–310. https://doi.org/10.1016/S0140-6736(86)90837-8

Bland, J. M., & Altman, D. G. (2010). Statistical methods for assessing agreement between two methods of clinical measurement. *International Journal of Nursing Studies, 47*(8), 931–936. https://doi.org/10.1016/j.ijnurstu.2009.10.001

Campbell, D. R., & Goldstein, H. (2021). Genesis of a new generation of telepractitioners: The COVID-19 pandemic and pediatric speech-language pathology services. *American Journal of Speech-Language Pathology, 30*(5), 2143–2154. https://doi.org/10.1044/2021_AJSLP-21-00013

Campbell, D. R., & Goldstein, H. (2022). Evolution of telehealth technology, evaluations, and therapy: Effects of the COVID-19 pandemic on pediatric speech-language pathology services. *American Journal of Speech-Language Pathology, 31*(1), 271–286. https://doi.org/10.1044/2021_AJSLP-21-00069

Cason, J., & Cohn, E. R. (2014). Telepractice: An overview and best practices. *SIG 12 Perspectives on Augmentative and Alternative Communication, 23*(1), 4–17. https://doi.org/10.1044/aac23.1.4

Coufal, K., Parham, D., Jakubowitz, M., Howell, C., & Reyes, J. (2018). Comparing traditional service delivery and telepractice for speech sound production using a functional outcome measure. *American Journal of Speech-Language Pathology, 27*(1), 82–90. https://doi.org/10.1044/2017_AJSLP-16-0070

Daub, O., Cunningham, B. J., Bagatto, M. P., Johnson, A. M., Kwok, E. Y., Smyth, R. E., & Cardy, J. O. (2021). Adopting a conceptual validity framework for testing in speech-language pathology. *American Journal of Speech-Language Pathology, 30*(4), 1894–1908. https://doi.org/10.1044/2021_AJSLP-20-00032

Dekhtyar, M., Braun, E. J., Billot, A., Foo, L., & Kiran, S. (2020). Videoconference administration of the western aphasia battery-revised: Feasibility and validity. *American Journal of Speech-Language Pathology, 29*(2), 673–687. https://doi.org/10.1044/2019_AJSLP-19-00023

Eriks-Brophy, A., Quittenbaum, J., Anderson, D., & Nelson, T. (2008). Part of the problem or part of the solution? Communication assessments of Aboriginal children residing in remote communities using videoconferencing. *Clinical Linguistics & Phonetic, 22*(8), 589–609. https://doi.org/10.1080/02699200802221737

Freckmann, A., Hines, M., & Lincoln, M. (2017). Clinicians' perspectives of therapeutic alliance in face-to-face and telepractice speech-language pathology sessions. *International Journal of Speech-Language Pathology, 19*(3), 287–296. https://doi.org/10.1080/17549507.2017.1292547

Fong, R., Tsai, C. F., & Yiu, O. Y. (2021). The implementation of telepractice in speech language pathology in Hong Kong during the COVID-19 pandemic. *Telemedicine Journal and E- Health, 27*(1), 30–38. https://doi.org/10.1089/tmj.2020.0223

George, D., & Mallery, P. (2016). *IBM SPSS Statistics 23 Step by Step: A Simple guide and reference* (14th ed.). Routledge. https://doi.org/10.4324/9781315545899

Goldman, R., & Fristoe, M. (2015). *Goldman-Fristoe Test of Articulation–Third Edition*. Pearson.

Grogan-Johnson, S., Schmidt, A. M., Schenker, J., Alvares, R., Rowan, L. E., & Taylor, J. (2013). A comparison of speech sound intervention delivered by telepractice and side-by-side service delivery models. *Communication Disorders Quarterly, 34*(4), 210–220. https://doi.org/10.1177/1525740113484965

Hall-Mills, S., Johnson, L., Gross, M., Latham, D., & Everhart, N. (2021). Providing telepractice in schools during a pandemic: The experiences and perspectives of speech-language pathologists. *Language, Speech, and Hearing Services in Schools*, 1–17. https://doi.org/10.1044/2021_LSHSS-21-00023

Hao, Y., Zhang, S., Conner, A., & Lee, N. Y. (2021). The evolution of telepractice use during the COVID-19 pandemic: Perspectives of pediatric speech-language pathologists. *International Journal of Environmental Research and Public Health, 18*(22), 12197. https://doi.org/10.3390/ijerph182212197

Hill, A. J., & Miller, L. E. (2012). A survey of the clinical use of telehealth in speech-language pathology across Australia. *Journal of Clinical Practice in Speech-Language Pathology, 14*(3), 110–117.

Hodge, M. A., Sutherland, R., Jeng, K., Bale, G., Batta, P., Cambridge, A., Detheridge, J., Drevensek, S., Edwards, L., Everett, M., Ganesalingam, K., Geier, P., Kass, C., Mathieson, S., McCabe, M., Micallef, K., Molomby, K., Ong, N., Pfeiffer, S., . . . 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. https://doi.org/10.1177/1357633X18776095

Jessiman, S. M. (2003). Speech and languages services using telehealth technology in remote and underserviced areas. *Journal of Speech-Language Pathology and Audiology, 27*(1), 45–51.

Keck, C. S., & Doarn, C. R. (2014). Telehealth technology applications in speech-language pathology. *Telemedicine and E-Health, 20*(7), 653–659. https://doi.org/10.1089/tmj.2013.0295

Kertesz, A. (2007). *The Western Aphasia Battery–Revised (WAB-R)*. Pearson. https://doi.org/10.1037/t15168-000

Kichloo, A., Albosta, M., Dettloff, K., Wani, F., El-Amir, Z., Singh, J., Aljadah, M., Chakinala, R. C., Kanugula, A. K., Solanki, S., & Chugh, S. (2020). Telemedicine, the current COVID-19 pandemic and the future: A narrative review and perspectives moving forward in the USA. Family Medicine and Community. *Health, 8*(3). https://doi.org/10.1136/fmch-2020-000530

Kirkpatrick, J., Stohr, P., & Kimbrough, D. (1990). *Moving across syllables*. Communication Skill Builders.

Lincoln, M., Hines, M., Fairweather, C., Ramsden, R., & Martinovich, J. (2015). Multiple stakeholder perspectives on teletherapy delivery of speech pathology services in rural schools: A preliminary, qualitative investigation. *International Journal of Telerehabilitation, 6*(2), 65–74. https://doi.org/10.5195/IJT.2014.6155

Madison, C. L., Kolbeck, C. P., & Walker, J. L. (1982). An evaluation of stimuli identification on three articulation tests. *Language, Speech, and Hearing Services in Schools, 13*(2), 110–115. https://doi.org/10.1044/0161-1461.1302.110

McClellan, M. J., Florell, D., Palmer, J., & Kidder, C. (2020). Clinician telehealth attitudes in a rural community mental health center setting. *Journal of Rural Mental Health, 44*(1), 62–73. https://doi.org/10.1037/rmh0000127

McGill, M., Siegel, J., & Noureal, N. (2021). A preliminary comparison of in-person and telepractice evaluations of stuttering. *American Journal of Speech-Language Pathology, 30*(4), 1737–1749. https://doi.org/10.1044/2021_AJSLP-19-00215

Mohan, H. S., Anjum, A., & Rao, P. K. (2017). A survey of telepractice in speech-language pathology and audiology in India. *International Journal of Telerehabilitation, 9*(2), 69–80. https://doi.org/10.5195/ijt.2017.6233

Oller, D. K., & Ramsdell, H. L. (2006). A weighted reliability measure for phonetic transcription. *Journal of Speech, Language, and Hearing Research, 49*(6), 1391–1411. https://doi.org/10.1044/1092-4388(2006/100)

Orlando, J. F., Beard, M., & Kumar, S. (2019). Systematic review of patient and caregivers' satisfaction with telehealth videoconferencing as a mode of service delivery in managing patients' health. *PLOS ONE, 14*(8), e0221848. https://doi.org/10.1371/journal.pone.0221848

Preston, J. L., Ramsdell, H. L., Oller, D. K., Edwards, M. L., & Tobin, S. J. (2011). Developing a weighted measure of speech sound accuracy. *Journal of Speech, Language, and Hearing Research, 54*(1), 1–18. https://doi.org/10.1044/1092-4388(2010/10-0030)

Raatz, M., Ward, E. C., Marshall, J., & Burns, C. L. (2021). Evaluating the use of telepractice to deliver pediatric feeding assessments. *American Journal of Speech Language Pathology, 30*(4), 1686–1699. https://doi.org/10.1044/2021_AJSLP-20-00323

Rauwerdink, A., Chavannes, N. H., & Schijven, M. P. (2019). Needed: Evidence based eHealth! *Clinical eHealth, 2,* 1–2. https://doi.org/10.1016/j.ceh.2019.01.001

Shriberg, L. D., Austin, D., Lewis, B. A., McSweeny, J. L., & Wilson, D. L. (1997). The percentage of consonants correct (PCC) metric. *Journal of Speech, Language, and Hearing Research, 40*(4), 708–722. https://doi.org/10.1044/jslhr.4004.708

Shriberg, L. D., Fourakis, M., Hall, S. D., Karlsson, H. B., Lohmeier, H. L., McSweeny, J. L., Potter, N. L., Scheer-Cohan, A. R., Strand, E. A., Tilkens, C. M., & Wilson, D. L. (2010). Extensions to the Speech Disorders Classification System (SDCS). *Clinical Linguistics & Phonetics, 24*(10), 795–824. https://doi.org/10.3109/02699206.2010.503006

Smith, A. C., Thomas, E., Snoswell, C. L., Haydon, H., Mehrotra, A., Clemensen, J., & Caffery, L. J. (2020). Telehealth for global emergencies: Implications for coronavirus disease 2019 (COVID-19). *Journal of Telemedicine and Telecare, 26*(5), 309–313. https://doi.org/10.1177/1357633X20916567

Snodgrass, M. R., Chung, M. Y., Biller, M. F., Appel, K. E., Meadan, H., & Halle, J. W. (2017). Telepractice in speech-language therapy: The use of online technologies for parent training and coaching. *Communication Disorders Quarterly, 38*(4), 242–254. https://doi.org/10.1177/1525740116680424

Sutherland, R., Hodge, A., Trembath, D., Drevensek, S., & Roberts, J. (2016). Overcoming barriers to using telehealth for standardized language assessments. *Perspectives of the ASHA Special Interest Groups, 1*(18), 41–50. https://doi.org/10.1044/persp1.SIG18.41

Sutherland, R., Trembath, D., Hodge, A., Drevensek, S., Lee, S., Silove, N., & Roberts, J. (2017). Telehealth language assessments using consumer grade equipment in rural and urban settings: Feasible, reliable and well tolerated. *Journal of Telemedicine and Telecare, 23*(1), 106–115. https://doi.org/10.1177/1357633X15623921

Sutherland, R., Trembath, D., & Roberts, J. (2018). Telehealth and autism: A systematic search and review of the literature. *International Journal of Speech-Language Pathology, 20*(3), 324–336. https://doi.org/10.1080/17549507.2018.1465123

Taylor, O. D., Armfield, N. R., Dodrill, P., & Smith, A. C. (2014). A review of the efficacy and effectiveness of using telehealth for paediatric speech and language assessment. *Journal of Telemedicine and Telecare, 20*(7), 405–412. https://doi.org/10.1177/1357633X14552388

Tohidast, S. A., Mansuri, B., Bagheri, R., & Azimi, H. (2020). Provision of speech-language pathology services for the treatment of speech and language disorders in children during the COVID-19 pandemic: Problems, concerns, and solutions. *International Journal of Pediatric Otorhinolaryngology, 138,* 110262. https://doi.org/10.1016/j.ijporl.2020.110262

Tucker, J. K. (2012). Perspectives of speech-language pathologists on the use of telepractice in schools: Quantitative survey results. *International Journal of Telerehabilitation, 4*(2), 61–72. https://doi.org/10.5195/ijt.2012.6100

Wales, D., Skinner, L., & Hayman, M. (2017). The efficacy of telehealth-delivered speech and language intervention for primary school-age children: A systematic review. *International Journal of Telerehabilitation, 9*(1), 55–70. https://doi.org/10.5195/ijt.2017.6219

Waite, M. C., Cahill, L. M., Theodoras, D. G., Busuttin, S., & Russell, T. G. (2006). A pilot study of online assessment of childhood speech disorders. *Journal of Telemedicine and Telecare. 12*(3_suppl), 92–94. https://doi.org/10.1258/135763306779380048

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, 41*(4), 445–458. https://doi.org/10.1044/0161-1461(2009/08-0131)

Wang, C. J., Ma, J., Zuckerman, B., & Car, J. (2020). The opportunities for telehealth in pediatric practice and public health. *Pediatric Clinics of North America, 67*(4), 603–611. https://doi.org/10.1016/j.pcl.2020.03.001

Werfel, K. L., Grey, B., Johnson, M., Brooks, M., Cooper, E., Reynolds, G., Deutchki, E., Vachio, M., & Lund, E. A. (2021). Transitioning speech-language assessment to a virtual environment: Lessons learned from the ELLA study. *Language, Speech, and Hearing Services in Schools, 52*(3), 769–775. https://doi.org/10.1044/2021_LSHSS-20-00149

Wright, A. J. (2020). Equivalence of remote, digital administration and traditional, in-person administration of the Wechsler Intelligence Scale for Children, Fifth Edition (WISC-V). *Psychological Assessment, 32*(9), 809–817. https://doi.org/10.1037/pas0000939

Zoom Video Communications, Inc. (2020). *Background noise suppression*. https://support.zoom.us/hc/en-us/articles/360046244692-Background-noise-suppression

Zoom Video Communications, Inc. (2021). *System requirements for Windows, macOS, and Linux*. https://support.zoom.us/hc/en-us/articles/360046244692-Background-noise-suppression

Zou, G. Y. (2013). Confidence interval estimation for the Bland-Altman limits of agreement with multiple observations per individual. *Statistical Methods in Medical Research, 22*(6), 630–642. https://doi.org/10.1177/0962280211402548