unlocking potential

Evidence for Equivalency in Remote Assessment Using the Clinical Assessment of Pragmatics (CAPs™)

Kristin R. Ratliff, PhD[1], and Adriana Lavi, PhD, CCC-SLP[2]
[1]Western Psychological Services, [2]Lavi Institute

The demand for valid and reliable remote assessment methods is greater than ever now that traditional, in-person testing is inhibited due to COVID-19 restrictions. Individually administered performance tests have not been standardized in a remote capacity, and the change from traditional, face-to-face methods to remote methods may impact performance and/or the clinician's ratings of performance. There is evidence of the effectiveness of telepractice as a delivery model for services in the school system (Gabel, Grogan-Johnson, Alvares, Bechstein, & Taylor, 2013; Grogan-Johnson, Alvares, Rowan, & Creaghead, 2010; Grogan-Johnson et al., 2011; Lewis, Packman, Onslow, Simpson, & Jones, 2008; McCullough, 2001). However, for clinicians to confidently present the results of remote assessment, it is helpful to present direct evidence of the equivalency of the scores obtained from remote administration with those from the standardized, in-person administration whenever possible.

The present study examines the equivalency of pragmatic language performance as measured by the Clinical Assessment of Pragmatics (CAPs) during an in-person administration compared to a remote administration. The study used a test–retest design where the same individual was tested twice, once within an in-person session and once within a remote session. The same examiner administered both sessions, and the order of which session occurred first (remote vs. in-person) was counterbalanced. The study aims to reveal any potential differences in performance that might occur when testing within a traditional, face-to-face administration compared to within a remote administration. Additionally, the study design could reveal any differences in the examiner's ratings of performance when testing remotely compared to in-person.

WPS provides general guidelines for administering assessments remotely, which should be reviewed prior to administering the CAPs in this way. Additionally, clinicians should take care to follow the methods described within this study to achieve results that are most parallel to the standardized procedures.

METHOD

Participants

The sample consisted of 11 children, aged 8 years, 3 months, to 15 years, 9 months. Table 1 displays the demographic characteristics of the sample, which was reasonably well-balanced for gender and age range. The ethnic representation of the sample was not representative of the general population but was more diverse in terms of socioeconomic status. Data were collected between March of 2018 and May of 2020 in California.

The sample included four typically developing individuals, two with specific language impairment, and five with high-functioning autism. Six state-licensed, ASHA-certified, school-based speech–language pathologists participated in the study. The examiners were compensated for their time, and the participants received gift cards as incentive for participation. The clinical sample was recruited from the examiners' school-based caseload, and the evaluations were conducted as part of their initial or triennial IEP (individualized educational program) assessment. The typically developing sample was referred to Go2Consult Speech and Language Services, a certified special education staffing company, for an IEE (independent educational evaluation) and agreed to participate in the study.

Materials and Procedures

All examiners and participants used either a laptop or desktop personal computer and headphones with a built-in microphone during the administration. Video communication was established via a secure, password-protected meeting using Zoom (with additional licensing for the examiner's account to ensure HIPAA compliance). All examiners used a hardwired internet connection, and the participants' parents were instructed to use the same. However, due to availability, about half used this option while the remaining participants used a secure, stable Wi-Fi connection.

The examiner accessed the online CAPs videos and played them for the participant using the screen-sharing feature on Zoom. After each item, the examiner paused the video and activated full-screen for the window showing the participant's face on the examiner's screen. Participants were instructed to wait to give their oral response only when the examiner said, "Go," to allow time for the examiner to activate the full-screen of the participant. This allowed for the examiner to see the participant's facial expressions more accurately. Then the examiner would activate full-screen for the window displaying the CAPs video again and play the next item, repeating this process until all items were administered.

There were two instances when an examiner had to discontinue testing: once because a participant appeared to be fatigued and once because a participant appeared to be distracted by a sibling. In both instances, the assessment session was rescheduled to a later time. There were a few instances of poor connection reported by the examiners, and they instructed the participant to stop for a few moments. After waiting a few seconds to ensure that the connection was stable, the examiner continued with the video and the questions, restarting the video if the disruption occurred during the initial viewing.

All participants completed a full CAPs administration during their first testing session within either a remote or in-person setting. A follow-up testing session was scheduled for approximately three weeks later, when the same participant was administered the CAPs again but, this time, in the opposite condition from the first (e.g., if they completed a remote administration in Session 1, then they completed an in-person administration in Session 2 and vice versa). To reduce recall bias, the examiners did not inform the examinees at the time of the first administration that they would be tested again. All retesting was done by the same examiner who administrated the test the first time.

This test–retest design was utilized so that the same participant experienced both an in-person administration and a remote administration. The order of the remote vs. in-person test sessions was counterbalanced to remove any effects of test familiarity on performance. The test–retest interval was an average of three weeks, ranging from 20 to 90 days. Over this interval, test scores are not expected to change appreciably due to any development of the underlying language abilities and have been demonstrated to be stable and reliable during this timeframe (Lavi, 2019). All administrations were conducted with parent consent.

For both testing sessions, the examiner used a paper Record Form to record the participant's responses. The Core Pragmatic Language Composite was calculated to explore any effect of remote administration on overall pragmatic language performance. Additionally, the Paralinguistic Index was calculated to examine any differences in performance and examiner ratings for reading and using nonverbal cues within a remote assessment setting. Raw scores were calculated by the examiner for each test administration and then converted to standard scores with a mean of 100 and standard deviation of 15. Test–retest reliability was conducted using the standard scores, while rater reliability was conducted using the raw scores.

RESULTS

The results of the test–retest reliability study are presented in Table 2, displayed first for the entire sample together and then separately, by the clinical diagnosis of the participant. As the focus of the analysis examines only the *difference* in scores from in-

person to remote assessment, all participants were grouped together for the primary analysis. For both the CAPs Core Pragmatic Language Composite and Paralinguistic Index, the reliability coefficients are .99. Similar results of high correlation between scores obtained in a remote and an in-person setting are found for the individual groups based on clinical diagnosis, ranging from .94 to 1.00 (.99 to 1.00 for corrected correlations).

To illustrate test equivalency in another way, Table 2 also shows the means and standard deviations for the in-person and remote standard scores, as well as the effect size of the difference between the means. The variance in means across groups reflects the expected range of performance for typically developing participants (ranging from 93.5 to 94.0) to those with specific language impairment (ranging from 89.0 to 89.5) and high-functioning autism (ranging from 68.0 to 68.6). The effect size was calculated as the difference between the mean standard scores of the two testing occasions, divided by the pooled standard deviation. By this method, an effect size of 0.2 is considered small, 0.5 is considered medium, and 0.8 is considered large (Cohen, 1992). The effect sizes range from 0.01 to 0.02 for the entire sample, and 0.00 to 0.19 across the individual clinical groups. All of the observed effect sizes are considered small, indicating negligible change between the two conditions of testing, remote and in-person.

There were no statistically significant differences found between in-person and remote administrations for CAPs standard scores using a paired samples t-test for the sample as a whole, (Core Pragmatic Language Composite, $t(10) = 0.43$, $p = 0.68$; Paralinguistic Index, $t(10) = 1.00$, $p = 0.34$). Further, no significant differences were found between in-person and remote assessments when looking at the clinical groups separately for the Core Pragmatic Language Composite (Typically Developing, $t(3) = 1.00$, $p = 0.39$; SLI, $t(1) = 1.00$, $p = 0.50$; ASD, $t(4) = 0.00$, $p = 1.00$) or for the Paralinguistic Index (Typically Developing, $t(3) = 0.00$, $p = 1.00$; SLI, $t(1) = 0.00$, $p = 1.00$; ASD, $t(4) = 1.00$, $p = 0.37$).

Another method of examining the reliability of the ratings across conditions was demonstrated by comparing the examiner's calculated raw scores for each participant during the in-person administration to the raw score they calculated for the remote administration. Rater reliability was conducted using the intraclass correlation coefficient, following the method outlined by Shrout and Fleiss (1979). The intraclass correlation coefficients were .99 for both the Core Pragmatic Language Composite and the Paralinguistic Index. These results indicate a very high level of agreement across the conditions of in-person and remote administrations for the same participant.

DISCUSSION

Participants were administered the CAPs on two separate occasions: one being a remote administration via an online platform and the other being an in-person administration. The order of remote and in-person administration conditions was counterbalanced to avoid any effects on performance due to test familiarity, and the same examiner conducted the assessment during both sessions. Test scores were compared for the two sessions with no significant differences found between the remote and in-person performances. Further, the reliability of scores was extremely high across the different methods of in-person and remote administrations using both raw and standard scores.

These results suggest that the test scores obtained through administering the CAPs remotely, via an online video conferencing platform with screen-sharing capabilities, are equivalent to the test scores obtained through the standardized, in-person administration. As such, remote assessment does not appear to hinder the individual's performance in understanding and using nonverbal cues, as evidenced by the stability in the Paralinguistic Index from in-person to remote administrations. Similarly, these results support the equivalence of the examiner's ability to adequately rate the nonverbal cues in the individual's response that are associated with pragmatic language. The same pattern of results was found for clinical groups of SLI and ASD as for typically developing individuals. Thus, equivalency of CAPs scores using remote assessment can be extended for use with individuals

who have clinical conditions or disabilities as well. Taken together, these results provide further support for the valid and reliable application of normative scores for individually administered performance tests that are adapted to a remote assessment platform, particularly when the remote administration captures the same mechanisms and constructs as the in-person assessment.

Further studies should examine the generalizability of these results within larger samples and across more varied clinical populations, particularly those who may have difficulties using the technology that is currently utilized in the digital realm. Motor and other physical impairments may play a significant role in performance for some individually administered performance tests when transitioning to remote assessment. Further study should investigate other assessments, online platforms, and digital delivery methods to better identify those factors that impact examinee performance when shifting to remote assessment. Such findings are needed to ensure the reliability and validity of remote assessment more generally and to further the development of digital test design itself.

REFERENCES

Gabel, R., Grogan-Johnson, S., Alvares, R., Bechstein, L., & Taylor, J. (2013). A field study of telepractice for school intervention using the ASHA NOMS K–12 database. *Communication Disorders Quarterly, 35*(1), 44–53.

Grogan-Johnson, S., Alvares, R., Rowan, L., & Creaghead, N. (2010). A pilot study comparing the effectiveness of speech–language therapy provided by telemedicine with conventional on-site therapy. *Journal of Telemedicine and Telecare, 16*, 134–139.

Grogan-Johnson, S., Gabel, R., Taylor, J., Rowan, L., Alvares, R., & Schenker, J. (2011). A pilot exploration of speech–sound disorder intervention delivered by telehealth to school-age children. *International Journal of Telerehabilitation, 3*, 31–41.

Lavi, A. (2019). *Clinical Assessment of Pragmatics (CAPs)*. Torrance, CA: Western Psychological Services.

Lewis, C., Packman, A., Onslow, M., Simpson, J., & Jones, M. (2008). A phase II trial of telehealth delivery of the Lidcombe Program of Early Stuttering Intervention. *American Journal of Speech–Language Pathology, 17*, 139–149.

McCullough, A. (2001). Viability and effectiveness of teletherapy for preschool children with special needs. *International Journal of Language and Communication Disorders, 36*, 321–326.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428.

Table 1. Demographic Characteristics of the CAPs Equivalency Sample

Characteristic	*n*	% of sample
Gender		
Male	8	72.7
Female	3	27.3
Age in Years		
8	3	27.2
9	2	18.2
10	1	9.1
11	1	9.1
12	1	9.1
14	2	18.2
15	1	9.1
Race/Ethnicity		
Hispanic	5	45.5
African American	2	18.2
Caucasian	4	36.3
Socioeconomic Status[a]		
Low (≤25k)	5	45.4
Medium (26–99k)	3	27.3
High (≥100k)	3	27.3

Note. $N = 11$.

[a]Income derived from zip codes, using IRS.gov SOI Tax Stats data for 2017 gross adjusted income by zip code.

Table 2. Equivalency of CAPs Standard Scores for In-Person vs. Remote Administrations: Corrected Correlations, Descriptives, and Effect Sizes

			In-Person		Remote				
	n	CAPs Scores	Mean	*SD*	Mean	*SD*	Effect Size	*r*	Corrected *r*[a]
All Participants	11	Core Pragmatic Language Composite	81.45	12.89	81.64	13.17	0.01	.99	.99
		Paralinguistic Index	81.36	12.82	81.09	13.16	0.02	.99	.99
Typically Developing	4	Core Pragmatic Language Composite	93.75	1.50	94.00	1.40	0.17	.94	1.00
		Paralinguistic Index	93.50	1.91	93.50	1.91	0.00	1.00	1.00
ASD	5	Core Pragmatic Language Composite	68.60	5.08	68.60	5.90	0.00	.94	.99
		Paralinguistic Index	68.60	5.08	68.00	5.34	0.12	.97	1.00
SLI	2	Core Pragmatic Language Composite	89.00	2.83	89.50	2.12	0.19	.99	1.00
		Paralinguistic Index	89.00	2.83	89.00	2.83	0.00	1.00	1.00

Note. Means, *SD*'s expressed in standard score units ($M = 100$, $SD = 15$).

Effect size (Cohen's *d*) = In-Person mean minus Remote mean, divided by pooled *SD*.

[a]The reliability coefficient was corrected for variability of normative group ($SD = 15$) based on the standard deviation obtained for In-Person, using Guilford's (1954) formula.