JSLHR

Research Article

Measuring Speech Comprehensibility in Students with Down Syndrome

Paul J. Yoder,[a] Tiffany Woynaroski,[b] and Stephen Camarata[b]

Purpose: There is an ongoing need to develop assessments of spontaneous speech that focus on whether the child's utterances are comprehensible to listeners. This study sought to identify the attributes of a stable ratings-based measure of speech comprehensibility, which enabled examining the criterion-related validity of an orthography-based measure of the comprehensibility of conversational speech in students with Down syndrome.
Method: Participants were 10 elementary school students with Down syndrome and 4 unfamiliar adult raters. Averaged across-observer Likert ratings of speech comprehensibility were called a *ratings-based* measure of speech comprehensibility. The proportion of utterance attempts fully glossed constituted an *orthography-based* measure of speech comprehensibility.
Results: Averaging across 4 raters on four 5-min segments produced a reliable (G = .83) ratings-based measure of speech comprehensibility. The ratings-based measure was strongly ($r > .80$) correlated with the orthography-based measure for both the same and different conversational samples.
Conclusion: Reliable and valid measures of speech comprehensibility are achievable with the resources available to many researchers and some clinicians.

The goal of expressive communication is to produce a comprehensible message (i.e., a message that can be understood by a listener). We refer to students' ability to produce conversational-level speech that can be understood by an unfamiliar listener as speech comprehensibility (Camarata, Yoder, & Camarata, 2006; Yoder, Camarata, & Gardner, 2005). We use the term *speech comprehensibility* as opposed to the more commonly used term *speech intelligibility* because at the measurement level, intelligibility has been defined differently, depending on the research lab. For example, some researchers measure speech intelligibility exclusively in contexts in which the intended message is known (see Kent, Miolo, & Bloedel, 1994, for a review). Well-designed and conducted efficacy studies with students who have speech disorders not caused by physical impairment, cleft palette, or hearing impairment rarely directly test whether speech therapy affects speech comprehensibility.

An Orthography-Based Measure of Speech Comprehensibility

Despite the evident face and ecological validity of speech comprehensibility, one reason that it is rarely tested as an outcome of speech therapy is the concern that ratings-based measures of speech comprehensibility reflect characteristics of the listener, not just the speech of the speaker. For example, speech comprehensibility ratings have been found to vary as a function of familiarity of the listener with the speaker (Brodkey, 1972; Doyle, Swift, & Haaf, 1989; McGarr, 1983; Platt, Andrews, Young, & Quinn, 1980; Tjaden & Liss, 1995a, 1995b; Weist & Kruppe, 1977; Wilcox, Kouri, & Caswell, 1990). In addition, Schiavetti (1992) noted that confidence intervals around ratings from the middle part of the rating scale for comprehensibility tend to be quite large, rendering individual values very difficult to interpret.

As an alternative to global ratings of speech comprehensibility, some researchers have defined speech comprehensibility by using orthography-based measures. For example, Kwiatkowski and Shriberg (1992) operationalized the extent to which students' speech could be understood as the proportion of *word* attempts fully glossed. However, this variable has the very real problem of not identifying how many words are attempted in highly incomprehensible speech.

[a]Special Education, Vanderbilt University, Nashville, TN
[b]Hearing & Speech Sciences, Vanderbilt University, Nashville, TN
Correspondence to Paul J. Yoder: paul.yoder@vanderbilt.edu

Editor: Jody Kreiman
Associate Editor: Julie Liss
Received April 20, 2015
Revision received August 25, 2015
Accepted October 7, 2015
DOI: 10.1044/2015_JSLHR-S-15-0149

Disclosure: *Paul J. Yoder and Stephen M. Camarata are the primary authors of the Broad Target Speech Recasts approach. Neither receives financial gain from the use of this treatment.*

Another *orthography-based measure* of speech comprehensibility that has been utilized in research is the proportion of *utterance* attempts that orthographers fully gloss from a conversational language sample (Yoder et al., in press). Semantic, grammatical, and terminal prosody information makes segmenting utterances versus words achievable (Miller & Chapman, 1993). Reasonable guesses at particular words that are inaccurately produced are supported by access to visual and auditory contextual information. Thus, the proportion of utterance attempts fully glossed might be a useful measure of comprehensibility.

Indeed, recent studies have found that this utterance-level, orthography-based index of speech comprehensibility has acceptable to very good interobserver reliability, even in studies with participant samples that have an average of only 50% comprehensible utterances. For example, the interobserver reliability for the proportion of utterance attempts fully glossed was .74 in a sample of preschoolers with speech and language impairments (Yoder et al., in press), .72 in a sample of preschoolers with Down syndrome (DS; Camarata et al., 2006), and .86 in a sample of elementary school–aged students with DS (Yoder, Camarata, & Woynarowski, in press).

Rationale for a Ratings-Based Measure of Speech Comprehensibility

This orthography-based index is derived from the transcripts of well-trained orthographers by using a detailed operational definition of *word approximation,* detailed utterance segmenting rules, and multiple stop-and-go listening passes before rendering a gloss. Thus, if our goal is to capture conversational speech comprehensibility in an ecologically valid context (i.e., the judgment of unfamiliar listeners in real time), it is reasonable to question the face validity of orthography-based measures of speech comprehensibility. Tension exists between the need for a face and ecologically valid measure of speech comprehensibility versus the need for this measure to reflect the speech of the child, not the characteristics of the listener.

Others have used the average of multiple informants' judgments to render improved stability relative to any individual rating (Conway, Jako, & Goodman, 1995). Averaging multiple raters' estimates has been particularly useful when measuring characteristics that involve subjective judgments of other's behavior (Messinger, Mahoor, Chow, & Cohn, 2009). In fact, using the average of multiple estimates of a phenomenon to improve the measurement of such phenomenon has a long and productive history (Galton, 1907; Rushton, Brainerd, & Pressley, 1983).

In addition, when measuring the outcome of speech intervention, we need a measure of speech comprehensibility that is stable across speech samples (i.e., that reflects a highly generalized characteristic). This issue is salient in children with severe speech disorders because it is not uncommon for such children to have inconsistent comprehensibility. Speaking rates, intensities of utterances, accuracy of production, and length and complexity of utterances vary considerably from session to session, which, in turn, affects comprehensibility scores (Flipsen, 2006). Averaging scores across multiple behavior samples has been shown to enable stable estimates of behavior, varying widely across sessions (Sandbank & Yoder, 2014). The question of how many sessions are needed to provide a stable estimate of speech comprehensibility as a generalized characteristic arises can be tested empirically.

Generalizability (G) theory clarified the logic behind why an aggregation of many estimates of individual differences on a construct is less influenced by characteristics of the examiner, rater, or procedure used to measure the construct (Cronbach, 1972). In Generalizability theory, a G coefficient is the proportion of total variance of scores divided into the among-participant variance on the dependent variable for the reliability sample (Chronbach, Rajaratnam, & Gleser, 1963; Shavelson & Webb, 1991). The higher the proportion, the more of the total variance in the reliability study is due to what we want to measure: individual differences in speech comprehensibility among participants. We interpret high G coefficients as an indication that an estimate is stable. When two aspects of the measurement system (e.g., number of raters and number of speech samples) are studied, G studies can be used to identify which aspect is the greater source of measurement error (e.g., the subjectivity of the rater or unrepresentativeness of the speech sample; Shavelson & Webb, 1991).

Another application of G theory, called a decision (D) study, uses an extension of Spearman's prophesy formula to decide how many estimates need to be averaged to render scores that are stable. For example, a past G and D study involving preschool children with severe speech disorders indicated that a 20-min session with a single orthographer was sufficient to produce a stable estimate of speech comprehensibility by using the *percentage* of utterances fully glossed (Bruckner, Yoder, & McWilliam, 2006). In contrast, in this same sample, it took at least five sessions with a single orthographer to derive a stable estimate of speech comprehensibility by using the *number* of fully glossed utterances (Bruckner et al., 2006). Thus, one can use G and D studies to guide our computation of average rating of speech comprehensibility by indicating how many sessions and how many raters one needs to average to yield a stable estimate of speech comprehensibility. The average of ratings will be referred to as the *ratings-based measure* of speech comprehensibility.

Past Research on Speech Comprehensibility of Students with DS Highlights the Need for the Current Study

Students with DS represent a clinical population needing a psychometrically sound measure of speech comprehensibility. People with DS often experience lifelong difficulties in making themselves understood in conversations (Kumin, 2006), so effective interventions that improve comprehensibility are needed. One speech therapy method,

Broad Target Speech Recasts (BTSR; Camarata, Yoder, & Camarata, 2006; Yoder, Camarata, & Gardner, 2005), facilitates increases in the orthography-based measure of speech comprehensibility in conversational samples of students with DS who had relatively high verbal imitation prior to treatment onset (Yoder, Camarata, & Woynaroski, in press). Because the speech samples in that study were videotaped, they can be used as the basis for determining the number of estimates needed to compute a reliable ratings-based measure of speech comprehensibility. These video samples can also be used to estimate the criterion-related validity of the orthography-based measure of speech comprehensibility relative to the more face and ecologically valid ratings-based measure of speech comprehensibility. If the validity of the two measures is supported, it supports inferences that (a) orthography-based measures of speech comprehensibility are valid estimates of what most people think of as speech comprehensibility, and (b) ratings-based measures of speech comprehensibility are valid tools that clinicians can use to assess their clients' progress on speech comprehensibility.

Research Questions

The research questions for the current study were as follows: (a) When rating comprehensibility of students with DS in speech samples, which aspect of the measure produces the most measurement error: different raters or different numbers of speech samples? (b) How many unfamiliar raters' estimates of speech comprehensibility and how many minutes of a speech sample are needed to obtain a stable estimate (i.e., reliability coefficient > .8) of speech comprehensibility for elementary school students with DS? (c) To what extent does the orthography-based measure of interest in this study correlate with the stable ratings-based measure? We consider this last analysis a test of criterion-related validity of the orthography-based measure, a measure that was promising as a research tool and that has been affected by speech therapy in some students with DS (Yoder et al., in press).

Methods

Participants

Children with DS Who Also Had Severe Speech Disorders

Ten students with DS were randomly selected from a pool of 51 students with DS who participated in a larger study that tested the efficacy of BTSR (Yoder et al., in press). The selection criteria for entry into the larger study were (a) presence of DS as indicated by appearance and parent report, (b) less than 75% of utterance attempts comprehensible and performance less than the 10th percentile on an articulation test (Arizona Articulation Proficiency Scale–Third Edition; Fudala, 2001), (c) hearing within normal limits in at least one ear, (d) production of at least 20 different comprehensible, referential words in a 20-min speech sample, (e) English as the primary language spoken at home, and (f) chronological age between 5 and 12 years. Exclusion criteria included (a) uncontrolled seizures, (b) diagnosis of attention deficit disorder, autism spectrum disorder, or apraxia, and (c) severely disruptive behavior, as reported by the parent. Students are described in Table 1. The scores for IQ, receptive vocabulary, and articulation were not available for the assessment period most relevant for the current study; therefore, these scores represent the participants' status at entry into the larger study on BTSR. The other variables were derived from the speech samples at the measurement period most relevant to thecurrent study (i.e., 9 months after entry into the larger study).

Raters

Raters of students' speech comprehensibility were graduate students and/or research staff in the Department of Hearing & Speech Sciences at Vanderbilt University. Raters all had some degree of experience in listening to individuals with speech sound disorders but varied in ways that may have affected their judgments of students' speech comprehensibility. Potentially relevant rater characteristics are summarized in Table 2.

Research Design

The measurement period at 9 months after entry into the larger study was selected because variability in speech comprehensibility was higher at this period than it was when the students entered the larger study. High variability in speech comprehensibility is desirable because G, D, and correlational analyses attempt to explain individual differences in speech comprehensibility.

Table 1. Sample means and standard deviations of descriptive variables for the 10 selected participants.

Variable	*M*	*SD*
Proportion of utterance attempts comprehensible[a]	0.61	0.17
Arizona Articulation Proficiency Scale standard score[b]	68	7.70
Mean length of utterance[a]	1.60	0.49
Number of different word roots[c]	79	45
IQ[d]	82	18
Receptive vocabulary standard score[e]	55	20
Chronological age in years at period for current study	13	1.60

[a]Average score over two 20-min conversational speech samples with an unfamiliar research staff at the current study's measurement period. [b]Arizona Articulation Proficiency Test–Third Edition (Fudala, 2000) at 9 months prior to period for current study; population *M* (*SD*) is 100 (15). [c]Average of scores over two 20-min conversational speech samples at the current study's measurement period. [d]Stanford Binet–Fourth Edition (Thorndike, Haden, & Sattler, 1986) at 9 months prior to period for current study; population *M* (*SD*) is 100 (15). [e]Peabody Picture Vocabulary Test–Fourth Edition (Dunn & Dunn, 2007) at 9 months prior to period for current study; population *M* (*SD*) is 100 (15).

Table 2. Characteristics of raters.

Characteristics	Rater 1	Rater 2	Rater 3	Rater 4
Sex	Female	Female	Female	Male
Age	29	24	27	32
Highest degree achieved	BA Linguistics and BS Communicative Disorders	BA Linguistics and French	BA Psychology	BA Psychology
Position at time of ratings	MS SLP student	MS SLP student	PhD SLP student	Research assistant in SLP laboratory
Formal coursework in articulation and phonology completed at time of rating	Undergraduate: phonetics and phonology, sounds of the world's languages, disorders of articulation and phonology, descriptive linguistics, assessment and treatment of children with communicative disorders, and speech science Graduate-level: articulation disorders and motor speech disorders	Graduate level: two phonetics courses, articulation and phonology	Graduate level: articulation and phonology	None
Other skills training in articulation and phonology	Transcription using IPA and administration and scoring of standardized measures of articulation and phonology	Transcription using IPA and administration and scoring of standardized measures of articulation and phonology	Transcription using IPA and administration and scoring of standardized measures of articulation and phonology	Transcription of language samples (not using IPA)
Estimated number hours experience in listening to individuals with impaired speech comprehensibility	50–75 hours as SLP graduate student clinician	50–75 hours as SLP graduate student clinician and 10–25 hours as SLP graduate research assistant	50–75 hours as SLP graduate research assistant and 100+ hours as a nanny, tutor, and sibling of people with speech sound disorders	100+ hours as research staff

Note. SLP = speech-language pathology; IPA = International Phonetic Alphabet.

G and D studies were conducted by using the ratings-based measure of speech comprehensibility from speech samples that were administered within 1 week of each other. The orthography-based measures of speech comprehensibility for the same speech samples had been derived in the larger study. Correlations between the ratings-based and orthography-based measures were computed.

G theory operationalizes classical test theory by using analysis of variance to estimate the amount of variance contributed by the participants versus different aspects of measurement. According to G theory, the mean squares for the various factors in the design can be used to separate *true* person variance (i.e., what we wish to measure) from that which is resultant from the way we estimate the construct of interest (e.g., number of speech samples and number of raters, error variation; Webb & Shavelson, 2005) in a G study. The ways we measure the construct are called measurement facets. The current G study used two measurement facets: raters and speech samples. The partitioned variance can then be used to calculate a generalizability (G) coefficient, a type of intraclass correlation that indicates the level of measurement stability achieved. The D studies use the variance estimates provided by a G study to calculate the number of speech samples or number of raters that we need to average to achieve a threshold level of stability (Shavelson & Webb, 1991). The results of our D study assume that particular raters and particular speech samples are equivalently valid.

Procedures

Speech Samples

Two 20-min speech samples using the same two sets of parallel-function toys were collected for each student. The examiners were unfamiliar with the students with whom they were interacting. Examiners responded to children's actions and communication and asked questions and delivered comments about the children's focus of attention and action. They were directed not to use imitation prompts, speech recasts, topic-initiating questions, or play directives during the sampling. The sessions were video- and audio-recorded for later transcription using a digital camcorder by Panasonic (model PV-GS500 3CCD/3DCC; Panasonic North America, Newark, NJ) in conjunction with an Azden Wireless Transmitter and Receiver Set with an EX503 LAV Microphone (Azden Corporation USA, Mount Arlington, NJ).

Orthography

Speech samples were transcribed by research staff who were unfamiliar with the children. These staff were trained to a criterion of at least 80% agreement for three

consecutive sessions and retrained whenever agreement fell below this criterion. Transcribers were allowed up to three listening passes per utterance. The proportion of utterance attempts that were fully glossed from all utterances was the orthography-based measure of speech comprehensibility. For example, consider that a child produces the utterance *ah wa du,* as he or she reaches for a pitcher of juice. This utterance would be fully glossed if a rater comprehended every word that the child attempted in the utterance and transcribed *I want juice*. This utterance would be only partially glossed if a rater, even after three listening passes, could only comprehend the final word in the utterance *du* and transcribed *x juice*. Transcription manuals are available from the first author. More details on the transcription process are available in Yoder et al. (in press).

Interobserver reliability of the orthography-based measure of speech comprehensibility variable was estimated by a intraclass correlation coefficient, which reflects variability among participants and among orthographers. The intraclass correlation coefficients for a single observer's scores from a single 20-min speech sample, based on a random sample of 29% of the larger study's sample, was .92 for the first speech sample and .90 for the second speech sample. Past work has shown that a single rater and a single 20-min speech sample can yield stable orthography-based speech comprehensibility scores in preschoolers with severe speech and language disorders (Bruckner et al., 2006). These data corroborate that view.

Rating of Speech Samples

To derive the ratings-based measure of speech comprehensibility, the two 20-min speech samples were divided into eight 5-min segments. These eight segments were shown to four raters, who were instructed to rate all eight segments for all 10 students with DS. The rating scale was a 5-point Likert scale with 5 representing the highest level of speech comprehensibility. To train raters, we created two video anchors each for 1, 3, and 5 values on the rating scale. Such a strategy has been used to yield reliable estimates of voice quality (Kreiman, Gerratt, Kempster, Erman, & Berke, 1993). To increase the probability that observers would rate the extent to which utterance attempts are comprehensible, not the number of comprehensible utterances, we selected six video anchors that had approximately the same number of intelligible utterances ($M = 64$, $SD = 8$) from the larger study. Using the minimum, maximum, and mean orthography-based measure of speech comprehensibility at each period, we identified .13–.27, .48–.58, and .82–.89 as the target range of orthography-based scores for 1, 3, and 5, respectively. The selected six video anchors had the characteristics listed in Table 3. After viewing the six video anchors with their assigned Likert scores, the assessment of the 10 participants' speech samples began. Raters were instructed to watch the entire 5-min segment and then rate on the 5-point scale their global impression of the comprehensibility of the participant's speech.

Table 3. Number and proportion of utterance attempts that were fully glossed for the six video anchors used in training the raters.

Likert score (number label for anchor)	Number of comprehensible utterances	Proportion of utterance attempts that were fully glossed
1 (1st)	54	.21
1 (2nd)	71	.26
3 (1st)	53	.50
3 (2nd)	69	.52
5 (1st)	68	.83
5 (2nd)	71	.89

Results

Preliminary Results

To enable a fully crossed G study, the number of segments was reduced to seven because one of the raters did not rate all eight segments for all participants. Thus, the actual G study had two measurement facets with four levels for raters and seven levels for segments. A facet is a term used in G and D studies to refer to persons as well as ways to measure (e.g., raters and segments) among person variation on the construct of interest (i.e., speech comprehensibility).

The Measurement Facet Producing Most Errors in Ratings-Based Speech Comprehensibility

The results of the G study are presented in Table 4. Using only one rater and one 5-min segment, the reliability of ratings-based speech comprehensibility was less than .5. That is, less than 50% of the variance in a single observer's ratings-based speech comprehensibility scores was due to individual differences in speech comprehensibility. Without a doubt, this is insufficient for research or clinical purposes. The largest source of measurement error was Rater. This can been shown by summing the variance accounted for by the main effect of Rater (12%) and the Persons × Raters (23%) interaction, which is 35%. By comparing this 35% to the measurement error attributed to Segments (0%) and the interaction between Segments and Persons (3%), it becomes clear that raters' variability was the larger source of measurement error.

Table 4. Source and percentage of variance explained in ratings-based speech comprehensibility scores.

Source of variance	Percentage of variance explained
Persons	47.5
Raters	11.6
Segments	0.0
Persons × Raters	23.2
Persons × Segments	2.5
Raters × Segments	0.0
Persons × Raters × Segments	15.1

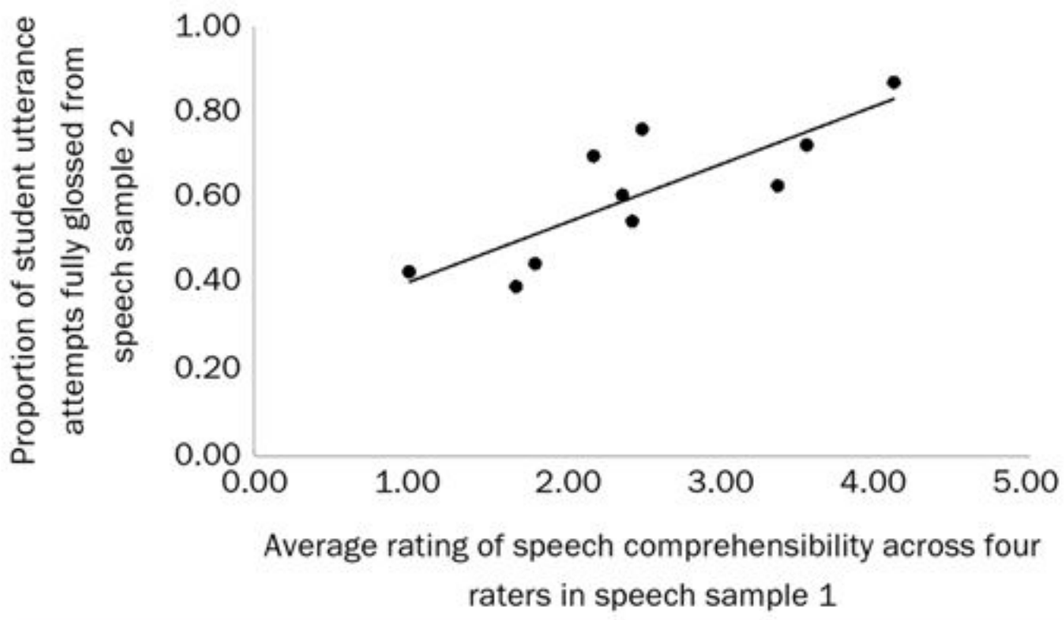
The Number of Raters Needed to Produce a Stable Ratings-Based Measure of Speech Comprehensibility

In the D study, the number of sessions was kept constant (i.e., at the length of one 20-min speech sampling session), and number of raters was varied because only the latter contributed noteworthy error variance to the ratings-based measure. The results of the D study are in Table 5. The reliability of one rater's average ratings-based estimate of speech comprehensibility across the four 5-min segments was .55. For school-age children with DS, it takes the average of four raters' scores across a 20-min speech sample to produce a G coefficient at or above our threshold of .8 stability.

The Correlation of the Stable Estimate of the Ratings-Based Measure with the Orthography-Based Measure of Speech Comprehensibility

Using the results of the D study to guide us, we averaged the estimates across four raters to produce the ratings-based estimates for a single 20-min speech sample for each of the 10 participants with DS (Yoder et al., in press). We examined the association between this ratings-based score with the orthography-based measure (the proportion of utterance attempts that were fully glossed), as derived from one trained orthographer's broad transcription for each of the two 20-min speech samples collected for each student. The Pearson's product–moment correlation for the ratings-based measure with the orthography-based measure was .82 for the same sessions and .84 for different sessions (both p values <.01). The data meet the assumptions for Pearson's r; however, the Spearman's rho for the same associations are .88 and .89, respectively. One of these associations is displayed as a scatterplot in Figure 1.

Figure 1. Scatter plot and regression line showing the association between two measures of speech comprehensibility from different speech samples.



Discussion

The purpose of this study was to examine the validity of an orthography-based measure of speech comprehensibility that has been found to be sensitive to changes produced by speech therapy in some students with DS. To examine the validity of this measure, we needed a stable ratings-based measure of speech comprehensibility. The results of a G and D study indicated that the average of four observers' global ratings of speech comprehensibility for a 20-min speech sample has acceptable reliability for research and clinical applications. Even when the ratings-based estimate is from a different speech sample as the orthography-based estimate, the correlation of the two types of measures of speech comprehensibility was very high (r >.8).

Table 5. Absolute G coefficients for scenarios that vary by number of raters of a 20-min speech sample.

Number of raters in scenario	G coefficient
1	.55
2	.70
3	.77
4	.82

Limitations

Potential contributions of these findings must be evaluated in the context of the limitations of the study. One limitation of the present study is that the sample size of students with DS was small (i.e., 10). The small sample size has potential ramifications for generalizing the estimated number of raters needed to achieve a stable ratings-based index of speech comprehensibility. In particular, the confidence interval around the G coefficient is wide because it is influenced by the number of participants. Thus, caution should be exercised when generalizing these findings to future research or clinical practice.

Another limitation of the present result is that there is a possibility of bias for the sessions facet of the G study. The bias might have been produced by requiring raters to score all 5-min segments from a particular participant before progressing to the next participant. Such instructions might have increased the probability that past ratings influenced future ratings within the same participants, thus artificially reducing the error variance attributed to speech segments. Thus, the basis for the inference regarding number of segments needed to produce a reliable ratings-based measure of speech comprehensibility is weaker than that for number of raters. This problem does not compromise the findings relevant for identifying that rater was the measurement facet that produced the most measurement error in the ratings-based measure.

It is possible that the present result overestimates the number of speech-language pathologists (SLPs) one would need to average to achieve a stable ratings-based estimate of speech comprehensibility in our population of interest. Relative to the graduate students (SLPs in training) and research staff who provided ratings of participants' speech comprehensibility in the present study, SLPs have more training on measurement and sometimes more

experience in listening to speech with errors. Logic suggests that such training and experience would produce more agreement in ratings of speech comprehensibility; therefore, fewer raters would be necessary to obtain a stable estimate of the construct of interest. Thus, four can be considered a conservative estimate of the number of raters for a single speech sample that one would require to derive a stable estimate of speech comprehensibility in students with DS.

Strengths

The D study provided an empirical method for estimating how many non-SLP raters are needed to achieve a stable ratings-based measure of speech comprehensibility. In addition, one of the correlations involved measures from different speech samples. This enabled showing that the two measures are estimates of a generalized characteristic instead of a characteristic of a particular speech sample.

Measures with poor reliability tend to produce attenuated correlations because measurement error is randomly distributed across participants (Yoder & Symons, 2010). Thus, the high and significant correlation between the two types of measures of speech comprehensibility occurred despite, not because of, the possible limitations of the G and D studies.

Ramifications of Findings

The high correlation between orthography-based and ratings-based measures of speech comprehensibility supports the inference that (a) the average of four observers' global ratings of speech comprehensibility and (b) the proportion of utterance attempts that are fully glossed by a trained orthographer are valid measures of speech comprehensibility in students with DS. The findings support the scientific value of the orthography-based measure of speech comprehensibility for research purposes. The findings also suggest that clinicians can use the average rating of four individuals experienced in some degree to listening to and judging disordered speech for a single speech sample to derive stable estimates of speech comprehensibility for students with DS. Knowing that the rating of a single experienced listener is insufficient to produce a reliable estimate of speech comprehensibility might help clinicians seek the ratings of their colleagues before talking to stakeholders about the speech comprehensibility of a particular student.

In conclusion, these findings are important for research and clinical reasons. We hope that the findings will improve the probability that future research and clinicians will use measures of speech comprehensibility when evaluating their therapies.

Acknowledgments

This research was funded by the Institute of Education Science (R324A100225; awarded to P. Yoder and S. Camarata) and supported by the National Institute for Child Health and Human Development through the Vanderbilt Kennedy Center (P30HD15052; awarded to E. Dykens). The content is solely the responsibility of the authors and does not necessarily represent the official views of the Institute of Education Science. We are very grateful to our staff (Elizabeth Gardner, Catherine Bush, Jenny Elrod, Amanda Johnson, Tricia Paulley, Rebecca Frey, Megan Ochab, and Meghan Weber), the families who trust us with their children, and the raters who genverously gave of their time and effort. P. J. Yoder and S. Camarata are the primary authors of the Broad Target Speech Recasts approach. Neither receives financial gain from the use of this treatment.

References

Brodkey, D. (1972). Dictation as a measure of mutual intelligibility: A pilot study. *Language Learning, 22,* 203–217.

Bruckner, C. T., Yoder, P. J., & McWilliam, R. A. (2006). Generalizability and decision studies: An example using conversational language samples. *Journal of Early Intervention, 28,* 139–153.

Camarata, S., Yoder, P., & Camarata, M. (2006). Simultaneous treatment of grammatical and speech-comprehensibility deficits in children with Down syndrome. *Down Syndrome Research and Practice, 11,* 9–17.

Conway, J. M., Jako, R. A., & Goodman, D. F. (1995). A meta-analysis of interrater and internal consistency reliability of selection interviews. *Journal of Applied Psychology, 80,* 565–579.

Cronbach, L. J. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York, NY: Wiley.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology, 16*(2), 137–163.

Doyle, P. C., Swift, R., & Haaf, R. G. (1989). Effects of listener sophistication on judgments of tracheoesophageal talker intelligibility. *Journal of Communication Disorders, 22,* 105–113.

Dunn, L. M., & Dunn, D. M. (2007). *Peabody Picture Vocabulary Test* (4th ed.). San Antonio, TX: Pearson.

Flipsen, P., Jr. (2006). Measuring the intelligibility of conversational speech in children. *Clinical Linguistics & Phonetics, 20,* 303–312.

Fudala, J. (2000). *Arizona Articulation Proficiency Scale* (3rd ed.). Los Angeles, CA: Western Psychological Corporation.

Fudala, J. (2001). *Arizona Articulation Proficiency Scale* (3rd ed.). Los Angeles, CA: Western Psychological Services.

Galton, F. (1907). One vote, one value. *Nature, 75,* 450–451.

Kent, R. D., Miolo, G., & Bloedel, S. (1994). The intelligibility of children's speech: A review of evaluation procedures. *American Journal of Speech-Language Pathology, 3,* 81–95.

Kreiman, J., Gerratt, B. R., Kempster, G. B., Erman, A., & Berke, G. (1993). Perceptual evaluation of voice quality: Review, tutorial and a framework for future research. *Journal of Speech and Hearing Research, 36,* 21–40.

Kumin, L. (2006). Speech intelligibility and childhood verbal apraxia in children with Down syndrome. *Down Syndrome Research and Practice, 10,* 10–22.

Kwiatkowski, J., & Shriberg, L. D. (1992). Intelligibility assessment in developmental phonological disorders: Accuracy of caregiver gloss. *Journal of Speech and Hearing Research, 35,* 1095–1104.

McGarr, N. (1983). The intelligibility of deaf speech to experienced and inexperienced listeners. *Journal of Speech and Hearing Research, 26,* 451–458.

Messinger, D. S., Mahoor, M. H., Chow, S. M., & Cohn, J. F. (2009). Automated measurement of facial expression in infant–mother interaction: A pilot study. *Infancy, 14,* 285–305.

Miller, J., & Chapman, R. (1993). *Systematic analysis of language transcripts*. Madison, WI: University of Wisconsin.

Platt, L. J., Andrews, G., Young, M., & Quinn, P. T. (1980). Dysarthria of adult cerebral palsy: Intelligibility and articulatory impairment. *Journal of Speech and Hearing Research, 23,* 28–40.

Rushton, J. P., Brainerd, C. J., & Pressley, M. (1983). Behavioral development and construct validity: The principle of aggregation. *Psychological Bulletin, 94,* 18–38.

Sandbank, M., & Yoder, P. (2014). Measuring representative communication in young children with developmental delay. *Topics in Early Childhood Special Education, 34,* 133–141.

Schiavetti, N. (1992). Scaling procedures for the measurement of speech intelligibility. In R. D. Kent (Ed.), *Intelligibility in speech disorders* (pp. 11–34). Philadelphia, PA: John Benjamins.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer* (Vol. 1). Thousand Oaks, CA: Sage.

Thorndike, R. L., Hagen, E. P., & Sattler, J. M. (1986). *Stanford-Binet Intelligence Scale* (4th ed.). Itasca, IL: Riverside Publishing Company.

Tjaden, K., & Liss, J. M. (1995a). The influence of familiarity on judgments of treated speech. *American Journal of Speech-Language Pathology, 4*(1), 39–47.

Tjaden, K., & Liss, J. M. (1995b). The role of listener familiarity in the perception of dysarthric speech. *Clinical Linguistics & Phonetics, 9,* 139–154.

Webb, N. M., & Shavelson, R. J. (2005). *Generalizability theory: Overview*. Wiley StatsRef: Statistics Reference Online.

Weist, R. M., & Kruppe, B. (1977). Parent and sibling comprehension of children's speech. *Journal of Psycholinguistic Research, 6,* 49–58.

Wilcox, M. J., Kouri, T. A., & Caswell, S. (1990). Partner sensitivity to communication behavior of young children with developmental disabilities. *Journal of Speech and Hearing Disorders, 55,* 679–693.

Yoder, P., Camarata, S., & Gardner, E. (2005). Treatment effects on speech intelligibility and length of utterance in children with specific language and intelligibility impairments. *Journal of Early Intervention, 28,* 34–49.

Yoder, P., Camarata, S., & Woynaroski, T. (in press). Treating speech comprehensibility in students with Down syndrome. *Journal of Speech, Language, & Hearing Research.*

Yoder, P. J., & Symons, F. (2010). *Observational measurement of behavior*. New York, NY: Springer.