JSLHR

Research Note

The Use of Generalizability Theory to Inform Sampling of Educator Language Used With Preschoolers With Autism Spectrum Disorder

Andrea L. B. Ford[a] and LeAnne D. Johnson[a]

Purpose: A myriad features can impact the nature, frequency, and length of adult–child interactions important for language learning. Empirical investigations of language learning opportunities for young children with autism spectrum disorder (ASD) provide limited generalizable insight, with inferences more constrained to the sample than is often considered. The aim of this study was to explore a multidimensional understanding of reliability and define optimal measurement procedures for a measurement approach used to examine the language of preschool educators interacting with children with ASD.

Method: We employed the logic of Generalizability Theory to differentiate sources of error for two measurement facets, *occasion* and *observer*. We video-recorded four 15-min occasions of educator–child interactions for 11 participants with ASD during free-play in their respective inclusive preschool classrooms. Two trained observers coded all videos for six educator language variables: open-ended questions/statements, choice questions, yes/no questions, imitation prompts, statements, and other talk.

Results: The generalizability studies illustrated that, across all variables measured, observer accounted for little to no error. Occasion, however, accounted for much of the error for all language variables. To determine the number of occasions needed to achieve stable estimates of the variables, we manipulated occasion in the decision study. Five to more than 15 occasions were needed to achieve stability in educator language variables.

Conclusion: To advance our understanding of the language learning environments of preschool classrooms that serve children with ASD, researchers must understand how aspects of the measurement design in those environments, such as occasion, impact the inferences they make.

Adult–child interactions are considered to be significant contributors to language development in young children through their direct influence, functional relations to language outcomes, and malleability (e.g., Hart & Risley, 1995; Rowe & Snow, 2020). For young children with autism spectrum disorder (ASD), however, characteristics inherent to the disorder can and do alter the nature of interactions between the child and adult (e.g., National Research Council, 2001; Wetherby et al., 2004). Adults, such as educators and parents, are "charged with remediating the impairment and minimizing the disruption to the partnership across multiple contexts" (Burgess et al., 2013, p. 429) by using language-facilitating behaviors, such as opportunities to respond, statements, and follow-in comments (e.g., Warren, 2015).

Approaches to Understanding Educator–Child Interactions

Of the studies researchers have conducted with children with ASD to understand and try to enhance language learning opportunities, many researchers have turned to systematic observation methods to capture form and quality features of the educator–child interactions. For example, Sanders et al. (2016) gathered and analyzed 30-min video-recorded observations during free-play of 42 children with ASD. In a later study, Qian (2018) analyzed the same database, but included a larger sample from both self-contained and inclusive classrooms, and analyzed only 15 min of the observed time. As examples, Sanders et al. (2016) and Qian (2018) have provided some preliminary information about

[a]Department of Educational Psychology, University of Minnesota, Twin Cities, Minneapolis

Correspondence to Andrea L. B. Ford: bohxx001@umn.edu

Editor-in-Chief: Stephen M. Camarata
Editor: Mary Alt

Received October 2, 2020
Revision received November 11, 2020
Accepted January 15, 2021
https://doi.org/10.1044/2021_JSLHR-20-00586

Disclosure: *The authors have declared that no competing interests existed at the time of publication.*

the form, frequency, and quality features of language learning opportunities that young children with ASD might experience. For both studies, the researchers further reported high levels of interrater agreement (IRA). Reporting IRA in varying forms is a necessary feature of observational research to confirm that the target of measurement is being reliably measured by multiple observers. This reliance on examining IRA exclusively, however, often leads to assumptions and inferences about the generalizability of the data beyond the sampling context, without exploring sources of variance within that sampling context (Bottema-Beutel et al., 2014; Cone, 1977).

A Tradition in Unidimensional Reliability

Demonstration of the reliability of the estimated frequencies of the behaviors is a necessary, though not sufficient, precondition for an argument of validity (Kane, 1982; Suen & Ary, 1989). More specifically, before researchers can deem their conclusions as having a high degree of validity or accuracy, they must work to maximize the precision with which they estimate a true score or measure of behavior and minimize error within the measurement system (Bottema-Beutel et al., 2014). To operationalize reliability, researchers must demonstrate that an observed measure of behavior is consistent with another observed measure of the same behavior, which they achieve by calculating the association between the two against an appropriate, allowable threshold. This measurement is traditionally—and often solely—represented as IRA, wherein researchers calculate the consistency of two or more observers using metrics such as kappa or percent agreement (Suen & Ary, 1989; Yoder et al., 2018). Though one could conclude that high IRA indicates precision and minimized measurement error, this unidimensional approach to examining the reliability of observational data may not be complete.

An important next step for researchers is to consider reliability in the broader contexts and conditions in which they collected the measurements and plan to generalize their findings. For example, although both Sanders et al. (2016) and Qian (2018) had multiple observers rating the language educators used, their analyses were based on observations of only a single session, in one learning context (i.e., free-play), and one classroom type (i.e., inclusive for Sanders et al.). The extent to which the additional measurement conditions—observation occasion, learning context, and classroom type—contributed measurement error to observed frequencies is unexplored. There is a necessary opportunity to empirically examine parameters that may impact the generalizability of measured findings so that researchers are able to address the validity of their inferences, not just the reliability of their measured samples.

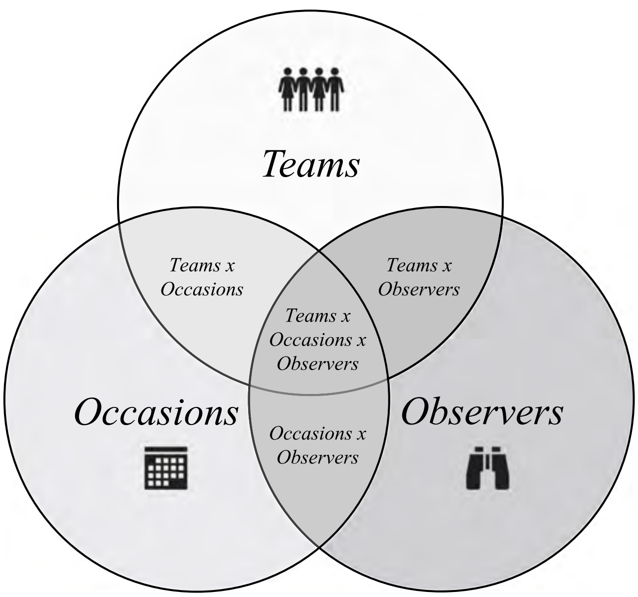
Utilizing Generalizability Theory to Examine Multidimensional Reliability

Generalizability Theory (G-Theory), initially developed and described by Cronbach et al. (1972) and expanded upon by Brennan (2001), may offer a potential solution to this problem. In the most basic description, G-Theory, and the studies within it (i.e., generalizability and decision), allow researchers to identify, quantify, and partition out potentially relevant sources of measurement error, called measurement facets (e.g., observer, occasion, item, setting, method, and dimension; Cone, 1977). Researchers can then compare the error attributed to the measurement facets to the error attributed to the object of measurement or differentiated facet, which is most often the individual or groups of individuals being observed (e.g., children, educators, teams of educators, classrooms; Brennan, 2001; Cronbach et al., 1972). As a variance attribution diagram (Brennan, 2001), Figure 1 illustrates the facets that likely contribute variance to observations of children's performance in unique and combined ways. Differentiating these sources of error gives researchers a means to understand how current and future configurations of measurement conditions may or may not promote stability, precision, and generalizability of estimated scores, frequencies, or durations of the desired behavior (Shavelson & Webb, 2006; Suen & Ary, 1989). Overall, G-Theory enables researchers to adopt a multidimensional approach to reliability, increasing the rigor of, comprehensiveness of, and empirical support behind their methodology to support the inferences made.

Generalizability Study to Understand Contribution

A generalizability study, or g-study, broadly entails separating and estimating the error that each identified facet contributes to a given set of gathered data (Brennan, 2001; Suen & Ary, 1989). Using an analysis of variance, a researcher calculates the percent of variance accounted for by the differentiated facet (i.e., *persons* or the object of measurement), measurement facet or facets (e.g., *occasion*), and

Figure 1. Total variance attribution diagram for the measurement design of the current study.



the interactions between them. The variance estimates for each source support researchers in identifying the impact of the measurement facets on (a) an individual or group's (e.g., team of educators) ranking against others for relative decisions or (b) the observed data when held against a specified criterion for absolute decisions. These estimates further allow researchers to quantify the precision with which they estimated the individual's or group's universe score (analogous to a true score; Shavelson & Webb, 2006).

Although the variance components promote an understanding of relative contribution, a clear benefit of G-Theory and a generalizability study is the ability to calculate absolute and relative reliability coefficients (Brennan, 2001). A reliability coefficient represents the extent to which the data gathered (i.e., the observed score) can be generalized to the larger universe of admissible observations, or all acceptable observed measures of behavior from a possible set of measurement conditions. To calculate the reliability coefficient, the variance of *persons* is divided by the variance of *persons* plus the variance from all other sources of variance (e.g., measurement facets; Brennan, 2001; Shavelson & Webb, 2006). One recommendation suggests that a coefficient of 0.80 or above represents a reliable observed score that provides a stable estimate of the larger universe (Cardinet et al., 2010).

Decision Study to Understand Optimization

Using the variance components estimated by the g-study, researchers can examine how future investigations could be optimized through a sampling approach that reduces the contribution of measurement error and results in more stable estimates of a behavior (Brennan, 2001). This exploration occurs as part of a decision study, or d-study. Through an iterative process, researchers systematically and statistically manipulate the number of levels of the measurement facets identified in the g-study (Cronbach et al., 1972). The researcher has flexibility in designing which facets to include within the d-study, as they may opt to include all the facets in the g-study or only subset (Suen & Ary, 1989). Researchers can also estimate absolute and relative reliability coefficients for each new set of conditions.

Study Purpose and Research Questions

Though a wealth of literature provides a strong empirical base for adult behaviors that can promote language gains in children with ASD (e.g., Hampton & Kaiser, 2016), we explored one potent language learning interaction behavior frequently used by early educators: *opportunity to respond* (OTR; Greenwood et al., 1994). When applied to language learning, a primary function of OTRs is to create an opportunity for a child to demonstrate some form of language related behavior (Ford et al., 2020; Pianta et al., 2009). For our investigation, we examined forms of explicit OTRs that educators used to evoke children's demonstrations of expressive language, referred to as opportunities for expressive language (OELs). Guided by the logic of G-Theory (Brennan, 2001; Cronbach et al., 1972), the primary purpose of this investigation was to examine a common approach to sampling the forms of OELs used by educators with children with ASD during free-play in an inclusive preschool classroom. First, we examined how a sampling approach related to the inferences made about educator–child interactions. Second, we statistically manipulated features of the sampling approach in ways that could inform enhanced reliability of measurements in future investigations. Two research questions guided this work:

1. When making inferences about the type of educator talk to which children with ASD are exposed, to what extent are the measurement facets of occasion and observer relevant?
2. Given these two measurement facets, under what conditions can the sampling methodology be optimized?

Method

Participants

We recruited educators and children from 11 early childhood inclusive preschool classrooms in a midwestern state who were already participating in a larger federally funded measurement study. Within each classroom, parents of the focal child with ASD and all educators consented to participate in procedures that were reviewed and approved by the university institutional review board.

Focal Children

We recruited only one child in each classroom who met the inclusion criteria for participation in the study, for a total sample size of 11. Although this sample size was small, we aligned it with other investigations that have gathered observational data and applied G-Theory models (e.g., Hill et al., 2012; Mantzicopoulos et al., 2018). To be included, the focal child needed to have qualified for early childhood special education services with primary eligibility in ASD. Additionally, the focal child needed to (a) use English as their first language, (b) use verbal language as a principal means of communication as reported by the speech-language pathologist, (c) have a goal for expressive communication in their Individualized Education Plan, and (d) have regular access to free-play. All participants were male and between the ages of 3 and 5 years ($M = 4.09$, $SD = 0.70$). The demographic breakdown of the focal children's race was 27.2% Asian, 9.1% Multiracial, and 63.6% White. One child identified as of Hispanic or Latino ethnicity. Table 1 provides the classroom demographic information for each child.

Teams of Educators

We wanted to examine the classroom language environment from the perspective of the focal child. To achieve this aim, we recruited all educators in the classroom of the focal child, yielding a total of 59 educators. Of this group, however, only 38 educators engaged in interactions with the focal child in their classroom during the recorded observations.

Table 1. Classroom demographic information for focal children.

		Nonfocal children		Educators		
Child	**Total number of children in class**	**Total number with IEP in addition to focal child**	**Disability areas identified on the IEP**	**Total number in the class**	**Total number that interacted with focal child**	**Role(s) of educators that interacted with focal child**
1	26	2	ASD; S/LD	5	3	GET, SEA, SET
2	13	6	ASD; DD	7	3	GET, SEA, SLP
3	15	3	DD; EBD; OHI; S/LD	5	2	GET, SEA
4	19	3	ASD; DD; S/LD	5	4	GEA, GET, SEA, SET
5	15	5	DD; S/LD	7	5	GET, SEA, SEA, SET, SLP
6	19	2	ASD; DCD; PI; S/LD	4	2	SEA, SET
7	18	1	ASD	7	4	GEA, GET, SEA, SEA
8	16	5	ASD; DD; S/LD	6	5	GEA, GET, SEA, SEA, SET
9	10	4	EBD;DD	4	2	SEA, SET
10	19	2	DD; EBD; S/LD	6	3	SEA, SEA, SET
11	20	3	DD	5	5	GEA, GET, SEA, SEA, SET

Note. The total number with an IEP is only known for other students that consented and provided enrollment information. IEP = individualized education plan; ASD = autism spectrum disorder; S/LD = speech/language disorder; GET = general education teacher; SEA = special education assistant; SET = special education teacher; DD = developmental delay; SLP = speech-language pathologist; EBD = emotional behavioral disorder; OHI = other health impaired; DCD = developmental cognitive delay; GEA = general education assistant; PI = physically impaired.

These educators were White and female, with 90% (n = 35) of participants indicating they were involved in a lot or all classroom routines. The rightmost column in Table 1 provides the breakdown by focal child, indicating the number of educators in the class and the number and role of educators whose interactions were captured across observation occasions.

Data Collection Procedures

We filmed each focal child for four sessions of 15 min each over 3 to 4 weeks, during free-play. To obtain a naturalistic sample of the focal child's language experience, we did not provide any specific instructions to the educators before filming; educators were told to interact as they normally do. Up to five educators who were regularly part of the free-play routine wore small microphones that were either attached with a clip to their shirts or on a lanyard. Filming for each session began when at least one educator was in the same play area of the classroom (e.g., kitchen area, art table, construction area) as the focal child. When an educator language behavior was (a) directed at the focal child and (b) fell into one of the educator behavior categories, we continued filming for an additional 15 min. Observers only coded from the onset of the language behavior to the end of the 15 min. A total of 44 videos (four videos from 11 classrooms) were collected and coded for specific study variables.

Study Variables

This investigation focused on verbally provided OELs that, (a) in their form, explicitly requested a response and (b) were directed to the focal child or to a group of children of which the focal child was part. Observers determined the type of OEL by the level of support provided within the question/statement/prompt. The following four types of OELs were coded: (a) open-ended questions/statements (e.g., "What color is it?"), (b) yes/no questions (e.g., "Do you like apples?"), (c) choice questions/statements (e.g., "Do you want green or blue?"), and (d) imitation prompts (e.g., "Say 'car'."). Because we were also interested in understanding how measurement error may be impacted when moving from highly specified variables to one broad variable, we aggregated these four codes into one category called *combined OELs*. Finally, recognizing educators frequently engaged in talk that was not explicit in evoking expressive language, we included two additional categories: statements and other. See Table 2 for definitions and examples of the mutually exclusive codes.

Data Coding and Reliability

To characterize the language-learning environment from the perspective of the focal child, we coded the directed language *any* educator used. We did not code the language or separate the data by the individual educator; rather, the data represents an aggregated frequency across educators, an approach consistent with previous investigations (e.g., Irvin et al., 2013; Sanders et al., 2016). For all coding of the study variables, we used Behavioral Observation Research Interactive Software (Friard & Gamba, 2016). This software allows for event logging of video recordings, allowing users to code both the occurrence of a discrete study variable and the time that variable occurred within the observation. During coding, users can pause and replay portions of the video as appropriate to ensure accurate coding.

During the observations, it became apparent that educators were frequently stringing multiple phrases and sentences together, which made determining the boundaries between codes challenging. Because the primary aim depended on characterizing the educator language to which a focal child was exposed, we decided it was necessary to

Table 2. Definitions and examples of educator language codes.

Code	Definition	Example
Open-ended	A question or statement used by the educator that is directed at the focal child and can be answered with numerous and unrestricted responses.	What do you want to eat? Where should we go? Tell me about your weekend.
Yes/no	A question used by the educator that is directed at the focal child and can be answered with a yes or no. This can be the formal (i.e., auxiliary-fronted) or informal questions use rising intonation or a tag at the end of the sentence. Tags may include isn't it, aren't they, don't you, and ok.	Do you like apples? It's hard, isn't it?
Choice	A question used by the educator that offers two or more explicit options from which the child may choose using verbal language that delineates the choice options OR visual supports (i.e., pictures, objects) that clearly delineate the choice options at the same time as the question.	Do you want trains or blocks? [Pointing to two options]: Which one is taller? Tell me if you want blue or green.
Imitation	A question or statement used by that requires a direct repetition of the word or phrase from the child. This can also represent a model for the child, with a prompt such as tell me or say.	Say "Ball." Can you say "more please?"
Statement	A comment used by the educator that is directed at the focal child that carries meaning in its use, such that it may label or describe, but is not used to evoke language. Statements can be single words, phrases, or sentences but must include at least one of the following parts of speech: nouns, verbs, pronouns, adverbs, adjectives, and/or prepositions.	That's a big tower. Let's clean up the toys.
Other	These are single words that do not fit in the categories described above. They may include common exclamations (e.g., whoa!), affirmations (e.g., yes, no, okay), greetings (e.g., hi, bye), or common sound effects (e.g., beep beep).	Uh-oh! Bummer!

Note. All codes are mutually exclusive. Open-ended, yes/no, choice, and imitation together represent combined opportunities for expressive language.

provide rules and guidelines for how to segment educator discourse that could then be behaviorally coded as discrete occurrences. For this purpose, we adopted Loban's (1966) notion of communication units, or c-units. C-units are frequently used in oral language analyses and are thought to preserve the meaning of interactions. This preservation is done by keeping the main clause and its modifiers together to inform the categorization of the type of interaction observed while accounting for pausing and intonation of adult talk to inform segmentation into discrete units (Eisenberg & Guo, 2013). Although c-unit is traditionally used when researchers work off transcription (e.g., Finestack et al., 2014), we adapted the logic of c-units for use while coding a video within the Behavioral Observation Research Interactive Software. That is, the two observers simultaneously segmented the educator language using the guidelines for c-units and categorized each communication unit following the variable definitions when coding. To promote efficiency, this segmentation and categorization co-occurred as part of a single activity completed by the observer during a single pass through the video.

The first author coded all videos. We also recruited one master's level data collector with experience in coding to independently code all videos and serve as the second observer. Before coding, the first author reviewed definitions, provided examples and nonexamples, and clarified questions. Each observer then independently coded a series of training videos until least 80% agreement on all individual codes across three consecutive videos was obtained.

We identified roughly 30% (n = 14) of the videos for IRA checks, spacing them out across the coding. Across all IRA checks for language coding, the observer mean percent agreement was 95% (range: 82%–100%) for open-ended questions/statements, 93% (range: 83%–100%) for yes/no questions, 97% (range: 75%–100%) for choice questions/statements, 98.6% (range: 80%–100%) for imitation, 85.4% (range: 71%–100%) for statements, and 85% (range: 60%–100%) for other.

Data Analysis Approach

Teams of educators were considered the differentiated facet (i.e., *persons* or the object of measurement) and is represented as *teams (t)* in all analyses. We had two measurement facets: (a) *occasion (o)* with four levels (i.e., the number of observations per focal child) and (b) *observer (r)* with two levels (i.e., the number of observers for each variable). The type of classroom and learning context were not included as measurements facets. Rather, they represent hidden facets as they only had one level, inclusive classroom and free-play, respectively.

Examining the Contribution of Occasion and Observer With Generalizability Studies

To answer the first research question, we modeled a fully crossed, *teams* × *occasions* × *observers* random effects analysis of variance for each study variable. Using EduG software (Swiss Society for Research in Education Working Group, 2012), we calculated the variance components for *teams*, *occasions*, *observers*, *teams* × *occasions* interaction, *teams* × *observers* interaction, *observers* × *occasions* interaction, and *teams* × *observers* × *occasions* interaction for each coded variable (see Figure 1).

Because of the fully crossed design, the variance components that were most relevant in the interpretation were *teams*, *teams* × *occasions* interaction, and *teams* × *observers* interaction (McWilliam & Ware, 1994). The interaction effect is important given that it quantifies the extent to which teams vary by occasion and observer. Although the main effects of *occasio*n and *observer* become less important in fully crossed designs when *teams* is not considered, the results of these sources of variance were included to provide a comprehensive picture of the error differentiation. The software also provided the percent variance accounted for by each source and the accompanying absolute reliability coefficients. We adopted the guideline for 0.80 or above as reliable when interpreting the results of the g-study (Cardinet et al., 2010).

Optimizing the Sampling Approach With Decision Studies

To answer the second research question, we conducted a d-study for each variable to determine the configuration of occasion and observers that would result in a sampling approach that provided precise, stable estimates of the specified behavior (Brennan, 2001). Again, using EduG software (Swiss Society for Research in Education Working Group, 2012), we systematically and statistically manipulated the levels of the measurement facets for those below criterion levels. This manipulation allowed us to forecast the number of occasions and the number of observers that, when combined, reduced the magnitude of the measurement error components, and in turn, optimized the reliability coefficients (Shavelson & Webb, 2006; Suen & Ary, 1989). We conducted this iterative process of data analysis until reliability coefficients were at or above 0.80 (Cardinet et al., 2010).

Results

Examining the Contribution of Occasion and Observer as Measurement Facets

Given that base rates and variability of measurement targets contribute to error, we first explored the average frequency with which each variable occurred when aggregated across all 44 sessions. Table 3 presents the descriptive statistics, as well as the variance estimations with percentages by source and absolute reliability coefficient, for each variable. *Teams* × *occasions* accounted for the largest percentage of total variability across all educator language variables (range: 53.6%–77.5%), while *teams* accounted for 13.6% to 45.6% of the total variance. For all variables, *teams* × *observers* accounted for less than 1% of the total variance. The estimated reliability coefficients were between 0.39 to 0.77, below the threshold of presumed reliability and suggesting a lack of precision in observed frequencies.

The percent variance accounted for by the interaction of *teams* × *occasions* for the aggregated category of combined OELs was the lowest among all educator language variables. This interaction, however, still accounted for more of the variability than *teams* and resulted in reliability coefficients that were below the 0.80 threshold. This variable also had higher reliability coefficients than each of the constituent variables (i.e., open-ended, yes/no, choice, and imitation).

Based on the results of the g-studies, we made two conclusions regarding the occasion and observer measurement facets. First, finding that the *teams* × *observers* interaction contributed little to no error across all measured variables indicates that any differences in the observed durations or frequencies between the two coders had little to no bearing on the total variability observed. Thus, observer was likely not a relevant measurement facet in this sampling approach. Second, the interaction of *teams* × *occasions* contributed a substantial amount to the measurement error for all variables; this contribution was higher than the contribution by *teams*. This result suggests that occasion was a relevant measurement facet when characterizing language adults use with children with ASD.

Optimizing the Sampling Approach

We conducted a series of planned decision studies on the same variables used in the g-studies. Given the results of the g-study, we defined occasion as the only universe of generalization and manipulated the levels within it to determine the minimum number of occasions needed to obtain a reliability coefficient of 0.80 (Cardinet et al., 2010). Although observer contributed little to no measurement error and could have been removed, we held it constant at one throughout all decision studies. This decision was made for two reasons: (a) to reflect the minimum number of observers required to code any data and (b) to be conversative in reporting optimal conditions for minimizing measurement error.

Figure 2 presents the absolute reliability coefficients for each of the educator language variables, from one to 15 occasions. Using the threshold of reliability of 0.80, it would take five to more than 15 occasions to produce stable estimates for some variables. More specifically, we would need to observe five occasions for combined OELs, seven occasions for imitation prompts, nine occasions for other language and yes/no questions, 12 occasions for open-ended questions/statements (0.81), and more than 15 occasions for statements and choices.

Discussion and Implications

By adopting the logic of G-Theory, researchers can differentiate sources of error and quantify the extent to which observed measures precisely represent the true measure (Cronbach et al., 1972). The current investigation aligned with this theory and had two primary aims: (a) understand

Table 3. Descriptive statistics, variance estimations by source, and reliability coefficient for each educator language variable.

	Specific types of educator language						
	Open-ended	**Yes/no**	**Choice**	**Imitation**	**Combined OELs**	**Statement**	**Other**
Descriptive statistics							
M (SD)	12.8 (7.0)	19.5 (9.9)	2.0 (2.5)	2.8 (3.2)	37.05 (15.2)	87.3 (42.0)	10.5 (7.3)
Range	0–32	0–54	0–11	0–16	12–91	19–201	1–34
Variance estimates (Percentages)							
Teams	13.8 (26.4)	34.3 (33.1)	0.9 (13.6)	3.9 (36.7)	111.8 (45.6)	391.0 (21.5)	18.6 (33.4)
Observers	0.0* (0)	0.0* (0)	0.0* (0)	0.0* (0)	0.1 (0)	0.0* (0)	0.0* (0)
Occasions	0.0* (0)	0.0* (0)	0.0* (0)	0.5 (0)	0.0* (0)	0.0* (0)	1.5 (2.6)
Teams × observers	0.1 (0.2)	0.1 (0.1)	< .01 (0)	0.1 (0.4)	0.3 (0.1)	5.1 (0.3)	0.4 (0.7)
Teams × occasions	37.9 (72.6)	67.8 (65.4)	5.7 (85.1)	5.8 (54.2)	131.4 (53.6)	412.9 (77.5)	31.0 (55.6)
Observers × occasions	0.1 (0.1)	0.2 (0.2)	0.0* (0)	0.0 (0.2)	0.2 (0.1)	0.0* (0)	0.5 (0.9)
Teams × observers × occasions	0.4 (0.7)	1.3 (1.3)	0.1 (1.2)	0.4 (3.9)	1.4 (0.6)	13.2 (0.7)	3.8 (6.8)
Total variance	52.2	103.7	6.7	10.6	245.2	822.2	55.7
Absolute reliability coefficient	0.59	0.67	0.39	0.70	0.77	0.52	0.68

Note. Percentages, in parentheses next to variance estimates, may not add up to 100 due to rounding. OELs = opportunities for expressive language.

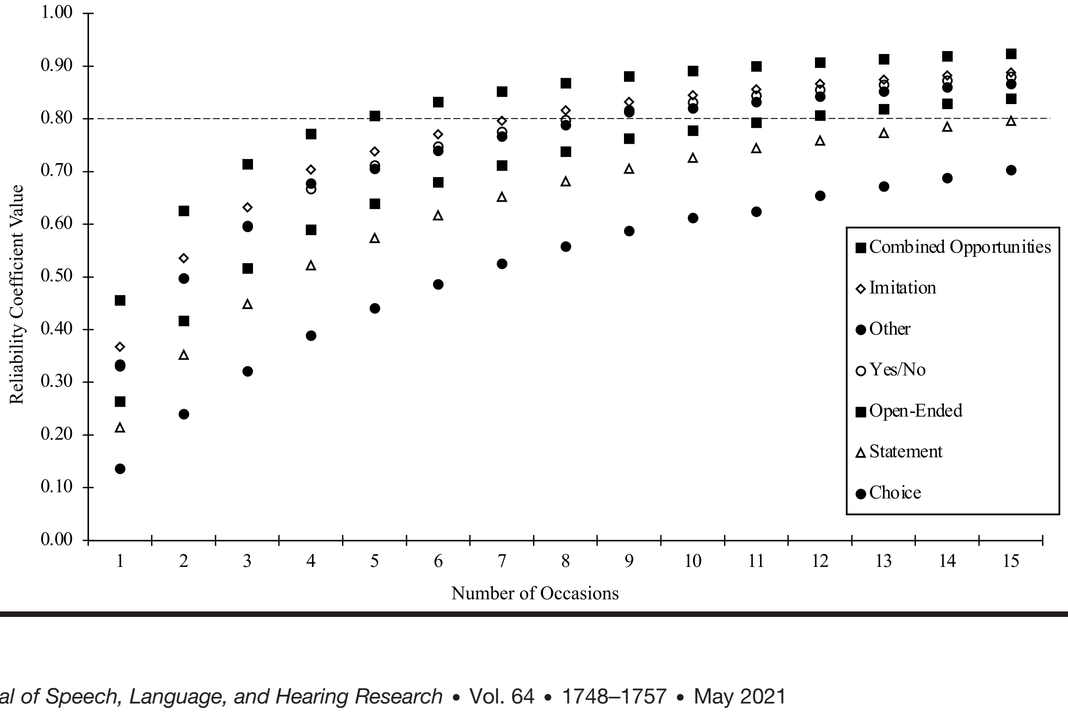
*Following guidance from Brennan (2001) and Cronbach et al. (1972), variance estimates that were negative were rounded to 0.00 and are indicated.

how occasion and observer contribute variability relevant to inferences made about the language educators use with children with ASD and (b) evaluate how the methodological approach could be enhanced in future studies.

The Need for a Multidimensional Approach to Reliability

Studies often rely solely on traditional, unidimensional approaches to reliability when evaluating the language learning environment (e.g., Qian, 2018; Sanders et al., 2016). Theoretically and practically, both occasion and observer, as measurement facets, have been observed to contribute measurement error to the gathered data (Bottema-Beutel et al., 2014; Brennan, 2001; McWilliam & Ware, 1994; Yoder et al., 2018). In fact, when compared to *teams* (the object of measurement), *teams* × *occasions* contributed the most variance for all variables, resulting in reliability coefficients below the recommended threshold of 0.80 (Cardinet et al., 2010). This result indicates that there was

Figure 2. Absolute reliability coefficients for educator language variables by number of occasions. Observer was held constant at one. The dashed line at 0.80 represents the reliability criterion.



limited precision and stability in the frequencies we estimated for all seven types of educator language.

The substantial contribution of occasion to measurement error when assessing interaction behavior, however, may not be all that surprising. First, previous literature suggests an apparent influence of occasion to observed behaviors of an individual, such that multiple observations may be recommended to achieve stable estimates for the context from which samples are gathered (Mantzicopoulos et al., 2018; Yoder et al., 2018). Second, the percentage of time in which at least one educator was available for interaction with the focal child was highly variable, which ranged from 20% of the session (i.e., 3.05 min) to 100% of the session (i.e., 15 min). This finding aligns other empirical investigations (e.g., Powell et al., 2008). Third, we coded and combined the directed language *any* educator used into a single aggregate frequency for the observation occasion. Two sources of variability—roles of the educators within the classroom (e.g., general educators, special educators, or assistants; Sawyer et al., 2018) and the number of different educators that could have interacted with the focal child in a given session (one to five educators)—were masked by the aggregation of language use across all adults into a single score. Although we aligned this method with current approaches in the field (e.g., Dykstra et al., 2013; Sanders et al., 2016), it may further explain why *occasion* contributes more variability to the observed scores and ultimately impacted the reliability (or lack of) we observed.

Though researchers may interpret their findings in ways that infer that their sampling of the language environment provided a true representation of the environment, the contribution of other measurement facets is a more significant problem than many consider (Yoder et al., 2018). In fact, occasion is frequently overlooked (Brennan, 2001) and thus a hidden facet in many investigations (e.g., Qian, 2018; Sanders et al., 2016). The role it played in the variance in the data in the current study warrants attention. Including occasion as a measurement facet, particularly when examining the language environment from the perspective of the focal child, seems to be supported not only by the data in this study and others (Hill et al., 2012), but also theoretically (Brennan, 2011). Beyond occasion, it would also behoove researchers to define the universe of generalization to include type of preschool classroom (e.g., levels of inclusive or self-contained) and learning context (e.g., levels of large group, small group, free-play, snack time) as additional measurement facets. The levels within each of these facets contribute variability in terms of the structural processes that create opportunities for educator–child interactions (e.g., Pianta et al., 2009). In combination, there is a need for researchers to adopt a multidimensional approach to their treatment of reliability.

Increasing the Precision and Generalizability of Measurements

Understanding how to enhance the sampling approach may offer important new avenues for researchers seeking ways to understand the language-learning environment for preschoolers with ASD. For example, if inferences were made that educators only used 1.99 choice questions in 15 min based on four brief observations, those inferences would be based on an unreliable measurement. The d-study provided important insight that, for some variables, more than 15 observations were required to achieve a stable frequency of a language behavior. Given a need to balance statistical rigor with practical cost, researchers must consider approaches to increase the precision and generalizability of observations.

One approach to increasing the stability of our measurements is to consider adjusting the type of OELs from the four prompting types in the current study to either (a) a more encompassing variable (e.g., combined OELs) or (b) being more selective of specific types. This logic is supported by the result that when the frequencies of these four variables were aggregated represent a single variable for combined OELs, the contribution of *teams* × *occasions* was less (53.6%) when compared to the contribution for each of the constituent variables (e.g., open-ended at 72.6%). This shift in the contribution of *teams* × *occasions* to the variance when moving from highly specified variables to a broad aggregated variable is further consistent with other literature (e.g., Bottema-Beutel et al., 2014; Hollo et al., 2020). Practically, this discussion of the contribution of occasion has implications for researchers when adopting either a molecular perspective or molar perspective to the measurement of language interaction behaviors (Baum, 2002). When measuring highly specified variables (i.e., molecular perspective), the study findings demonstrate that it may be necessary for researchers to increase their sampling across multiple occasions. On the other hand, as researchers examine broader groupings of language interaction behaviors (i.e., molar perspective), findings from this study demonstrate that fewer observations may be needed to make inferences about the language-learning environment.

Limitations

Several aspects of this study present limitations. First, the sample size was small with only 11 participants and 88 observations, though within the range of other investigations (Hill et al., 2012; Hollo et al., 2020; Mantzicopoulos et al., 2018). Simulated studies have shown that 50 to 300 data points can be robust enough for estimation of the variance components and *g*-coefficient (Atilgan, 2013), but the results should be considered preliminary. Second, we explicitly recognized two hidden facets, such that we conducted this investigation in one type of classroom (i.e., inclusive) and one learning context (i.e., free-play). The potential role or relevance of these two facets or item, method, and dimension as additional facets was not explored (Cone, 1977). As such, additional facets may have existed that were not only unknown to or unexplored, but were also distinct from, linked to, or confounded the observer or occasion facets in the current study. Third, demographic information on focal children was limited. Without information on the child's specific expressive language needs and developmental status, another potential source of variability remains unexplored—the interaction between a focal child's communication development and the educational team's delivery of certain types of OELs. Fourth, because segmentation and categorization occurred concurrently in coding, disentangling the cause of any reliability estimates that fell below expectations was difficult; despite this, IRA during coding was strong.

Conclusions

Issues in the reliability of the measurement procedures can thwart the validity of a researcher's inferences about a construct of interest. To examine the contribution of these procedures to the error, employing the logic of generalizability theory can be both useful and advantageous. In this study, we conducted a series of g-studies on measured educator variables, which highlighted occasion as a substantial contributor to measurement error and observer as virtually negligible for all variables. Depending on the variable of interest, a minimum of five and up to 15 observation sessions were required to obtain stable estimates of educator variables. To balance practical cost and technical rigor, it may be necessary to consider what is measured and empirically evaluate the impact of how it is measured in pursuit of accurately characterizing the language learning environments for children with ASD.

Acknowledgments

This work was supported by the National Center for Special Education Research, Institute of Education Sciences (Grant #R324A170032, PI: LeAnne Johnson). The statements do not necessarily reflect the views of our funders and all errors and omissions are our own. There are no conflicts of interest.

We would like to thank the educators, children, and families that supported this work. Without their support, this work would not have been possible. In addition, we would like to thank Drs. Frank Symons, Scott McConnell, Danielle Dupuis, Lizbeth Finestack, and Sheri Stronach at the University of Minnesota for their contributions, feedback, and support of this work as part of the first author's doctoral dissertation. Their mentorship was unparalleled and greatly appreciated.

References

Atilgan, H. (2013). Sample size estimation of g and phi coefficients in generalizability theory. *Eurasian Journal of Educational Research, 51,* 215–227.

Baum, W. M. (2002). From molecular to molar: A paradigm shift in behavior analysis. *Journal of the Experimental Analysis of Behavior, 78*(1), 95–116. https://doi.org/10.1901/jeab.2002.78-95

Bottema-Beutel, K., Lloyd, B., Carter, E. W., & Asmus, J. M. (2014). Generalizability and decision studies to inform observational and experimental research in classroom settings. *American Journal on Intellectual and Developmental Disabilities, 119*(6), 589–605. https://doi.org/10.1352/1944-7558-119.6.589

Brennan, R. L. (2001). *Generalizability theory*. Springer. https://doi.org/10.1007/978-1-4757-3456-0

Brennan, R. L. (2011). Generalizability theory and classical test theory. *Applied Measurement in Education, 24*(1), 1–21. https://doi.org/10.1080/08957347.2011.532417

Burgess, S., Audet, L., & Harjusola-Webb, S. (2013). Quantitative and qualitative characteristics of the school and home language environments of preschool-aged children with ASD. *Journal of Communication Disorders, 46*(5–6), 428–439. https://doi.org/10.1016/j.jcomdis.2013.09.003

Cardinet, J., Johnson, S., & Pini, G. (2010). *Applying generalizability theory using EduG*. Routledge. https://doi.org/10.4324/9780203866948

Cone, J. D. (1977). The relevance of reliability and validity for behavioral assessment. *Behavior Therapy, 8*(3), 411–426. https://doi.org/10.1016/S0005-7894(77)80077-4

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements*. Wiley.

Dykstra, J. R., Sabatos-DeVito, M. G., Irvin, D. W., Boyd, B. A., Hume, K. A., & Odom, S. L. (2013). Using the Language Environment Analysis (LENA) system in preschool classrooms with children with autism spectrum disorders. *Autism, 17*(5), 582–594. https://doi.org/10.1177/1362361312446206

Eisenberg, S. L., & Guo, L.-Y. (2013). Differentiating children with and without language impairment based on grammaticality. *Language, Speech, and Hearing Services in Schools, 44*(1), 20–31. https://doi.org/10.1044/0161-1461(2012/11-0089)

Finestack, L. H., Payesteh, B., Disher, J. R., & Julien, H. M. (2014). Reporting child language sampling procedures. *Journal of Speech, Language, and Hearing Research, 57*(6), 2274–2279. https://doi.org/10.1044/2014_JSLHR-L-14-0093

Ford, A. L. B., Elmquist, M., Merbler, A. M., Kriese, A., Will, K. K., & McConnell, S. R. (2020). Toward an ecobehavioral model of early language development. *Early Childhood Research Quarterly, 50,* 246–258. https://doi.org/10.1016/j.ecresq.2018.11.004

Friard, O., & Gamba, M. (2016). BORIS: A free, versatile open-source event-logging software for video/audio coding and live observations. *Methods in Ecology and Evolution, 7*(11), 1325–1330. https://doi.org/10.1111/2041-210X.12584

Greenwood, C. R., Hart, B., Walker, D., & Risley, T. (1994). The opportunity to respond and academic performance revisited: A behavioral theory of developmental retardation and its prevention. In R. Gardner III, D. M. Sainato, J. O. Cooper, T. E. Heron, W. L. Heward, J. W. Eshleman, & T. A. Grossi (Eds.), *Behavior analysis in education: Focus on measurably superior instruction* (pp. 213–223). Thomson Brooks/Cole Publishing Co.

Hampton, L. H., & Kaiser, A. P. (2016). Intervention effects on spoken-language outcomes for children with autism: A systematic review and meta-analysis. *Journal of Intellectual Disability Research, 60*(5), 444–463. https://doi.org/10.1111/jir.12283

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Brookes.

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher, 41*(2), 56–64. https://doi.org/10.3102/0013189X12437203

Hollo, A., Staubitz, J. L., & Chow, J. C. (2020). Applying generalizability theory to optimize analysis of spontaneous teacher talk in elementary classrooms. *Journal of Speech, Language, and Hearing Research, 63*(6), 1947–1957. https://doi.org/10.1044/2020_JSLHR-19-00118

Irvin, D. W., Hume, K. A., Boyd, B. A., McBee, M. T., & Odom, S. L. (2013). Child and classroom characteristics associated with the adult language provided to preschoolers with autism spectrum disorder. *Research in Autism Spectrum Disorders, 7*(8), 947–955. https://doi.org/10.1016/j.rasd.2013.04.004

Kane, M. (1982). A sampling model for validity. *Applied Psychological Measurement, 6*(2), 125–160. https://doi.org/10.1177/014662168200600201

Loban, W. (1966). *The language of elementary school children: A study of the use and control of language and the relations among speaking, reading, writing, and listening.* National Council of Teachers.

Mantzicopoulos, P., French, B. F., Patrick, H., Watson, J. S., & Ahn, I. (2018). The stability of kindergarten teachers' effectiveness: A generalizability study comparing the Framework For Teaching and the Classroom Assessment Scoring System. *Educational Assessment, 23*(1), 24–46. https://doi.org/10.1080/10627197.2017.1408407

McWilliam, R. A., & Ware, W. B. (1994). The reliability of observations of young children's engagement: An application of generalizability theory. *Journal of Early Intervention, 18*(1), 34–47. https://doi.org/10.1177/105381519401800104

National Research Council. (2001). Educating children with autism. In C. Lord & J. P. McGee (Eds.), *Commission on behavioral and social sciences and education*. National Academy Press. http://www.nap.edu/catalog/10017.html

Pianta, R. C., Barnett, W. S., Burchinal, M. R., & Thornburg, K. R. (2009). The effects of preschool education: What we know, how public policy is or is not aligned with the evidence base, and what we need to know. *Psychological Science in the Public Interest, 10*(2), 49–88. https://doi.org/10.1177/1529100610381908

Powell, D. R., Burchinal, M. R., File, N., & Kontos, S. (2008). An eco-behavioral analysis of children's engagement in urban public school preschool classrooms. *Early Childhood Research Quarterly, 23*(1), 108–123. https://doi.org/10.1016/j.ecresq.2007.04.001

Qian, X. (2018). Differences in teachers verbal responsiveness to groups of children with ASD who vary in cognitive and language abilities. *Journal of Intellectual Disability Research, 62*(6), 557–568. https://doi.org/10.1111/jir.12495

Rowe, M. L., & Snow, C. E. (2020). Analyzing input quality along three dimensions: Interactive, linguistic, and conceptual. *Journal of Child Language, 47*(1), 5–21. https://doi.org/10.1017/S0305000919000655

Sanders, E. J., Irvin, D. W., Belardi, K., McCune, L., Boyd, B. A., & Odom, S. L. (2016). The questions verbal children with autism spectrum disorder encounter in the inclusive preschool classroom. *Autism, 20*(1), 96–105. https://doi.org/10.1177/1362361315569744

Sawyer, B., Atkins-Burnett, S., Sandilos, L., Scheffner Hammer, C., Lopez, L., & Blair, C. (2018). Variations in classroom language environments of preschool children who are low income and linguistically diverse. *Early Education and Development, 29*(3), 398–416. https://doi.org/10.1080/10409289.2017.1408373

Shavelson, R. J., & Webb, N. M. (2006). Generalizability theory. In J. L. Green, G. Camili, & P. B. Elmore (Eds.), *Handbook of complementary methods in education research* (3rd ed., pp. 309–322). American Educational Research Association.

Suen, H. K., & Ary, D. (1989). Reliability: The generalizability approach. In *Analyzing quantitative behavioral observation data* (pp. 131–156). Erlbaum.

Swiss Society for Research in Education Working Group. (2012). *EduG User Guide (6.1)*. https://www.irdp.ch/data/secure/1968/document/EduGUserGuide.pdf

Warren, S. F. (2015). *Right from birth: Eliminating the talk gap in young children*. LENA Foundation. https://www.lena.org/wp-content/uploads/2016/10/RightFromBirth_Warren_5.12.2015_v.3.pdf

Wetherby, A. M., Woods, J., Allen, L., Cleary, J., Dickinson, H., & Lord, C. (2004). Early indicators of autism spectrum disorders in the second year of life. *Journal of Autism and Developmental Disorders, 34*(5), 473–493. https://doi.org/10.1007/s10803-004-2544-y

Yoder, P., Lloyd, B. P., & Symons, F. (2018). *Observational measurement of behavior*. Brookes.