

The Use of Generalizability Theory to Inform Sampling of Language Learning Environments for
Young Children with Autism Spectrum Disorder

A Dissertation
SUBMITTED TO THE FACULTY OF THE
UNIVERSITY OF MINNESOTA
BY

Andrea Lynn Boh Ford

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Dr. LeAnne Johnson, Advisor

April 2020

© Andrea Lynn Boh Ford, 2020

Acknowledgements

It takes a village to raise a speech-language pathologist and doctoral student into a Ph.D. In that vein, there are a number of individuals whom I would like to thank and acknowledge for not only this specific work, but also for their support in my process to acquire my doctoral degree.

First, words cannot describe my appreciation of my advisor, mentor, and colleague, Dr. LeAnne Johnson. From the beginning of my program, you have provided me with numerous opportunities that shaped my interests, deepened my knowledge and perspective, and provided me with the necessary skills to successfully engage in research. You have instilled a sense of practical purpose such that I will always work to keep the needs and interests of the students and educators we serve at the forefront. Finally, you believed in my work and, maybe mostly importantly, in me. I am eternally grateful.

I am sincerely grateful to my dissertation chair, mentor, and colleague, Dr. Frank Symons. You have taught me “writing is editing” and pushed me to think like a scientist. You were instrumental in guiding the design of this study. As we sat in LeAnne’s office and discussed options for my dissertation, I could not have predicted how much your casual mention of generalizability theory would forever change the way I think about measuring behavior.

I want to extend deep appreciation to my committee member, mentor, and colleague, Dr. Scott McConnell. When I joined your language seminar, I had no idea how much the content of the course, the in and out of class discussions, and our eventual paper would shape the way I conceptualize language development. It not only served as a guide for this work but will also guide my future research and teaching endeavors.

I am deeply appreciative of my committee members and colleagues, Dr. Danielle Dupuis and Dr. Lizbeth Finestack, who were willing to join our crew in the eleventh hour. Your advice, guidance, and support on this project and in my other research activities cannot be overstated.

I gratefully acknowledge all the children, parents, and educators who made this work possible. Your willingness to let me into your everyday routines is so appreciated and does not go unnoticed. Without out you, this work would not have been possible.

I would also like to thank my writing group, the B-JAMMers. You held me accountable to my goals and provided me with all the humor, fun, love, strategies, notebooks, and reinforcement a PhD could ever want.

I am incredibly appreciative of my dear friends, Erin Olson and Kimberly Kreines, for their edits and feedback to this paper. Your way with words, astute writing skills, and thought-provoking commentary only served to strengthen my argument and descriptions. I am so fortunate to have you in my village.

I am extremely grateful to my mentor, colleague, and friend, Dr. Linda Carpenter. From the moment I stepped foot in the Communication Sciences and Disorders program at UW-Eau Claire, you provided a level of mentorship that expanded my critical thinking and developed my appreciation for the world of research. You knew I was destined for academia and were always there to ever so slightly nudge me in that direction. Thank you for being that person.

(continued on next page)

I am especially appreciative of the friends and family who walked this path with me (whether you really wanted to or not). Your collective, constant, and unending support and positivity never faltered. You reminded me that there is life outside of a doctoral program and that life has wine.

I am eternally grateful to my sister, Kim, and my mom, Dawn. For my entire life and without fail, you have both been at my side in times of excitement and joy and in times when life does not go exactly the way I had planned it. You have taught me to take things one day at a time and that things will eventually all work out exactly how they should. You demonstrate the kind of strength, perseverance, positivity, and compassion it takes to find success in exactly what you want in life. I am so glad the universe gave me you as my family.

Finally, I am deeply indebted to partner in life, Brent. For the past six years, you have not only worked to feed our family while I was a student, but you also provided unending humor, love, fun, compassion, support, and sense of adventure. From ordering pizza and teaching me statistics to jumping in to help with coding of my dissertation videos when all hope was lost, you are a partner in every sense of the word. I cannot imagine this program and my life without you...and, Bernie too. I am so looking forward to our next adventure when I can (hopefully) get my weekends back.

Dedication

To my dad, Larry

Because your time in this world was too short to allow you to receive your own doctoral degree, without hesitation, I dedicate my dissertation and share the honor with you. Though you are no longer here, you have continued to influence not only my work, but also my outlook on this one great life we have.

You taught me that there are always solutions, but you need to be patient in finding them.

You taught me to channel my inner fire, take flight, and find my sense of adventure.

You taught me to love, respect, and approach people and world with care and kindness.

You taught me that you are never too old to be silly.

You taught me to be a leader and make my voice heard.

You taught me to always be willing to do something crazy and unexpected.

You taught me to always work hard, have fun, and do my very best.

So dad, I worked hard, had fun, and did my very best. I hope I have made you proud.

“Tis the gift to be simple, tis the gift to be free.”

Abstract

Features of autism spectrum disorder (ASD) can impact the nature, frequency, and length of adult-child interactions that are important for language learning. Empirical investigations of these interactions in preschool classrooms are limited and often provide minimal insight into the reliability of the observations beyond inter-rater agreement. To promote a multidimensional understanding of reliability and define optimal measurement procedures, the researcher employed the logic of Generalizability Theory to differentiate sources of error, namely persons (i.e., educator or child participants) and two measurement facets or conditions, occasion and observer. The researcher video-recorded 4, 15-minute occasions of educator-child interactions for each of the 11 participants with ASD during free play in their respective inclusive preschool classrooms. Two trained observers coded all videos for variables of educator proximity, educator language (i.e., open-ended, choice, yes/no, imitation, statement, and other), and child language. The generalizability studies illustrated that, across all variables measured, observer accounted for little to no error. Occasion, however, accounted for the majority of the error for all behaviors except child language. To determine the number of occasions needed to achieve stable estimates of the variables, the researcher manipulated occasion in the decision study. Although researchers would need only three occasions to reliably estimate child language, five to more than 15 occasions were needed for educator language and proximity variables. With a need to balance statistical rigor with pragmatics, refining, eliminating, or identifying new variables may be a necessary step toward making inferences about the language learning environments of young children with ASD.

Table of Contents

Acknowledgements	i
Dedication.....	iii
Abstract.....	iv
Table of Contents.....	v
List of Tables	vi
List of Figures.....	vii
Chapter 1: Introduction.....	1
Chapter 2: Literature Review	7
Chapter 3: Method.....	29
Chapter 4: Results.....	50
Chapter 5: Discussion.....	67
References	87

List of Tables

Table 1: <i>Demographic Information for Target Children</i>	29
Table 2: <i>Classroom Demographic Information for Target Children</i>	31
Table 3: <i>Demographic Information for Educators</i>	33
Table 4: <i>Definitions and Examples of Language Codes for Educator and Child Variables</i>	41
Table 5: <i>Variance Estimations, Percentages, and Standard Errors by Source for the Proximity Variable</i>	56
Table 6: <i>Spearman’s Rho Correlations between the Proximity Variable and Each Language Variable</i>	57
Table 7: <i>Variance Estimations by Source for the Educator Language Variables</i>	58
Table 8: <i>Variance Percentages by Source and Reliability Coefficients for the Educator Language Variables</i>	58
Table 9: <i>Variance Standard Errors by Source for the Educator Language Variables</i>	59
Table 10: <i>Variance Estimations, Percentages, and Standard Errors by Source for Child Verbalization</i>	60

List of Figures

Figure 1: Conceptual Model for the Relationship between Interaction Behaviors, Adult-Child Interaction Context, and Factors that Impact Interaction	8
Figure 2: Conceptual Model of Factors that Impact Measurement of Educator-Child Interactions in Preschool	18
Figure 3: Decision-Making Hierarchy for Proximity Coding	37
Figure 4: Example of Classroom Play Area Map.....	38
Figure 5: Decision-Making Hierarchy for Directed Language Coding	39
Figure 6: Variance Attribution Diagram for the Current Investigation.....	46
Figure 7: Box Plot for Proximity Variable	50
Figure 8: Density Plot for Proximity Variable	50
Figure 9: Box Plots for Educator Language Variables.....	52
Figure 10: Ridgeline Plot for Educator Language Variables	52
Figure 11: Box Plot for Child Language Variable.....	53
Figure 12: Density Plot for Child Language Variable.....	54
Figure 13: Relative and Absolute Reliability Coefficients with Observer Manipulated.....	62
Figure 14: Relative and Absolute Reliability Coefficients for Proximity	63
Figure 15: Relative Reliability Coefficients for Educator Language Variables.....	64
Figure 16: Absolute Reliability Coefficients for Educator Language Variables.....	64
Figure 17: Relative and Absolute Reliability Coefficients for Child Language Variables	65

Chapter 1: Introduction

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., Bronfenbrenner, 2001; Dunst et al., 1990; Hart & Risley, 1995; Hoff, 2006; Pruden et al., 2006; Rowe & Snow, 2019; Warren, 2015; Warren & Walker, 2005; Weisleder & Fernald, 2013). For many children who are developing in typical ways, their interactions with adults occur naturally, frequently, and are sustained. For young children with autism spectrum disorder (ASD), however, characteristics that are inherent to the disorder can and do alter the nature of interactions between the child and adult (Girolametto et al., 2000; National Research Council, 2001; Rimm-Kaufman et al., 2003). In previous research, young children with ASD were observed to (a) respond less frequently to bids for interaction; (b) initiate interaction less frequently; (c) develop idiosyncratic or unconventional communication behaviors (e.g., self-injurious behavior, aggression, or tantrums); (d) have reduced frequency and complexity of sounds, words, and sentences; and (e) engage in fewer conversational turns (Carpenter & Tomasello, 2000; Warren et al., 2010; Wetherby et al., 2004).

It is these deficits that require additional supports and scaffolds for young children with ASD to achieve the benefits that other children obtain through routine and regularly occurring interactions (Haebig et al., 2013). Adults, such as educators and parents, are “charged with remediating the impairment and minimizing the disruption to the partnership across multiple contexts” (Burgess, Audet, & Harjusola-Webb, 2013, p. 429) through the use of critical behaviors within language learning interactions, such as opportunities to respond, statements, expansions, and follow-in comments (McDuffie & Yoder, 2010; Warren, 2015). A relatively large number of studies have worked to capture information about interactions with parents (e.g., Casenhiser et al., 2013; Freeman & Kasari, 2013; Hudry et al., 2013; Siller & Sigman, 2008; Warren et al., 2010). Fewer studies, however, have worked to capture and understand how educators are promoting language development through the use of these critical behaviors broadly and opportunities to

respond in particular (e.g., Dykstra et al., 2013; Irvin et al., 2013; Qian, 2018; Sanders et al., 2016).

Approaches to Understanding Educator-Child Interactions

Of the limited number of studies researchers have conducted with children with ASD in preschool environments, researchers have generally adopted one of two methods to measure these educator-child interactions. Some researchers have used automated audio recording systems, such as Language ENvironment Analysis (LENA), which are worn by young children throughout their day (Dykstra et al., 2013; Irvin et al., 2013). The LENA system automatically processes the captured audio and provides quantitative information on three primary, global variables—adult word count, child vocalizations, and child turns, or the alternation between adult and child talk (Warren et al., 2010). This approach may provide an efficient means for gathering and summarizing relevant data on important aspects of language learning and support the generalization of findings to a broader scope of contexts. The device, however, records any language within a 6-foot radius of the target child. With this level of inclusivity, researchers cannot obtain a level of specificity around what was actually directed to, involved, and thus supported the language development of the target child (Irvin et al., 2013). Unfortunately, mere exposure to or overhearing speech is not predictive of later vocabulary knowledge (Hoff, 2006; Kuhl et al., 2003; Shneidman et al., 2013). Likewise, although the audio recordings can be transcribed and later analyzed, the automated summaries, with a focus on quantity, do not provide detailed information about another essential element of educator-child interaction for language learnings: the quality of the educator language. That is, these summaries do not provide information about the specific conceptual, linguistic, and interactional features of the educator input, such as the level of abstractness or difficulty of the questions asked, grammatical complexity or level of support in the language used, or responsiveness of the input (Rowe & Snow, 2019; Warren et al., 2010).

To offset these limitations, and when interested in understanding quality features of the educator-child interactions, researchers frequently turn to systematic observation methods. This approach involves behavioral coding of key variables functionally related to or associated with the achievement of the specified outcome (Bakeman & Gottman, 1997; Yoder et al., 2018). For example, Sanders et al. (2016) gathered and analyzed video-recorded observations of 42 children with ASD that were collected as part of a more extensive efficacy study. In an effort to understand the type of response opportunities children with ASD encountered, this team of researchers behaviorally coded the level of cognitive complexity to questions that any educator in the inclusive classroom (e.g., lead teacher, paraprofessionals, or related service providers) directed to a focal child during one, 30-minute session of free play. In a later study, Qian (2018) analyzed the same database, but included a larger sample of both self-contained and inclusive classrooms and analyzed only 15 minutes of the session. In addition, this researcher instead coded the form of educator input when following the focus of attention of the student with ASD, or the form of their responsiveness in terms of follow-in directive for behavior, follow-in directive for language, and follow-in comments.

As a result of their analyses, both Sanders et al. (2016) and Qian (2018) have provided some preliminary information about the preschool language environment broadly and opportunities to respond specifically in which young children with ASD participate. First, Sanders et al. (2016) found that educators directed primarily management questions (e.g., “Can you put your shoe on?”) to their students with ASD, with an average of 19.97 questions per session. In contrast, they observed educators directing only 10.00 cognitively challenging questions (e.g., “Who has more crayons?”) and 14.22 less challenging questions (e.g., “Do you want the blue or red crayon?”) to their students with ASD. Second, Qian (2018) found that on average and per minute, educators used 0.93 follow-in directives for language (e.g., “What color do you want?”), 0.86 follow-in comments (e.g., “You have a blue truck”), and 0.84 follow-in directives for behavior (e.g., “Put the blue truck on the table”). Although their work did not relate

observed frequencies to key proximal or distal language outcomes to guide intervention efforts, these investigations are the first to provide descriptors of the quality features of language and the frequencies with which specific types of opportunities to respond were occurring.

Statement of the Problem

For both studies described above, the researchers further reported high levels of inter-rater agreement (IRA). This result was interpreted as a demonstration of the reliability of the estimated frequencies of the behaviors, which is a necessary, though not sufficient, precondition for an argument of validity (Gast & Ledford, 2014; 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 in this context, a true 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—conceptualized as inter-rater agreement (IRA), wherein researchers calculate the consistency of two or more observers using metrics such as kappa or percent agreement (Gast & Ledford, 2014; McWilliam & Ware, 1994; Suen & Ary, 1989; Yoder et al., 2018). Applying this operationalization specifically to the investigations by Sanders et al. (2016) and Qian (2018), one could conclude that estimates of educator behavior from both were precise and that measurement error was minimized, given the high degree of reliability between observers. Although this conclusion would not be inaccurate and is aligned with traditional and unidimensional conceptualizations of the reliability of observational data (McWilliam & Ware, 1994; Suen & Ary, 1989), it may not be complete.

Researchers and readers may be remiss if their conclusions about the reliability of these particular measurements—and even more broadly, any measurement taken—end with this classic

conceptualization. That is, rather than considering reliability as unidimensional, it would behoove researchers to consider reliability in the broader contexts and conditions in which they collected the measurements and plan to generalize. 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 type of classroom (i.e., inclusive for Sanders et al.). As such, the extent to which the additional measurement conditions—observation occasion, learning context, and type of classroom—contributed measurement error to observed frequencies is unclear and unexplored. This lack of empirical information may call into question the precision and generalizability of the gathered data and in turn, the validity of the inferences made.

A Potential Solution

Generalizability Theory (G-Theory), initially developed and described by Cronbach, Gleser, Nanda, and Rajaratnam (1972) and expanded upon by Brennan (2001), may offer a potential solution to this problem of precision and generalizability. 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., scorer/observer, time/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, most often the persons or participants (Brennan, 2001b; Cronbach et al., 1972). Differentiating these sources of error through an analysis of variance (ANOVA) gives researchers a means to understand how current and future configurations of measurement conditions may or may not promote stability 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 in ways that critically support the kinds of inferences they plan to make.

Study Purpose and Research Questions

The previous efforts of Sanders et al. (2016) and Qian (2018) and the logic of G-Theory (Brennan, 2001b; Cronbach et al., 1963, 1972) guided the design of the current study. The purpose of this investigation was to examine an approach to sampling key variables related to the language learning environment of young children with ASD during free play in an inclusive preschool classroom. Given this overarching purpose, the specific aims of the study were two-fold. First, the researcher examined how a sampling approach intended to measure the language used by educators and children with ASD related to the inferences made about educator-child interactions important for promoting language development. Equipped with this understanding, the researcher secondarily aimed to systematically and statistically manipulate features of the sampling approach in ways that could enhance the reliability of measurements in future investigations. To that end, the researcher addressed the following research questions in this investigation:

1. When observing interactions of preschoolers with ASD with adult educators during free play to make inferences about the language learning environment, 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?

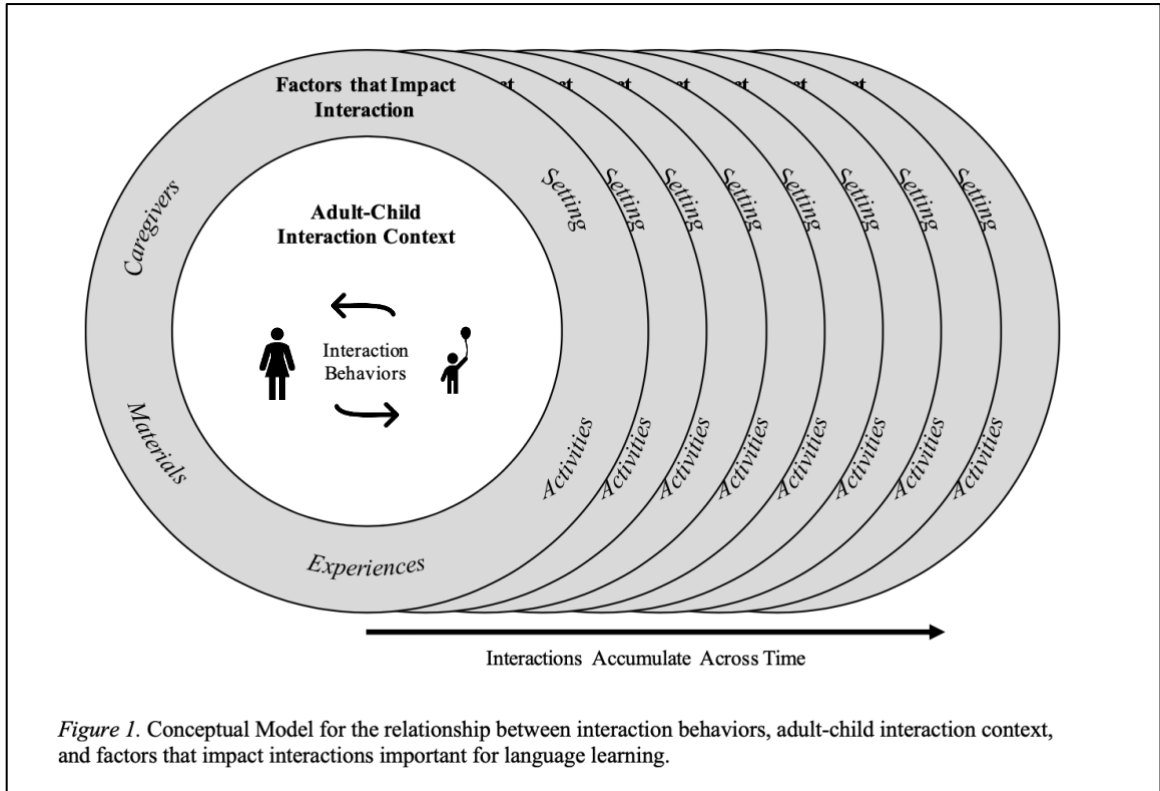
Chapter 2: Literature Review

Language development is an essential milestone in early childhood that has implications for participation in daily activities, social interactions, and later achievement (Fernald & Weisleder, 2011; Hammer et al., 2010; Light & McNaughton, 2014). An extensive body of literature exists that seeks to identify, understand, and examine variables that relate to and causally impact this development (e.g., Hart & Risley, 1995; Hoff, 2005; Romeo et al., 2018; Rowe & Snow, 2019). Identified variables such as brain processes (Kuhl, 2010), attention and memory mechanisms (Hollich et al., 2000), and experience within social systems (Bronfenbrenner, 2001; Fernald et al., 2013; Tomasello, 2003) are important to our comprehensive understanding of language development. These variables, however, may not provide a clear path to intervention when disparities and challenges in language acquisition arise. Adult-child interactions *as a context* and interaction behaviors within them *as the mechanism* provide an actionable context for intervention given their malleability and proximal impact on language outcomes (Hart & Risley, 1995; Rowe & Snow, 2019; Warren, 2015).

Adult-Child Interactions as the Context

It is generally agreed that communicative interactions with adults (e.g., educators, parents, grandparents, caregivers) are essential for a young child's language development (Dunst et al., 1990; Golinkoff et al., 2018; Hart & Risley, 1995; Romeo et al., 2018; Rowe & Snow, 2019; Warren, 2015). Some describe adult-child interactions as the “engines of development” (Bronfenbrenner & Morris, 2006, p. 996). Consistent with a bio-ecological approach to human development broadly (Bronfenbrenner, 2001) and an ecobehavioral approach to language development specifically (Ford et al., 2020), a range of factors can influence the timing, frequency, duration, and content of behaviors within the interactional context. These factors are more distal to the actual behaviors within interaction and may include, but are not limited to the caregiver behaviors, knowledge, and beliefs (Rowe, 2008); environment and resource availability (e.g., books and daily routines; Dunst et al., 2011; Rodriguez et al., 2009); and policies and

practices (e.g., parental leave and high-quality early care; Berger et al., 2005; Christina & Goodman, 2005). Figure 1 provides a conceptual model that illustrates the relationship between the adult-child interactions and the distal factors that are pertinent to the current investigation.



The conceptual model in Figure 1 recognizes and integrates specific elements of the ecobehavioral model presented in Ford et al. (2020)—namely, adult-child interactions, time, caregiver behaviors, and environmental influences. First, the model places adult-child interactions at the center, with the influential variables surrounding this crucial context. Second, the model illustrates the dynamic and accumulating nature of these interactions over time. That is, the influential factors may not only vary from one instance to the next but also build from a single instance (i.e., one set of circles) to a series of instances. Over time, these adult-child interactions are hypothesized to promote complexity and a greater number of communicative interactions (Sameroff, 2009; Warren, 2015).

Different from the Ford et al. (2020) ecobehavioral model, however, the model presented in Figure 1 provides a more pointed focus on caregiver behaviors and environmental factors that may be particularly influential on language learning. When examined relative to children's language learning, the model specifically identifies the caregiver behaviors as interaction behaviors (these behaviors will be described in more detail in the next section). In addition, this model identifies environmental influences in terms of their potential source of variability when trying to characterize the interaction context created by adult behaviors. These influences include, but are not limited to, setting (e.g., home, school, and community), materials (e.g., books, household items), experiences (e.g., zoo, museums), and activities (e.g., mealtime, bath time, circle time). Finally, the goal of this conceptual model is not to characterize the broader systems in which interactions take place. Instead, the goal is for it to provide a more focused and nuanced description of the specific variables that impact the frequency, rate, and duration of the interaction behaviors. In that way, this model may be used to not only inform under what conditions current practices and interventions may be useful and generalizable (Gutiérrez & Penuel, 2014), but also inform how we measure the interaction behaviors themselves.

Interaction Behaviors as a Critical Mechanism

Generally, the adult-child interaction context can set the occasion for a mutually shared experience. From a theoretical perspective, this shared experience is an essential element that supports the back-and-forth volley of interactions or the bi-directional and transactional exchange of information (Kublin et al., 1998; Owens, 2016; Rowe & Snow, 2019; Sameroff, 2009; Tomasello & Farrar, 1986). The arrows going between the child and adult illustrate this element in Figure 1. On a fine-grained level, these adult-child interactions also set the stage for an important mechanism in language acquisition: adult interaction behaviors. These behaviors carry specific features, dimensions, or qualities that, when maximally supportive of young children's involvement, serve as a fundamental, necessary, and critical precursor to language acquisition (Dunst et al., 1990; Rowe & Snow, 2019).

A wealth of literature exists that provides a strong empirical-base for adult interaction behaviors that have been shown to promote language gains in children with a range of delays and disabilities (e.g., see reviews by Akamoglu & Meadan, 2018; Hampton & Kaiser, 2016; Kaiser et al., 2001; Meadan et al., 2009). The goal here is not to review each of these strategies in detail, but rather provide a brief description of one critical interaction behavior: opportunity to respond. The primary function of opportunities to respond is to set the occasion for interaction and evoke language behaviors from a child as a way to promote and sustain interactions; it is a powerful and potent contributor to language learning (Dunst et al., 1990; Greenwood et al., 1984) and thus, worthy of exploration.

Opportunity to Respond

According to Greenwood, Delquadri, and Hall (1984), an opportunity to respond (OTR) is defined as “the interaction between: (a) an educator formulated instructional antecedent stimuli (i.e., the material presented, prompts, questions asked, signals to respond, etc.), and (b) their success in establishing the academic responding desired or implied” (p. 64). Although the goal of an OTR is to successfully evoke a desired verbal or non-verbal behavior of a child, the researcher in this investigation is particularly interested in OTRs that evoke verbal language or set the occasion for a child to respond with a form of expressive language. Early childhood educators often provide these OTRs to a child or a group of children in the form of questions or statements that, in their structure, explicitly request a response. Examples of these OTRs may include, but are not limited to: (a) open-ended, wh- questions (e.g., “What is the weather today?” or “Where did he go?”); (b) choice questions (e.g., “Do you want milk or water to drink?”); (c) yes/no questions (e.g., “Do you want to go outside?”); and (d) imitation prompts (e.g., “Say *ball*”) (Gest et al., 2006; Hart & Risley, 1975; Hepting & Goldstein, 1996; Kontos, 1999; Rowe et al., 2017).

For young children, the structure of the OTR needs to be challenging, while still within a child’s zone of proximal development to promote success and learning (Rowe & Snow, 2019; Vygotsky, 1978). The zone of proximal development refers to the level of functioning or set of

skills a child can achieve with guidance and support provided by adults but is not yet ready to do independently (National Research Council, 2001; Shonkoff & Phillips, 2000; Vygotsky, 1978). Pragmatically, the form of the OTR approximates this learning condition through the use of modifying questions and verbal directions that educators match to a child's developmental level. The form of this modification unfolds through a system of least-to-most prompts, wherein the educator plans and delivers support beginning with limited prompting (i.e., open-ended questions) and builds in intensity (i.e., choice, yes/no, imitation) to achieve a response from the child (Doyle et al., 1988). From the perspective of the child, this increasing level of support also serves to constrain his/her response options. For example, open-ended questions invite a more extensive range of and potentially longer responses, whereas choice, yes/no, and imitation invite a limited range and shortened length of response (De Rivera et al., 2005; Rowe et al., 2017; Wittmer & Honig, 1991). Researchers have found these different types of opportunities for expressive language—provided with appropriate scaffolding and attuned to the child's developmental level—to be effective in building vocabulary, developing complex language and question forms, and supporting verbal responses in young children (Cristofaro & Tamis-LeMonda, 2012; Hall et al., 1982; Rowland et al., 2003).

The Impact of Autism Spectrum Disorder on Interaction Behaviors

The interaction behaviors of adults, such as OTRs for expressive language, can be altered by the presence of specific child characteristics (Dunst et al., 1990; Dykstra et al., 2013; Landa, 2007; Spiker et al., 2002). For children with ASD, characteristics inherent to the disorder, such as impairments in communication, social reciprocity, and social interaction, impact both the nature and success of adult interaction behaviors (American Psychiatric Association, 2013). Briefly, these impairments often translate into a child's limited or lack of involvement in an interaction with an adult, as he/she may have difficulty: (a) responding to interaction behaviors initiated by parents, educators, caregivers, and/or peers; (b) initiating an interaction with others; or (c) sustaining the turn-taking or back and forth nature of communicative exchanges (American

Speech-Language-Hearing Association, 2017; Landa, 2007; Wetherby et al., 1998). In addition, some language used by children with ASD is self-directed or not directed at a level expected of their typically developing peers. In effect, it serves no functional communicative purpose that would indicate their involvement in interaction with their communication partner (Paul, 2008).

The need for the field to understand the frequency, form, and success of OTRs used by adults interacting with children with ASD cannot be overstated. This kind of descriptive research not only informs what problems need to be solved and areas of intervention, but also supports researchers in developing and evaluating interventions that promote adult use of interaction behaviors known to promote outcomes (Loeb et al., 2017). Toward that end, investigations of the frequency, form, and success of OTRs have begun, with a large number of studies working to capture information about interactions with parents (e.g., Casenhiser et al., 2013; Freeman & Kasari, 2013; Hudry et al., 2013; Siller & Sigman, 2008). Only a small number of studies, however, have aimed to capture and understand the nature of these differences for educators in preschool classroom environments (e.g., Dykstra et al., 2013; Irvin et al., 2013; Qian, 2018; Sanders et al., 2016).

Observed Impact in the Preschool Classroom

Researchers have recently begun to explore the interactions between children with ASD and their educator by measuring the amount of language used in the preschool environment through automated audio recording systems (i.e., Language ENvironment Analysis [LENA]). For example, using 40 children with ASD to gather information on adult-child language, Dykstra et al. (2013) found that, on average, children were exposed to over 1,700 adult words across one preschool day, exceeding language heard within the home environment (Warren et al., 2010). The researchers suggested that the difference found between the settings may be reflective of the nature of preschool environments themselves, as there are additional adults with varying roles (e.g., general educator, special educator, paraprofessional) who are interacting with each child. Extending the work of Dykstra et al., Irvin and colleagues (2013) found that for children with

ASD and comorbid significant cognitive delays, educators, on average, used fewer total number of words than with children with less severe cognitive delays. Whether this decrease is reflective of the characteristics of the child which impact their interactions or that educators may have been using fewer words to match the child's language level, the direct impact on language development was unexplored and remains to be seen.

Although the investigations by Dykstra et al. (2013) and Irvin et al. (2013) are undoubtedly useful in characterizing the preschool language environment of children with ASD, the inferences made from these findings are restricted mainly to the quantity of language input. As described by Rowe and Snow (2019), “while the quantity of input clearly plays a role, the quality of input is often found to matter more” for language learning (p. 5). Rowe and Snow describe language quality on three levels: (a) conceptual content, such as topics of conversation introduced by adults that adapt to the child's developmental stage; (b) linguistic complexity and redundancy, such as the phonological, lexical, and grammatical features of the adults' language attuned to the child's developmental stage; and (c) interactional features, such as adults being responsive and contingent upon the child's behavior, creating a mutually shared experience, and providing opportunities to interact or respond (e.g., asking questions).

Though limited, researchers are beginning to more closely examine an element of language quality—the interactional features of language broadly and opportunities to respond specifically—used by educators in preschool classrooms with children with ASD. To explore this specific quality feature, researchers have frequently turned to transcript analysis or coding of behaviors within observations of adult-child interactions. For example, using LENA recorders to identify segments of interactions that researchers transcribed and analyzed, Burgess, Audet, & Harjusola-Webb (2013) found that 67% of the utterances used in some center-based classrooms with children with ASD ($n = 10$) represented statements which may not evoke language. Only 25% of adult utterances were questions that evoked more complex communication (i.e., not yes/no questions). Investigating more closely the kinds of questions used by educators, Sanders et

al. (2016) evaluated the behaviors of 29 licensed educators working with 42 children with ASD between 3-5 years. These researchers determined that questions asked of the children related to regulation of behavior (e.g., management), rather than cognitively challenging questions that promote more complex language development about events and materials present or not present within the environment. Analyzing a similar database, but including a large sample within it, Qian (2018) found that children with lower levels of cognitive and language ability received fewer directives for language (commensurate with cognitively challenging questions in Sanders et al., 2016) and more directives for behavior.

These quality characteristics may explain how, despite efforts to identify and intervene on communication impairments early, some estimates suggest that 25% to 30% of children with ASD remain non-verbal or minimally-verbal upon kindergarten entry (Anderson et al., 2007; Norrelgen et al., 2015). Some long-term estimates even indicate that between 5% - 45% of individuals with ASD never develop the ability to speak (Brignell et al., 2018). These estimates are particularly troublesome in light of research, which suggests strong relationships between verbal ability in preschool and later reading achievement (e.g., Blese et al., 2016; Scarborough, 2001). In a follow-up study of 58 children with ASD, eight years after their initial evaluation in a clinical setting, Venter et al. (1992) identified the presence of verbal language before the age of 5 years as a significant predictor of academic achievement and adaptive development. It seems possible that these early deficits will have an impact on the development of future academic skills and educational attainment for the more than 20,000 preschoolers with ASD at risk for significant expressive language delays (U.S. Department of Education, 2018).

The Need for More Reliable Information

With more than 83,307 preschoolers who are served nationally under the category of ASD through the Individuals with Disabilities Education Improvement Act (IDEIA; U.S. Department of Education, 2018), the work of early childhood educators—general educators, special educators, teaching assistants, and related service providers—is directly affected. With

their expertise in child development and regular connection to children with ASD served in the preschool setting, early educators have a clear opportunity for engaging in interaction behaviors that are supportive of language development. Specifically, early childhood educators have a clear opportunity to provide children with ASD frequent, intense, contextually relevant, and appropriately challenging OTRs that invite the child's participation (Barton & Smith, 2015; Conn-Powers et al., 2006; DEC/NAEYC, 2009; Dunst et al., 1990; Rowe & Snow, 2019). Unfortunately, for researchers using existing literature to describe the current language learning environment as a means to identify areas where intervention is indicated, their inferences may be limited by the sampling methods employed within investigations.

When researchers, educators, and policymakers seek to draw conclusions about the language learning environment of children with ASD from existing research, an understanding of the sampling approach is vital. That is, specific environments and settings, constrained by specific aspects of time, impact the inferences researchers make about interaction behaviors and the language environment (Brennan, 2001b; Cone, 1977). The methodological designs of Sanders et al. (2016) and Qian (2018) demonstrate how these conclusions can be constrained. Explicitly highlighted as a limitation within Sanders et al., the researchers based their conclusions about the frequency with which educators use OTRs—e.g., questions or follow-in directives for language—upon a single observation, in a single learning context (i.e., free play), and in a single type of preschool classroom (e.g., inclusive classroom; Sanders et al., 2016). Similarly, Qian based her conclusions upon a single observation in a single learning context (i.e., free play). Their sampling approach was not uncommon as current methods frequently gather only a snapshot of all the interactions to which a child is exposed (Dykstra et al., 2013; Warren et al., 2010). Nevertheless, with only one observation, in one learning context, and within one type of preschool classroom (i.e., Sanders et al., 2016), the researchers' approach introduced potential, unaccounted for measurement error that may go beyond what researchers would reasonably expect when measuring interaction behaviors of educators. For example, with only one occasion (i.e.,

observation session), the researchers were unable to (a) quantify the extent to which their observed frequency represented the true frequency, (b) identify the relative contribution of measurement conditions to variability, and (c) determine the degree to which that particular occasion of data collection was representative of other occasions taken on a different day under the same conditions (Brennan, 2001b; Cronbach et al., 1972). A similar logic can additionally be applied to the measurement conditions of learning context and setting.

Given these methodological shortcomings, readers should interpret the findings of Sanders et al. (2016) and Qian (2018) with greater caution and limit the inferences they draw for two reasons. First, external validity—or the ability to generalize findings to contexts that were not measured—is of interest to many researchers (Cone, 1977; Yoder et al., 2018). Yet, the degree to which researchers can generalize the results to the broader spectrum of settings, activities, and time is impacted when there is a limited understanding of how the sampling approach may or may not have contributed variability to obtained estimates (Cone, 1977; Kane, 1982). Second, with an increased emphasis on rigor and quality of research design (Gersten et al., 2003; Kratochwill et al., 2013; What Works Clearinghouse, 2008), researchers must demonstrate that they have reliably measured the behavior or outcome. With limited evidence that the observed scores represent true scores given a single observation occasion, it is logical to question the reliability of Sanders et al. and Qian estimated frequencies of OTRs (Kane, 1982; Yoder et al., 2018). Given these limitations coupled with knowledge of factors hypothesized to influence language interaction behaviors (Figure 1), there is a clear need for exploration into sampling approaches that *provide more precise, reliable measurement* of the relevant behaviors and *promote generalization* of findings.

Considerations When Measuring Language Used in the Preschool Environment

Obtaining reliable estimates of behavior—or maximizing the precision with which one estimates a true score and minimizing error within the measurement system—is an integral part of observational research (Bottema-Beutel et al., 2014; What Works Clearinghouse, 2008; Yoder et

al., 2018). As a way to operationalize reliability, researchers demonstrate that a measurement is consistent with another measurement of the same behavior by quantifying the strength of the relation between the two and comparing it against a pre-determined, acceptable threshold. In their examination of the frequency of educator use of OTRs and predictors of that use, both Sanders et al. (2016) and Qian (2018) reported high levels of inter-rater agreement (IRA) across 20% of the observations that were double-coded for reliability purposes. This level of reporting is consistent with conventional procedures for indicating the degree of consistency in ratings between observers (Gast & Ledford, 2014). It further aligns with a unidimensional—albeit traditional—approach to addressing the precision of a measurement (Brennan, 2001b; Kane, 1996). 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). Though necessary, it is not sufficient to rely solely on the IRA to demonstrate that an observed frequency represents the true frequency and thus is reliable. Given the dynamic nature of the preschool environment and a need to extend findings beyond the collected data, researchers would be remiss if their conclusions about reliability and exploration of sources of variability ended with measurements of IRA.

When researchers view reliability as a reflection of conditions beyond the conditions in which the behavior was measured and more than IRA, they adopt a multidimensional approach. This perspective allows for more careful attention to the multitude of conditions that may impact inferences made and sheds light on the broader reliability of the sampling approach (Brennan, 2001b; Cone, 1977; Kane, 1996). Within observational research methods and with respect to the preschool environment, the conditions of measurement that are often hypothesized to contribute variability to observed scores and thus impact reliability include, but are not limited to: (a) observers or the individuals coding the behavior, (b) setting/context or the types learning

contexts, types of preschool classrooms, and educators, and (c) occasions or the frequency of the observation (Bottema-Beutel et al., 2014; McWilliam & Ware, 1994; Yoder et al., 2018).

Figure 2 provides a conceptual model of the relationship between (a) these three measurement conditions and (b) the adult-child interactional context and interaction behaviors in the preschool classroom. First, the two binoculars at the top represent the measurement condition of observer. Their placement illustrates that, in general, more than one individual is observing and rating the behaviors of interest across time and settings. Each of these observers contributes error to the observed scores through variability in their rating of the adult and child interaction behaviors. The shaded circle surrounding the interaction context represents the dynamic elements of the classroom ecology (i.e., setting/context) that contribute to measurement error through variability with the types of preschool classrooms, types of learning contexts, and characteristics of the educators. Recognizing that adult and child performance is variable over time, the overlapping circles represent the measurement error that may be related to the occasion on which the observation of the adult-child interaction occurred.

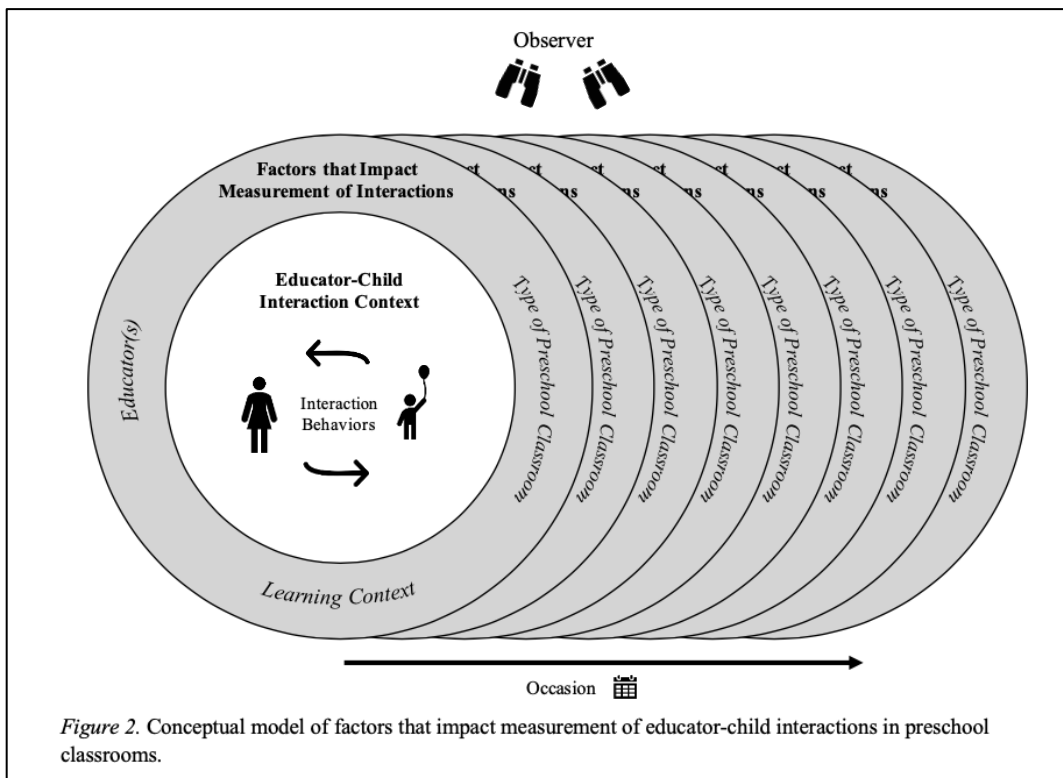


Figure 2. Conceptual model of factors that impact measurement of educator-child interactions in preschool classrooms.

Within Figure 2, the adult and child interaction behaviors are placed in the center to indicate their core role not as a condition of measurement but rather, as the object of measurement. It is these observable interaction behaviors that are of particular interest to researchers (e.g., Qian, 2018; Sanders et al., 2016). The type of interaction behaviors, however, serve as yet another source of variability and error for which researchers must account. Thus, as researchers make inferences about the true frequency, duration, or rate with which interaction behaviors occur in preschool environments, it is necessary to explore the multiple sources of variability—the objects *and* conditions of measurement—that may impact the degree to which the observed measure(s) represents the true measures. Toward that end, Generalizability Theory, and the studies within it, may serve as a potential avenue for empirically examining and disentangling these contributors to error, ultimately informing researchers approaches to sampling.

Generalizability Theory as a Means to Inform Sampling Approaches

In classical test theory (CTT; Bakeman & Quera, 2011; Yoder et al., 2018) and in observational research, an observed score, X , represented by a metric such as frequency, rate, or duration of a behavior is postulated to represent an individual's true score, T , plus error, E , as seen in the following equation (Lord & Novick, 1968):

$$X = T + E \tag{1}$$

The error, however, is undifferentiated, as specific sources of error and the degree to which they contribute to the overall error remains unknown.

As an outgrowth of classical test theory, Cronbach et al. (1963, 1972) developed Generalizability Theory, or G-Theory, to guide researchers in differentiating and quantifying the contribution of various identified sources of error. Utilizing the logic of G-Theory, researchers estimate “the components of observed-score variance contributed by the object of measurements, the [measurement] facets, and their combinations” (Shavelson & Webb, 2006, p. 310). Said another way, researchers can differentiate sources of error by estimating the magnitude with which (a) variability in the persons (i.e., the object of measurement), (b) variability in the

conditions or attributes of measurement (i.e., measurement facets), and (c) variability of the interaction of the persons with the conditions contribute to an observed measure of behavior. Though persons is a source of error, it is not a source of measurement error, as it is their variability that is of interest (Kane, 2002); persons, thus, are considered a differentiated facet and not a measurement one (Yoder et al., 2018).

By adopting such an approach, researchers shift from defining reliability solely as an absolute and unidimensional construct based on the agreement between observers to instead, considering reliability as relative and multidimensional based on how researchers sample the behaviors of interest. Reliability is now dependent upon and relative to the contexts in which it is being interpreted and to which it is being generalized, with an increased ability to understand the extent to which an observed score represents the true score (Brennan, 2001b; Suen & Ary, 1989).

Benefits and Considerations within G-Theory

With G-Theory, researchers are afforded an opportunity to bolster the reliability of their measurements by examining sources of error, a critical pre-requisite for an argument of validity (Kane, 1982, 1996). Researchers can represent the precision of measurement in ways that go beyond traditional measures of reliability, such as IRA (Suen & Ary, 1989; Yoder et al., 2018). By providing empirical information about the contribution of various measurement conditions, such as the time or settings the measurements were taken, researchers can ascertain the extent to which an observed score may provide a stable estimate of the true score over varying conditions (Suen & Ary, 1989).

Within G-Theory, researchers can quantify the extent to which their collected sample of observations (a) represent the larger universe of admissible observations and (b) generalize to the desired universes of generalization (Kane, 2002; Shavelson & Webb, 2006; Suen & Ary, 1989; Yoder et al., 2018). The universe of admissible observations refers to all acceptable observed measures of behavior from the universe of possible measurement conditions. In other words, the researcher's observed score could be interchangeable with another score and validly represent the

construct (Shavelson & Webb, 2006; Yoder et al., 2018). This universe is most classically associated with the generalizability study (described later). The universe of generalization serves as an entity to which observed scores are later generalized and is classically associated with the decision study (described later; Cronbach et al., 1972). It represents a collection of sources of observational measurement error—the measurement facets (Shavelson & Webb, 2006; Yoder et al., 2018).

These measurement facets frequently reflect Cone's (1977) description of the possible universes of generalizations and include (a) scorer/observer, (b) item, (c) time/occasion, (d) setting/context, method, and (e) dimension. As researchers define the universe of generalization, they determine which measurement facets may be relevant to and impact the reliability of the observed scores, frequencies, or durations. Using previous literature, preference, and feasibility in measurement to inform this decision, researchers seek to identify conditions they hypothesize will impact the eventual inferences made within their investigation (Brennan, 2000, 2001b; Li et al., 2015). This method of determination can introduce its own error and result in potential overgeneralization, as researchers may or may not be able to account for, plan for, or even recognize all conditions that impact measurement (Li et al., 2015). In this way, researchers must recognize the existence of hidden or implicit facets, or those facets that contribute to variability but are unaccounted for (i.e., the data have only one level). For example, in a study investigating adult-child interactions solely in free play, the measurement of these interactions may be confounded by measurement in only one context. Though it is impossible to identify and estimate the error attributable to hidden facets, without some recognition of their potential, researchers can overestimate the reliability of their data and in turn, overgeneralize their findings (Brennan, 2000; Kane, 1996).

Measurement facets have specific properties that are critical to a researcher's understanding of the structure of the data, the data analysis design, and the generalizations. First, within each identified facet, the researcher must also identify the levels or the number of

conditions. For example, if researchers identify occasion as a relevant facet, the researcher must indicate the number of times the object of measurement was scored or observed. The extent to which participants are exposed to levels of the facets allows researchers to determine whether the facet is crossed (i.e., all participants are exposed to all levels of a facet) or nested (i.e., participants are exposed to only some levels of the facet; Yoder et al., 2018). For example, a nested facet may result when data from two observers is used, but only one observer codes all data, and another observer codes only a subset of all data. In addition, a facet can have one of two types of an effect that relates to the sampling pattern: (a) a random effect, wherein the levels within the universe of generalization is assumed to be infinitely large, levels are randomly selected, and the researcher generalizes beyond their own data, or (b) a fixed effect, wherein levels are finite or limited to only ones included in the data, with generalizations limited to the data itself (Bruckner et al., 2006; Li et al., 2015). In the end, researcher determination of the facets and their properties have implications for the design, data analysis, and ultimately the inferences made as part of a generalizability study (Brennan, 2001a; Shavelson & Webb, 2006).

Generalizability Study

A generalizability study, or g-study, broadly entails separating and estimating the error that each facet contributes in the universe of admissible observations given a set of gathered data (Brennan, 2001b; Cronbach et al., 1972; Shavelson & Webb, 2006; Suen & Ary, 1989). Through this process, researchers can identify the impact of the measurement conditions on an individual participant's ranking against (a) other persons for relative decisions or (b) specified criteria for absolute decisions (Brennan, 2003; Shavelson & Webb, 2006; Yoder et al., 2018). The variance estimates for each source of variance support researchers in determining if occasion impacts all people within the sample in the same way and to the same degree *or* if there are systematic differences. Harkening back to the discussion of a need for reliability in estimates of behavior, in a g-study, researchers quantify the precision with which they estimated a person's true score and generalize across a set of measurement conditions specifically tested (Shavelson & Webb, 2006).

With the number of facets and the properties of each determined at the outset, a researcher can extend outward from classical testing theory to parse apart an observed score into more than a true score plus error. This observed score, X , represents the (a) grand mean, or average score across all persons and all levels of all facets, μ ; (b) the main effect of the differentiated, person facet based on the person's universe score (analogous to the "true score" in classical testing theory) across all facets, μ_p ; (c) the main effect of the measurement facet(s) based on μ_f for facet, f , across all persons; and (d) the interaction effect of all facets (Brennan, 2001b; Shavelson & Webb, 2006). Breaking this down mathematically and for an example of a one-facet, fully-crossed $p \times o$ (*person x occasion*) design where persons, p , and occasions, o , are randomly selected, the following equation represents the observed score for one person on one occasion, denoted as X_{po} ¹:

$$\begin{aligned}
 X_{po} = & \mu && \text{[grand mean]} && (2) \\
 & + (\mu_p - \mu) && \text{[persons main effect]} \\
 & + (\mu_o - \mu) && \text{[occasions main effect]} \\
 & + (X_{po} - \mu_p - \mu_o + \mu) && \text{[interaction effect]}
 \end{aligned}$$

Each of these facets—person and occasion—and the interaction ($p \times o$) between them contribute to the variability of an observed score. First, there likely exist systematic differences in persons, such as their specific skill, knowledge, behavior, or attitudes. Second, there may be variability in the occasions, such that the materials and activities available during one occasion may be different from the materials available at the next occasion. When the conditions of both the person and the occasion interact, an additional layer of variability, called the residual or interaction effect, results (Brennan, 2001b; Shavelson & Webb, 2006). It is this interaction effect that is of interest to researchers when making relative or absolute decisions, as it quantifies the

¹ Adapted from Shavelson & Webb (2006)

extent to which a person varies by the occasion with reference to others' performance or without reference to others' performance, respectively (Webb et al., 2006).

These three variabilities and their magnitude, quantified as variance components, can be estimated by a *person x occasion* random effects Analysis of Variance (ANOVA). When combined, these variance components—*persons* (σ_p^2), *occasions* (σ_o^2), and the interaction with the residual ($\sigma_{p_o,e}^2$)—represent the variance of all the observed scores, $\sigma_{X_{po}}^2$, and is denoted as:

$$\sigma_{X_{po}}^2 = \sigma_p^2 + \sigma_o^2 + \sigma_{p_o,e}^2 \quad (3)$$

Using this information, the researcher can first calculate the percent of variance accounted for by the differentiated and measurement facet, as well as the interaction of *persons x occasions*. When a design is fully crossed, the interaction of *persons x occasions* is of particular interest to researchers wanting to understand the contribution of the identified measurement condition.

Although the variance components promote an understanding of relative contribution, a clear benefit of G-Theory is the ability to calculate a coefficient of reliability, akin to intra-class correlation coefficient (ICC), a metric commonly used by researchers (Bottema-Beutel et al., 2014; Hallgren, 2012). Cronbach et al. (1972) described these coefficients as an ICC, such that it represents the mean of the ratio of universe-score variance (i.e., *persons*) to the actual observed-score variance for all applications of the design. Mathematically, reliability coefficients—either index of reliability or generalizability coefficient—represent the ratio of the variance observed between persons relative to the total variance (Brennan, 2001b; Shavelson & Webb, 2006). The total variance is the sum of the variance of *persons*, σ_p^2 , and the variance of the interaction effect, $\sigma_{p_o}^2$ for relative decisions, for the index of dependability, $E\rho^2$, which mathematically is:

$$E\rho^2 = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_{p_o}^2} \quad (4)$$

When researchers make absolute decisions, the variance of the measurement facet, σ_o^2 , is also included in the total variance, as the main effect does influence absolute performance. As such, the generalizability coefficient, Φ , can be represented in the following equation:

$$\Phi = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_o^2 + \sigma_{po}^2} \quad (5)$$

Similar to other coefficients, these reliability coefficients carry a range of 0 to 1 and are interpreted as a measure of the stability or precision of an observed score given one random occasion in the universe of admissible observations (Li et al., 2015; Suen & Ary, 1989). The literature suggests some flexibility in a minimum coefficient threshold. Although there has yet to be consensus within the field, one recommendation is that a coefficient of 0.80 or above suggests a reliably observed score (Cardinet et al., 2010). On a pragmatic level, these results will support the researchers in knowing the degree to which the observed behavioral frequency, duration, or score is a reliable estimate of a true score that generalizes to their large universe score taken under all potential measurement conditions (Brennan, 2001b; Cronbach et al., 1972).

Decision Study

In addition to considering the contribution sources of error within a specific investigation, it may be advantageous for researchers to examine how future investigations could be optimized to reduce the contribution of measurement error and increase how reliably the observed scores generalize to the true scores (Brennan, 2001b; Cronbach et al., 1972). As part of a decision study, or d-study, researchers have this benefit. Through an iterative process, researchers systematically manipulate the number of levels of some or all of the facets identified in the g-study and estimate a reliability coefficient for each new set of conditions (Brennan, 2001b; Cronbach et al., 1972). The researcher has flexibility in designing the universe of generalization within the d-study, as they may opt to include all the facets they considered in the g-study or only subset that may be adjusted given practical constraints (Suen & Ary, 1989).

To conduct the d-study, the researcher utilizes the variance components estimated by the g-study to calculate the new reliability coefficients (Li et al., 2015; Shavelson & Webb, 2006). The reliability coefficients continue to represent the proportion of observed score variance given the total, universe-score variance but with one variation (Shavelson & Webb, 2006). The variance component for the measurement facet manipulated is now divided by the number of levels the researcher specifies, or n' . Using the previous example of the one-facet, fully-crossed $p \times o$ (*person x occasion*) design, Equation 6 represents how the index of dependability, $E\rho^2$, is calculated for relative decisions, where the variance of *persons* is σ_p^2 , and the variance of the interaction effect is σ_{po}^2 .

$$E\rho^2 = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_{po}^2}{n'}} \quad (6)$$

Equation 7 represents how the generalizability coefficient, Φ , is calculated for absolute decisions, where the variance of *persons* is σ_p^2 , the variance of occasions is σ_o^2 , and the variance of the interaction effect is σ_{po}^2 .

$$\Phi = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_o^2}{n'} + \frac{\sigma_{po}^2}{n'}} \quad (7)$$

The number of reliability coefficients estimated is a direct reflection of the number of manipulations the researcher conducted. Although the number of manipulations is boundless, researchers frequently conduct d-studies intending to achieve combinations of measurement conditions that result in reliability coefficients at or above the recommended threshold of 0.80 (Bottema-Beutel et al., 2014; Brennan, 2001b; Cardinet et al., 2010; McWilliam & Ware, 1994). The process is further informed by the magnitude of the variance component estimations within the g-study, as it provides information about the relative contribution of measurement facets to error (Li et al., 2015). The degree to which a facet contributes measurement error not only highlights for researchers which facets that may be the most relevant to manipulate, but also

serves as an indicator of the extent to which manipulations of the levels may impact the obtained reliability coefficients (Brennan, 2001b).

As child language researchers seek to make inferences about language learning interactions within the preschool environment, by using this iterative process of d-studies they have a clear advantage in determining the most efficient method that equally supports reliable estimates (Bottema-Beutel et al., 2014; Cronbach et al., 1972). With the explorations of the language learning environment for children with ASD increasing (e.g., Dykstra et al., 2013; Qian, 2018; Sanders et al., 2016), researchers must take advantage of these benefits to ensure sampling approaches can support reliable estimates of interaction behaviors.

Summary

By incorporating G-theory into their own methodology, researchers can move beyond a unidimensional approach to reliability. Instead, researchers can adopt a multidimensional approach that is concerned with how reliably the scores generalize to and represent the true scores under a variety of measurement conditions (Suen & Ary, 1989; Yoder et al., 2018). They can begin to quantify the individual contribution of identified measurement procedures to their observed measures of behavior. Researchers interested in young children in preschool have begun to use this methodology to understand child-level skills such as engagement (McWilliam & Ware, 1994) and social behavior (Chafouleas et al., 2007). To the knowledge of this researcher, however, studies examining the frequency of educators' specific use of OTRs and widespread use of interaction practices to support language development for young children with ASD in the classroom have not yet incorporated the logic and methods associated with G-Theory. Examining the extent to which other measurement conditions, such as occasion, contribute error and impact the precision of measurements made, remains mostly unexplored, yet necessary to informing conclusions that researchers and readers may draw from any given study. The current investigation seeks to provide preliminary information to fill this gap and demonstrate the utility

of conducting g- and d-studies as part of the observational methods used to examine and understand language learning for children with ASD in preschool classrooms.

Chapter 3: Method

Rooted in the logic of G-Theory (Brennan, 2001b; Cronbach et al., 1972), the purpose of this study was to develop and examine a method for characterizing the language learning environment young children with ASD experience within their inclusive preschool classroom. To that end, the first aim was to understand the impact of the current investigation's sampling procedures (i.e., occasion and observer) on the inferences researchers could make about crucial interaction behaviors related to the language experience of these young learners. Following those analyses, the second aim was to empirically evaluate different combinations of sampling procedures that researchers use to optimize the ability to obtain reliable estimates of these critical variables in future investigations.

Participants

For this investigation, the researcher recruited educators and children from 11 early childhood inclusive classrooms who were already participating in a more extensive measurement study funded by the Institute for Education Sciences (IES, Award #R324A170032). Within each classroom, parents of the target child with ASD and all educators consented to participation in procedures that were reviewed and approved by the University Institutional Review Board.

Target Children

The researcher 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, the researcher aligned it with other investigations that have used gathered observational data and applied G-Theory models (e.g., Hill et al., 2012; Mantzicopoulos et al., 2018; Praetorius et al., 2014). To be included in this investigation, the target child needed to have previously qualified for early childhood special education services with a primary educational identification of ASD. Additionally, the target child needed to meet the following inclusion criteria: (a) the parent identified English as their first language, (b) the child used verbal language as a principal means of communication, (c) the child had a goal for expressive communication in their Individualized

Education Plan, and (d) the child had regular access to and participation in free play in their inclusive preschool classroom. The lead educator in each classroom supported the researcher in identifying a participating target child who met inclusion criteria.

Participant Demographics. Table 1 provides the breakdown of demographic information by the target child. All participants were male and between the ages of 3 and 5 years ($M = 4.09$, $SD = 0.70$). The demographic breakdown of the target children’s race was 27.2% Asian, 9.1% Multi-Racial, and 63.6% White. Child 4 in Table 1 additionally identified as of Hispanic or Latino ethnicity.

Table 1
Demographic Information for Target Children

Child	Age	Race	Expressive Language Ability				
			M_{cu} (SD_{cu})	Range of CUs	Occasion with Highest Number of CUs	Percentage of CUs with Single Words	Classification of Language Ability
1	4	WH	35.00 (17.40)	33-61	1	72.88	Single Words
2	4	MR	111.88 (24.34)	73-130	2	23.20	Phrases
3	5	AS	149.25 (52.97)	67-203	2	10.64	Phrases
4	5	WH	122.13 (56.29)	68-208	3	69.39	Single Words
5	5	WH	27.63 (8.18)	17-40	2	9.09	Phrases
6	4	WH	32.38 (15.64)	14-54	4	79.07	Single Words
7	4	WH	14.13 (4.09)	9-22	2	91.67	Single Words
8	4	WH	78.88 (12.57)	61-102	2	13.64	Phrases
9	4	WH	58.75 (23.47)	31-98	2	71.25	Single Words
10	3	WH	61.88 (18.52)	43-82	4	93.75	Single Words
11	3	AS	41.63 (21.34)	15-69	2	59.62	Single Words

Note. AS=Asian; MR=Multi-Racial; WH=White; CU=Communication Unit; M_{cu} = Mean Number of Communication Units per Occasion; SD_{cu} = Standard Deviation of the Communication Units per Occasion.

Expressive Language. To ensure that the target children met the inclusion criteria of using at least single words and to characterize each child’s expressive language ability, the researcher calculated the total number of words per verbalization (TNW; Miller, 1981). To calculate TNW, the researcher identified the observation occasion (described below) with the largest number of coded child verbalizations to ensure ample opportunities for calculations. The researcher then classified each verbalization as either a: (a) single word (i.e., TNW for the

utterance was equal to 1) or (b) phrase/sentence (i.e., TNW for the utterance was greater than or equal to 2). The researcher then divided the total number of verbalizations identified as a single word by the total number of verbalizations for the participant. This fraction was converted to a percentage to indicate how often each child used single words to communicate in the observation session. For target children who used mostly single words (a percentage of 50% or higher), the researcher classified their expressive language as "single words." For children who used single words less than 50% of the time, the researchers classified their expressive language as "phrases or sentences." Based on this method, a total of seven participants used primarily single words ($M_{\text{percent single words}} = 76.8\%$, ranged from 59.62 to 93.75%) and a total of four participants used primarily phrases or sentences ($M_{\text{percent phrases/sentences}} = 85.86\%$, ranged from 76.8 to 90.91%) during at least one of the four sessions observed. Table 1 provides the breakdown of the percentage of single words and classification of language ability by the participant.

Although the calculation of the mean length of utterance (MLU) or mean length of communication unit (MLCU) are the conventional methods for characterizing young children's expressive language ability (Miller, 1981; Paul, 2007), the researcher determined this measure was inappropriate and unreliable for the current study. Collective guidance in the field suggests that researchers use systematic methods to obtain samples, collect samples presumed to be representative, and obtain at least 100 utterances per sample (Casby, 2011; Pavelko & Owens Jr., 2017). In this case, the researcher chose broad characterization, as informed by TNW (Miller, 1981), as a viable alternative to characterize each participant's expressive language level within a single observation session. The researcher made this choice for two reasons: (a) only seven of the 11 participants wore a microphone to allow for accurate collection of their language sample, and (b) additional procedures for collecting samples were not included in the broader study of which this investigation was part. As such, the researcher is discouraging broader judgments and generalizations about the expressive language ability for these target children.

Classroom Demographics. Table 2 provides the classroom demographic information for

each target child, including a description of non-target students and educators. All 11 classroom educators indicated that their instructional approach represented a balance between child-directed and adult-directed activities.

Table 2
Classroom Demographic Information for Target Children

Child	Total Number of Students in Class	Non-Target Students		Educators	
		Total Number with IEP	Disability Areas Identified	Total Number in the Class	Total Number that Interacted with Target Student
1	26	2	ASD; SLI	5	3
2	13	6	ASD; DD	7	3
3	15	3	DD; EBD; OHI; SLI	5	2
4	19	3	ASD; DD; SLI	5	4
5	15	5	DD; SLI	7	5
6	19	2	ASD; DCD; PI; SLI	4	2
7	18	1	ASD	7	4
8	16	5	ASD; DD; SLI	6	5
9	10	4	EBD; DD	4	2
10	19	2	DD; EBD; SLI	6	3
11	20	3	DD	5	5

Note. IEP = Individualized Education Plan; ASD = Autism Spectrum Disorder; DCD = Developmental Cognitive Delay; DD = Developmental Delay; EBD = Emotional Behavioral Disorder; OHI = Other Health Impaired; PI = Physically Impaired; SLI = Speech-Language Impairment. The total number with an IEP is only known for other students that consented and provided enrollment information. The total number with an IEP, as well as the disability area identified, does not include the target student in the total or in the disability category identified.

Teams of Educators

The researcher wanted to examine the classroom environment from the perspective of the target child, an approach that was consistent with previous literature (Dykstra et al., 2013; Irvin et al., 2013; Sanders et al., 2016). To achieve this aim, the researcher recruited all educators in the classroom who interacted with each target child to allow for a more comprehensive picture of the language environment to which the child was exposed. The researcher succeeded in recruiting all educators who interacted with each target child in the classroom ($n = 1-5$), yielding a total of 59 educators. This group of educators included general education teachers ($n = 10$), special education teachers ($n = 11$), general education teaching assistants ($n = 9$), special education teaching assistants ($n = 18$), speech-language pathologists ($n = 7$), occupational therapists ($n = 3$), and other ($n = 1$). Of the 59 educators recruited for participation, however, only 38 educators engaged in interactions with the target children in their classroom during the recorded

observations. The rightmost column in Table 2 provides the breakdown by the participant, indicating the number of educators in the class and the number whose interactions were captured by the researcher across all occasions.

Additional demographic information is provided only for the 38 participants who interacted with the target child for at least some portion of the recorded observations. Table 3 summarizes this information, including role, race, ethnicity, gender, highest degree earned, and level of involvement in the classroom.

Consent and Confidentiality Process

Video-recording, to be described in the data collection procedures section below, was the primary data collection procedure for this investigation. The researcher worked to obtain consent from each child, target and otherwise, in the classroom, as well as all adults (i.e., teaching assistants, special educators, speech-language pathologists, occupational therapists) that were present in the classroom. All parents of target children consented to have their child participate in the study. If parents of non-target children did not agree to their child's participation or if an educator in the classroom did not agree, the researcher ensured that the individual was not recorded. If a child or educator who had not consented to participate appeared on video, that person's images were blurred, and the researcher deleted raw video footage before any coding occurred.

Table 3

Demographic Information for Educators who Interacted with the Target Student

Demographic Category	Frequency
Role in the Classroom	
General Education Teacher	7
General Education Teaching Assistant	4
Special Education Teacher	10
Special Education Teaching Assistant	15
Speech Language Pathologist	2
Occupational Therapist	0
Other	0
Race	
White	38
Ethnicity	
Hispanic or Latino	1
Not Hispanic or Latino	34
Gender	
Female	38
Highest Degree Earned	
High School Diploma	6
A.A.	8
B.A./B.S.	12
M.Ed./M.A./M.S.	11
Above Master's Degree	1
Level of Involvement in Classroom Routines	
Limited, 1-2 specific routines a week	2
Some, 1-2 specific routines each day	1
A lot, 2-3 routines most days	6
All the time, all routines every day	29

Note. Totals may not add up to 38 per demographic category, as some individuals did not respond to all questions.

Facets of the Investigation

Persons as the Differentiated Facet

The participants described above—target children and educators—served as the object of measurement; they were not a measurement facet or source of measurement error, as the researcher was interested in their variability (Kane, 2002; Yoder et al., 2018). Thus, in the current investigation, these participants were considered the differentiated facet and represented *persons* (*p*) in all analyses (Yoder et al., 2018).

Occasion as a Measurement Facet

The researcher identified occasion as the first measurement facet, with four levels. Previous investigations have demonstrated that increasing the frequency of observations (i.e., occasions) is more effective than increasing the duration of a single observation to reliably estimate a behavior (McWilliam & Ware, 1994; Yoder et al., 2018). As such, the researcher chose to film each target child for four sessions of 15 minutes each over three and four weeks. A total of 44 videos (four videos for each of the 11 classrooms) were collected and coded for specific study variables. This facet is identified as *occasion* (*o*) in all analyses.

Observer as a Measurement Facet

The researcher identified *observer* as the second measurement facet, with two levels. The researcher recruited two master's level data collectors with experience in coding to code all videos along with the researcher. The researcher served as the first observer for both the educator proximity code and the child and educator language codes. One observer was trained for the proximity coding and served as the second observer for that variable only; the other observer was trained for the language coding and served as the second observer for all educator and child language variables. Although there was a total of three observers in the study, only two of the observers coded each variable for all videos. This facet is identified as *observer* (*r*) in all analyses.

Type of Preschool Classroom and Learning Context as Hidden Facets

For the study, the researcher recruited inclusive early childhood classrooms from public schools within a metropolitan area in a midwestern state. These classrooms serve children with disabilities alongside their typically developing peers. Self-contained classrooms, or those that exclusively serve children with disabilities, were not included in this study. Given that only one level of the type of preschool classroom—inclusive classrooms—was selected as the learning setting for this study, type of classroom was not included as an identified measurement facet included in this investigation; it represents a hidden facet.

The researcher filmed each target child during one type of child-directed, naturally occurring classroom activity, known as free play. This classroom routine was chosen for several reasons: (a) it was consistently offered across classrooms, (b) it offered a more consistent length for filming purposes across classrooms and occasions, and (c) it increased the likelihood of ample opportunities for each coded language variable (defined in the study variables section below). The researcher asked the lead teacher or an educator familiar with the classroom to identify the start of the free play routine. Given that only one level of classroom routine—free play—was selected as the context for examining the language environment, the learning context was not included as a measurement facet under consideration in this study; it represents a hidden facet.

Data Collection Procedures

To obtain a naturalistic sample of the target child's language experience, the researcher did not provide any specific instructions to the educators before filming. The educators were told to position themselves and 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. In each classroom, the lead educator also worked with the target child to encourage him to wear the same kind of microphone. Despite multiple attempts by the educators to have the target children wear microphones, four participants refused. In these instances, the educators and researcher attempted to strategically place the microphone near the

target children during the play session to increase the likelihood of clearly recording their verbalizations.

Filming for each session began at the onset of an educator language behavior that was (a) directed at the target child (see description in the study variables section) and (b) fell into one of the educator behavior categories. This coded educator language behavior triggered the start of at least 15 minutes of video to ensure there was the potential for at least one coded interaction behavior for use in the generalizability study.

Study Variables to Characterize the Language Learning Environment

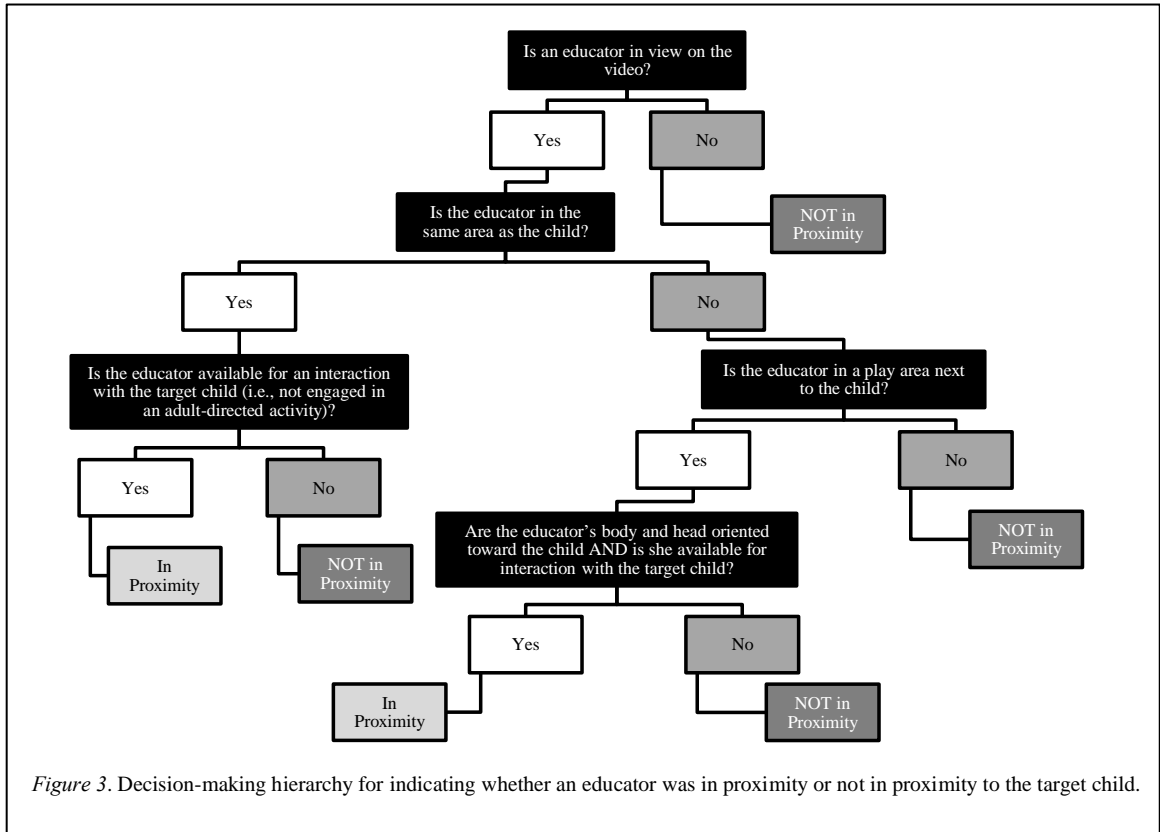
The researcher was interested in three broad categories of variables in the current investigation: (a) educator proximity, (b) educator language, and (c) child language. The following section will detail each of these variables and the specific codes within them.

Educator Proximity

For children with ASD to receive benefit from language learning opportunities provided by educators, the researcher theorized that an adult must be in close proximity to and available for an interaction with them. Thus, the quantification of educator proximity was necessary to better understand potential variability in occasions for the educator and child language variables. Extrapolating from studies that have examined proximity and identified its relation to child engagement (Powell et al., 2008; Sam et al., 2016; Singer et al., 2014; Walker & Berthelsen, 2008), this researcher hypothesized that the presence and availability of an educator to each target child (a) may vary from occasion to occasion and (b) may impact frequency and rates of educator and child behaviors.

The researcher and observer coded for the presence of at least one educator in the play context and their opportunity for interaction with the target child. A duration code represented the proximity of an educator with two mutually exclusive states: (a) educator in proximity and (b) educator out of proximity. An educator was coded as "in proximity" when they met three criteria. These criteria were: they (a) appeared on video, (b) were within the same play area as the child

OR one play area away but oriented toward the target child, and (c) were available for interaction and not engaged in an adult-directed activity that excluded the target child. Figure 3 provides a more comprehensive description of the decision-making hierarchy the researcher and observers used when coding for the educator proximity variable.



The researcher and lead educator identified play areas within the classroom and translated them into classroom play area maps (see Figure 4 for an example). Observers used these maps to distinguish when educators were in the same or different play area as the target child. Consistent with how filming began, an educator was in proximity to the target child at the start of the observation session. A minimum of 5 seconds in the new state needed to elapse before the observers coded a new state.

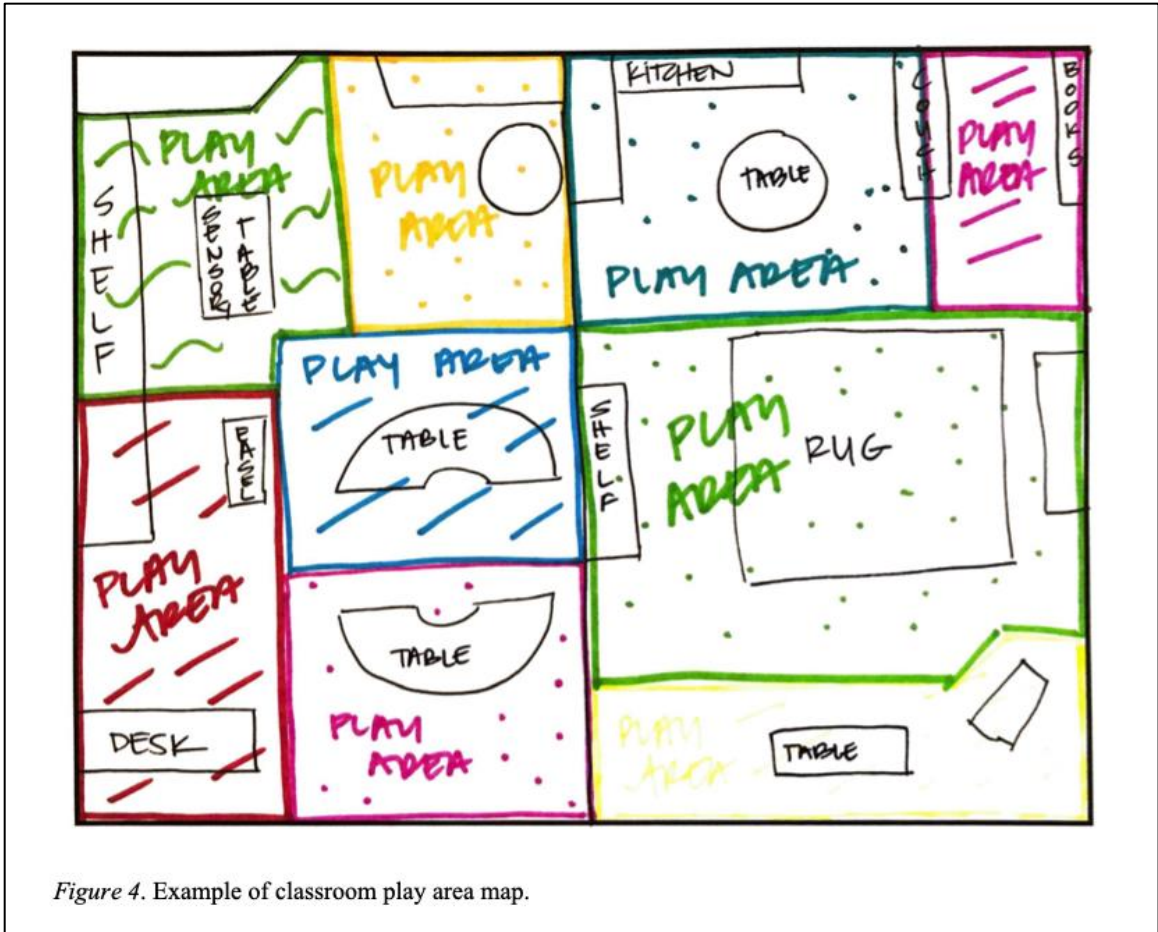
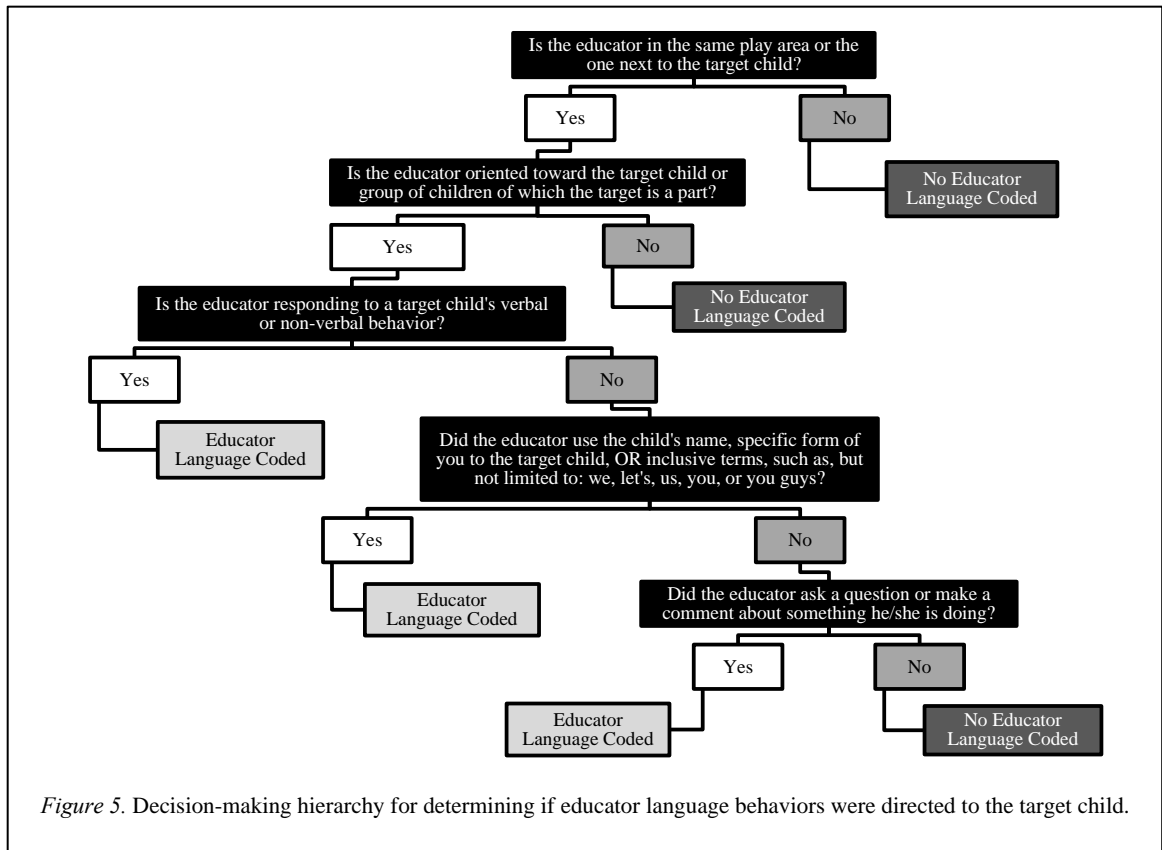


Figure 4. Example of classroom play area map.

Educator Language

The researcher was interested in the type of language used by educators while interacting with the target children in the same play area. To capture and code all language used by educators interacting with each target child, the researcher recruited all educators in each classroom. The researcher coded for specific types of language that were either directed individually to the target child or to a group of children that included the target child. This determination of whether educators directed their language to the target child served as a foundation for using the educator language codes. Any language the observers deemed not to be directed to the target child was not coded. The observers used the decision-making hierarchy in Figure 5 to support observers in this process.



Once the observers determined that the language was directed at the target child or to the group including the target child, the observers used timed-event recording to determine the frequency with which educators were using specific types of language. In contrast to discontinuous interval recording, timed-event recording has been shown to result in increased accuracy in the estimates of educator language behaviors (Cunningham et al., 2019). Cunningham et al. (2019) hypothesized that the discontinuous system might overestimate frequencies, as educator language may span multiple utterances and result in being double or tripled counted. Thus, the researcher chose continuous recording as the most robust method for the purposes of this investigation.

Similar to Sanders et al. (2016), this investigation sought to understand the questions verbal children with ASD experience in their inclusive classroom that explicitly evoke expressive language. Different from their study, which classified a question asked based on the degree of cognitive challenge, this investigation focused on the structure of the expressive language

opportunity. Observers determined the type of expressive language opportunity by the level of support provided within the question that could contribute to a child's success in responding. The following four types of opportunities for expressive language were coded: (a) open-ended questions/statements, (b) yes/no questions, (c) choice questions/statements, and (d) imitation. Because the researcher was also interested in the extent to which educators were broadly providing opportunities for expressive language to the target student with ASD, regardless of type, the researcher aggregated these four codes into one category called *combined opportunities for expressive language*. Finally, with a recognition that educators frequently engaged in talk that was not explicit in evoking expressive language, to further characterize the language to which children with ASD were exposed, the researcher included two additional categories: statements and other. See Table 4 for specific definitions of codes and examples.

Child Language

Child verbalization was the only form of child language behavior that was timed-event coded. The observers coded a child verbalization when the target child used intelligible verbal words, phrases, and sentences in the presence of at least one educator who was in the same play area as the child. When a target child used immediate echolalia, unintelligible jargon, or unconventional noises or sounds (e.g., crying, whining, shrieking), no behavior was coded. See Table 4 for the specific definition and example of this code.

Table 4

Definitions and Examples of Codes within the Educator and Child Language Variables

Code	Definition	Example
Educator Language		
Open-Ended	A question or statement used by the educator that is directed at the target child and can be answered with numerous and unrestricted responses.	What do you want to eat? Where should we go?
Yes/No	A question used by the educator that is directed at the target 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 2 or more explicit options from which the child may choose through the use of 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?
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."
Statement	A statement or comment used by the educator that is directed at the target child that carries meaning in its use, such that it may label or describe. 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!
Child Language		
Child Verbalization	An intelligible spoken word approximation, word, phrase, or sentence that the target child says.	I want ball. My turn.

Note. Open-ended, yes/no, choice, and imitation together represent combined opportunities for expressive language.

Coding Procedures

For all coding of the study variables, the researcher used Behavioral Observation Research Interactive Software (BORIS; Friard & Gamba, 2016) to record the time of and the specific individual code.

Communication Units for Segmenting Language Codes

During the observations, it became apparent that educators and children were frequently stringing multiple phrases and sentences together, which made determining the boundaries between codes challenging. Because the primary aim of the researcher depended on characterizing the language to which a target child was exposed, the researcher decided it was necessary to provide rules and guidelines for how to segment educator and child discourse that could then be behaviorally coded. For this purpose, the researcher adopted Loban's (1966, 1976) 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, given their logical division and accounting for pausing and intonation of child and adult talk (Craig et al., 1998; Eisenberg & Guo, 2013).

One communication unit represents an independent clause and its modifiers, such as a dependent clause (Loban, 1966, 1976). Independent clauses can stand alone because they include both a subject and a predicate (e.g., "He likes blueberries"; "I want a turn"). Dependent clauses, however, do not include a complete thought and therefore, cannot stand alone (e.g., "Because it is raining outside"; "Before you can have a cookie"). For example, the statement, "When you have finished with the puzzle (dependent clause), you can go outside (independent clause)" would count as one communication unit. In contrast, the statement, "First we are going to read the book (independent clause), and then you can play with the trains (independent clause)" would count as two communication units, with the division of the units occurring between the word "book" and the word "and." Following the logic of communication units, the researcher and observer segmented the child and educator language and then categorized each communication unit

following the variable definitions. This segmentation and categorization co-occurred and were not two distinct activities to promote efficiency in the coding process. Although this concurrent nature of coding impacts the ability to distinguish which element (i.e., segmentation or categorization) may have contributed to any reliability estimates that fell below expectations, the reliability was strong (see Training and Reliability section).

Training and Reliability

Before coding, the primary researcher conducted a training session with each observer to review definitions, provide examples and non-examples, and clarify questions. Each observer then independently coded three videos identified as training videos. During this training phase, the researcher and observer made revisions to improve upon the clarity and precision of the definitions for the study variables. Once at least 80% agreement on all individual codes across three consecutive videos was obtained, independent coding proceeded.

To control for and reduce observer drift, the researcher identified roughly 30% ($n = 14$) of the videos for inter-rater reliability (IRA) checks and spaced them out across the entire coding process. The researcher used the kappa values and percentage agreement for each code within a session to resolve any disagreements and support areas of re-training before the researcher and observer continued to code independently. If proximity codes dropped below a kappa value of 0.60 (i.e., substantial agreement; Hallgren, 2012), the researcher and observer planned to discuss disagreements, re-train, and code a new training video until kappa values were above 0.60. No re-training was necessary for proximity coding, as all kappa values obtained during IRA checks were above 0.60. If two or more specific language codes dropped below 80% agreement within a video designated for IRA, the researcher and observer discussed disagreements, re-trained, and then coded a new training video to at least 80% agreement across all individual codes. This re-training was necessary following five videos, which had only two codes drop below 80%.

The researcher completed the agreement checks during training and throughout the coding process using the Multi-Option Observation System for Experimental Studies for all

variables (MOOSES; Tapp et al., 1995). The researcher used the kappa values to determine agreement for educator proximity codes, which she measured as a duration. Across all reliability checks for proximity coding, the mean kappa value was 0.92, with a range of 0.75-0.99, suggesting substantial to near-perfect agreement (Hallgren, 2012).

While kappa values account for chance agreement (Bakeman & Gottman, 1997), the non-occurrence of behavior is undefined in timed-event sampled data, which is a requirement of kappa calculation. The researcher determined that point-by-point percentage agreement ratios for a target code per observation session were a reasonable and supported alternative for determining the reliability of the educator language and child language coding (House et al., 1981). Across all reliability checks for language coding using this alternate method, mean 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, 85% (range, 60% - 100%) for other, and 80% (range, 70% - 88%) for child verbalization. Overall, the mean agreement for each variable was within expectations (Gast & Ledford, 2014).

Data Analysis Approach

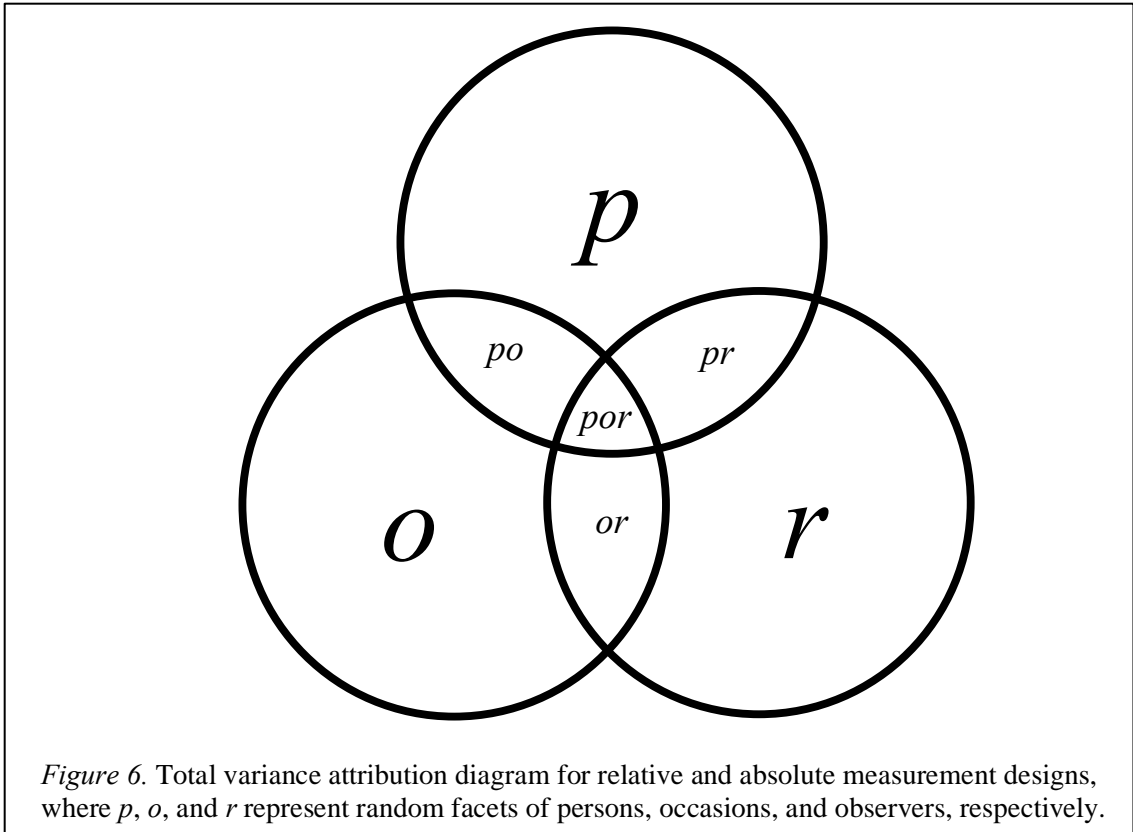
To describe and summarize the data resulting from the classroom observation sessions, the researcher used two software programs, MOOSES (Tapp et al., 1995) and R Studio (R Core Team, 2015). MOOSES was used to calculate the duration an educator was in proximity to the target child in a given session, along with the frequency counts for each of the educator and child language variables in a given session for each observer. Using the data from both observers, the researcher then used R Studio to calculate and visualize descriptive statistics for the entire dataset, such as measures of central tendency and variability. The researcher then analyzed the summarized data to answer the two research questions:

1. When observing interactions of preschoolers with ASD with adult educators during free play to make inferences about the language learning environment, to what extent are the measurement facets of occasion and observer relevant?
2. Given these measurement facets, under what conditions can the sampling methodology be optimized?

Examining the Relevance of Occasion and Observer as Measurement Facets

The researcher conducted a generalizability study, or g-study, to explore the relevance of the two researcher-identified measurement facets (i.e., occasions and observers) for making inferences about the language learning environment of children with ASD. A g-study disentangles and estimates the error that measurement facets contribute to a set of gathered data with respect to the universe of admissible observations (Brennan, 2001b; Cronbach et al., 1972; Suen & Ary, 1989). With a fully-crossed g-study, a researcher can determine if an identified facet impacts all persons within the sample to the same degree or if there are systematic differences that limit the generalizability of observed data to the universe score.

Because nested designs are often less informative to the primary goals of a g-study (Suen & Ary, 1989; Yoder et al., 2018), the researcher used a fully-crossed design. This design included *persons* (i.e., the child or educator participants) and two measurement facets: (a) *occasion* with four levels (i.e., the number of observations per target child) and (b) *observer* with two levels (i.e., the number of observers for each variable). As such, the researcher modeled a *person x occasion x observer* random effects Analysis of Variance (ANOVA) for each of the study variables to support generalization from the observed durations or frequencies of educator and child behaviors to the larger universe of admissible observations. The variance partition diagram in Figure 6 displays how the researcher broke down the relative and absolute total variance, such that the effects of *persons* (*p*), *occasions* (*o*), *observers/raters* (*r*), and the interactions between them (*pr*, *po*, *or*, and *por*) can be visually illustrated.



The researcher conducted a series of g-studies in EduG software (Swiss Society for Research in Education Working Group, 2012). With this software, the researcher calculated the variance components for *persons*, *occasions*, *observers*, *persons x occasions* interaction, *persons x observers* interaction, *observers x occasions* interaction, and *persons x observers x occasions* interaction for each coded variable. The software also provided the percent variance accounted for by each source and the accompanying reliability coefficient for the individual g-study. Because of the fully-crossed design, the variance components that were most relevant in the interpretation were *persons*, *persons x occasions* interaction, and *persons x observers* interaction (McWilliam & Ware, 1994); it is the interaction effect that quantifies the extent to which persons vary by occasion and observer. While the main effects of occasion and observer become essentially meaningless when persons are not taken into account, the results of these sources of variance were included to provide a comprehensive picture of the error differentiation.

A reliability coefficient, which represents the extent to which the data gathered can be generalized to the universe score, was calculated for each variable. The researcher used the following equation (Brennan, 2001b; Cronbach et al., 1972) to calculate this coefficient:

$$\text{reliability coefficient} = \frac{\text{true score variance}}{\text{true score variance} + \text{error variance}} \quad (8)$$

To support both absolute and relative decisions, the researcher calculated both types of reliability coefficients for each variable (Cardinet et al., 2010). To calculate the absolute (criterion- or domain-referenced) reliability coefficient or index of dependability, Φ , where the variance for *persons* is σ_p^2 , *occasions* is σ_o^2 , *observers* is σ_r^2 , the *persons x occasions* interaction is σ_{po}^2 , the *persons x observers* interaction is σ_{pr}^2 , and the *persons x occasions x observers* interaction with the residual is $\sigma_{pro,e}^2$, the researcher used the following equation:

$$\Phi = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_r^2 + \sigma_o^2 + \sigma_{po}^2 + \sigma_{pr}^2 + \sigma_{pro,e}^2} \quad (9)$$

A similar equation was used to calculate the relative reliability coefficient or *generalizability* coefficient, $E\rho^2$ for each variable. The denominator, however, does not include the variance of the facets themselves (i.e., σ_r^2 or σ_o^2).

$$E\rho^2 = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_{po}^2 + \sigma_{pr}^2 + \sigma_{pro,e}^2} \quad (10)$$

One recommendation suggests that a coefficient of 0.80 or above represents a reliably observed score (Cardinet et al., 2010). The researcher adopted this guideline for interpreting the results of the g-study.

Optimizing the Sampling Approach

Following the g-study, the researcher conducted a decision study, d-study, for each variable to determine the optimal conditions of the measurement facets that led to stability in observed scores. Within each d-study, the researcher systematically and statistically manipulated the levels of the measurement facets for those below criterion levels. This manipulation allowed

the researcher to forecast the number of occasions and the number of observers that, when combined, reduced the magnitude of the error components and in turn, optimized the reliability coefficient (Shavelson & Webb, 2006; Suen & Ary, 1989). The researcher conducted this iterative process of data analysis until reliability coefficients were at or above 0.80 (Cardinet et al., 2010). Using EduG software (Swiss Society for Research in Education Working Group, 2012), the researcher calculated the absolute (Φ) and relative ($E\rho^2$) reliability coefficients for each manipulation for each variable using the same equations (Equations 9 and 10) as in the g-studies.

Chapter 4: Results

The purpose of this investigation was to develop and examine a method for characterizing the language learning environment for young children with ASD in inclusive preschool classrooms. The researcher observed 11 children across four occasions with key language interaction behaviors coded by two observers. Aligning the methodology with the tenets of Generalizability Theory (Brennan, 2001b; Cronbach et al., 1972), the researcher was interested in first examining the extent to which this sampling approach impacted inferences made about the language learning environment of children with ASD. Guided by these results, the researcher then empirically and systematically evaluated how the sampling approach could be optimized to increase the reliability of the measurements. The researcher explicitly addressed the following research questions in this investigation:

3. When observing interactions of preschoolers with ASD with adult educators during free play to make inferences about the language learning environment, to what extent are the measurement facets of occasion and observer relevant?
4. Given these measurement facets, under what conditions can the sampling methodology be optimized?

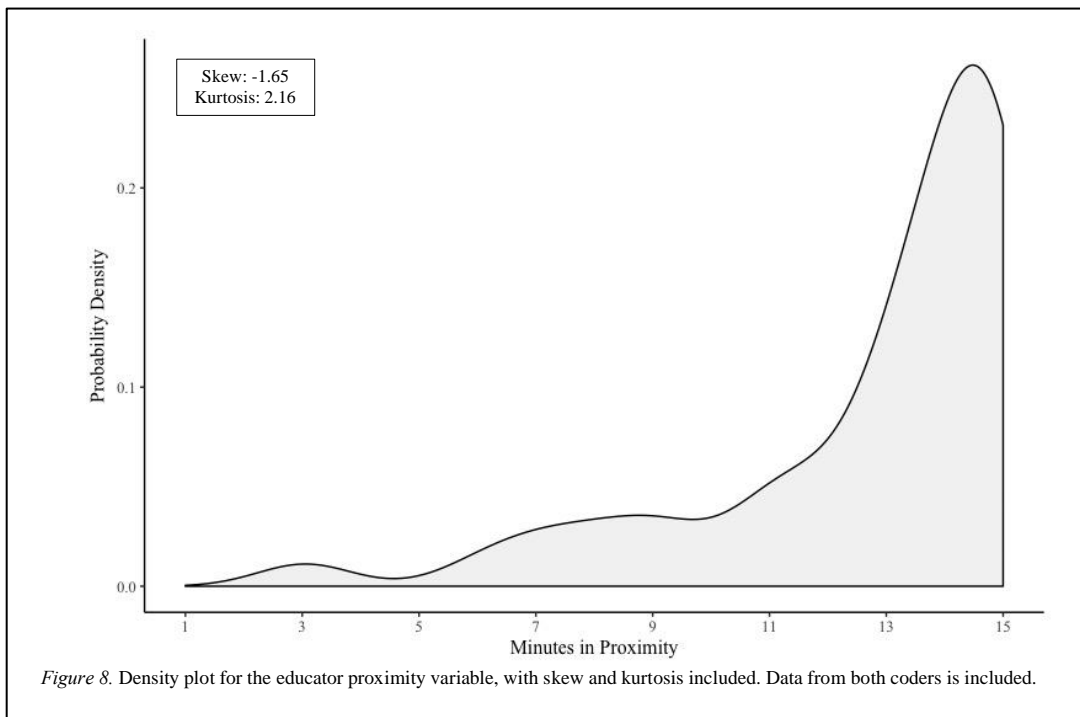
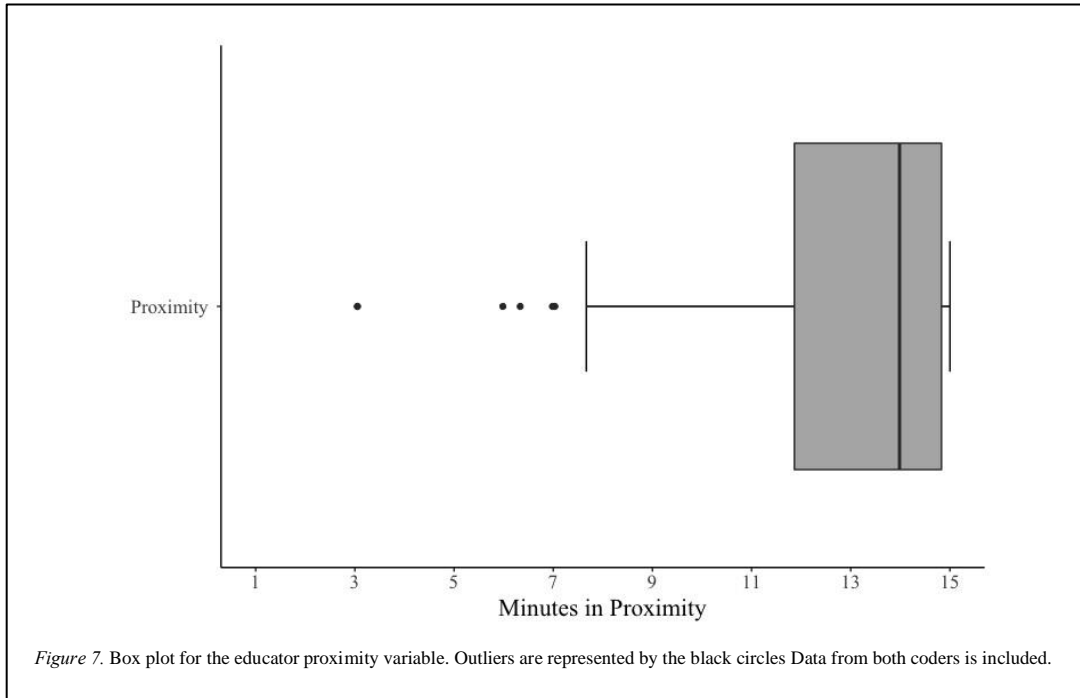
Descriptive Statistics for Language Learning Environment Variables

Before answering the primary research questions, the researcher explored descriptive information about the language environment to understand central tendency, variability, and extent of normality in the distribution of each variable measured. The descriptive statistics provided represent the full dataset, meaning data reported includes that from both observers for each participant across the four recorded occasions.

Proximity

The researcher measured the duration of time during which at least one educator was in proximity to the target child. On average, at least one educator was in proximity to a target child roughly 12.81 minutes ($SD = 2.80$ minutes; range = 3.05-15 minutes) out of 15 minutes. The

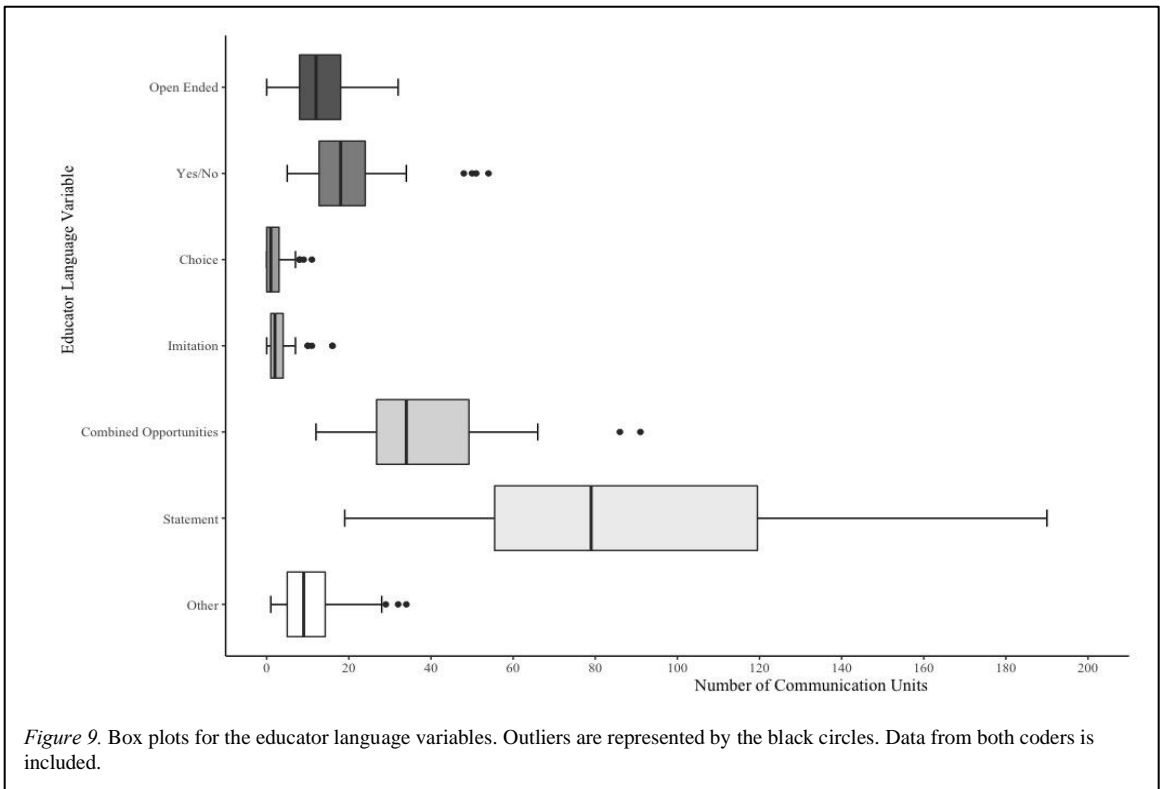
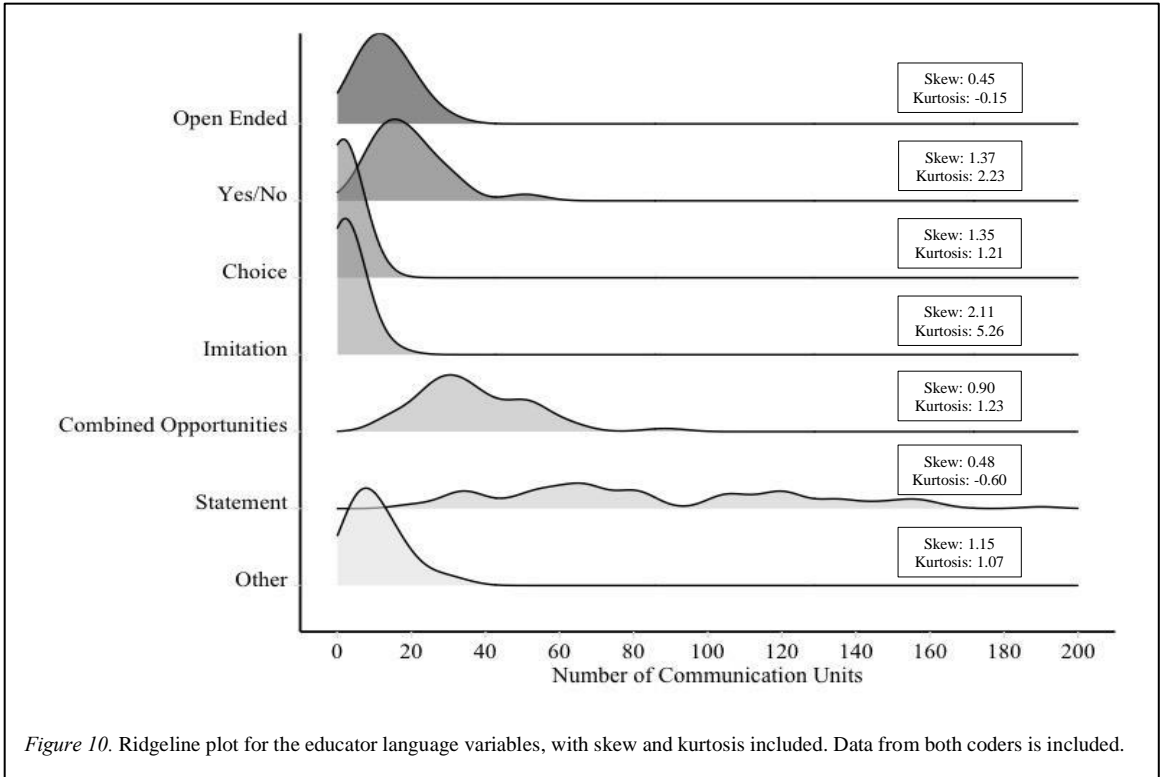
boxplot in Figure 7 shows the range for minutes in proximity, as well as the median (13.98), quartiles, and outliers. Figure 8 illustrates the density plot and indicates the skew and kurtosis values; it suggested that the proximity variable had a non-normal distribution, such that it is skewed to the left, and there is extremeness in the data.



Educator Language

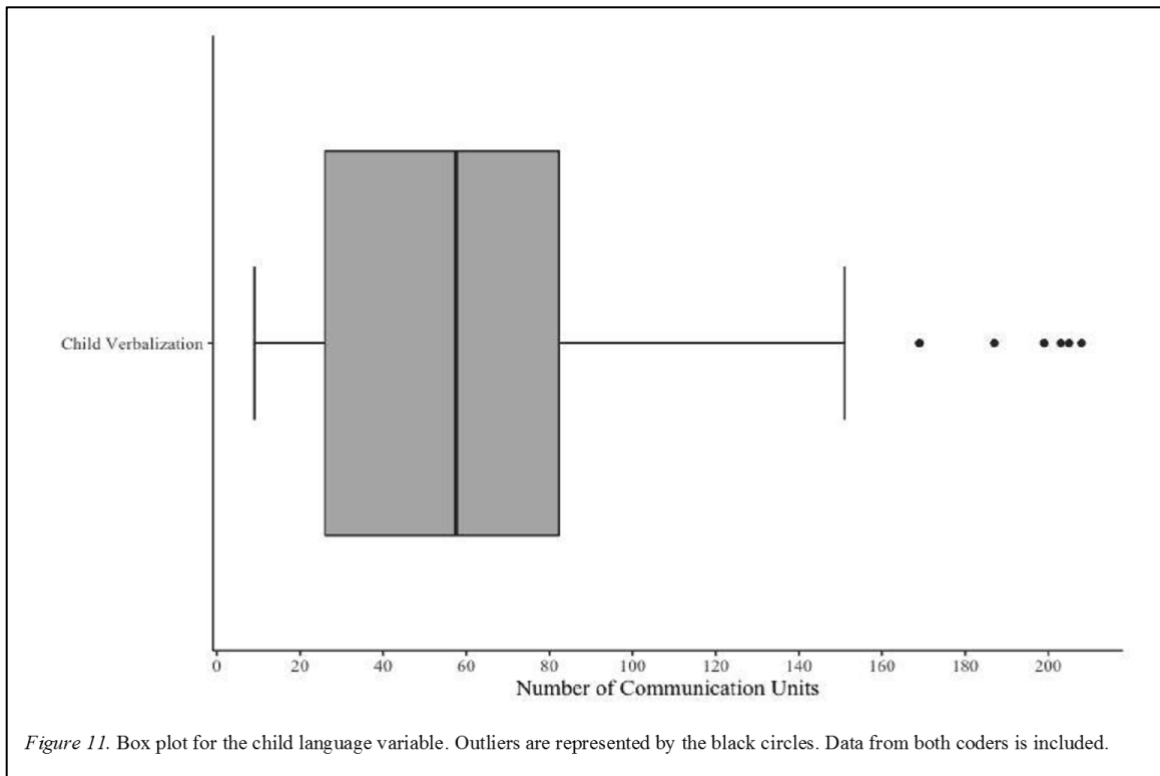
The researcher measured the frequency with which educators directed specific types of communication units to children with ASD. Averaged across all sessions for all target children in a 15-minute free play session, educators used a mean (in communication units) of 12.83 open-ended questions/statements ($SD = 6.95$, ranged from 0 to 32), 19.45 yes/no questions ($SD = 9.93$, ranged from 0 to 54), 1.99 choice questions ($SD = 2.51$, ranged from 0 to 11), 2.77 imitation prompts ($SD = 3.18$, ranged from 0 to 16), 87.31 statements ($SD = 41.95$ ranged from 19 to 201), and 10.5 other language ($SD = 7.30$, ranged from 1 to 34). Four of these educator language variables (open-ended, yes/no, choice questions, and imitation statements) were aggregated to represent the educator language variable: combined opportunities for expressive language. The mean number of combined opportunities for expressive language was 37.05 ($SD = 15.20$, range of 12 to 91).

Figure 9 provides the boxplots for each variable. Figure 10 provides the ridgeline plots for each educator language variable, consisting of individual density plots with skew and kurtosis indicated. Of the seven variables, open-ended question/statement was the only variable that closely approximated a normal distribution, with skew and kurtosis in acceptable ranges (i.e., between -1 and 1 and close to 0, respectively), the mean and median close together, and no outliers or extremeness in the data. The variables of yes/no question, choice, imitation, combined opportunities, and other were all skewed to the right. They had a considerable concentration of data around the mean, while also showing extremes. Though the statement variable was not highly skewed, it was highly variable and is a borderline bi-modal distribution.



Child Language

The researcher measured the frequency of communication units each target child with ASD used within 15 minutes (i.e., the child language variable). On average, children used 66.68 communication units ($SD = 49.72$). The median number of communication units was 57.50, with a range of 9 to 208 communication units, suggesting significant variability. Figure 11 provides the boxplot, and Figure 12 provides a density plot for this child language variable. Similar to several of the educator language variables, the child language variable was skewed to the right and had a considerable concentration of data around the mean, while also having extremes.



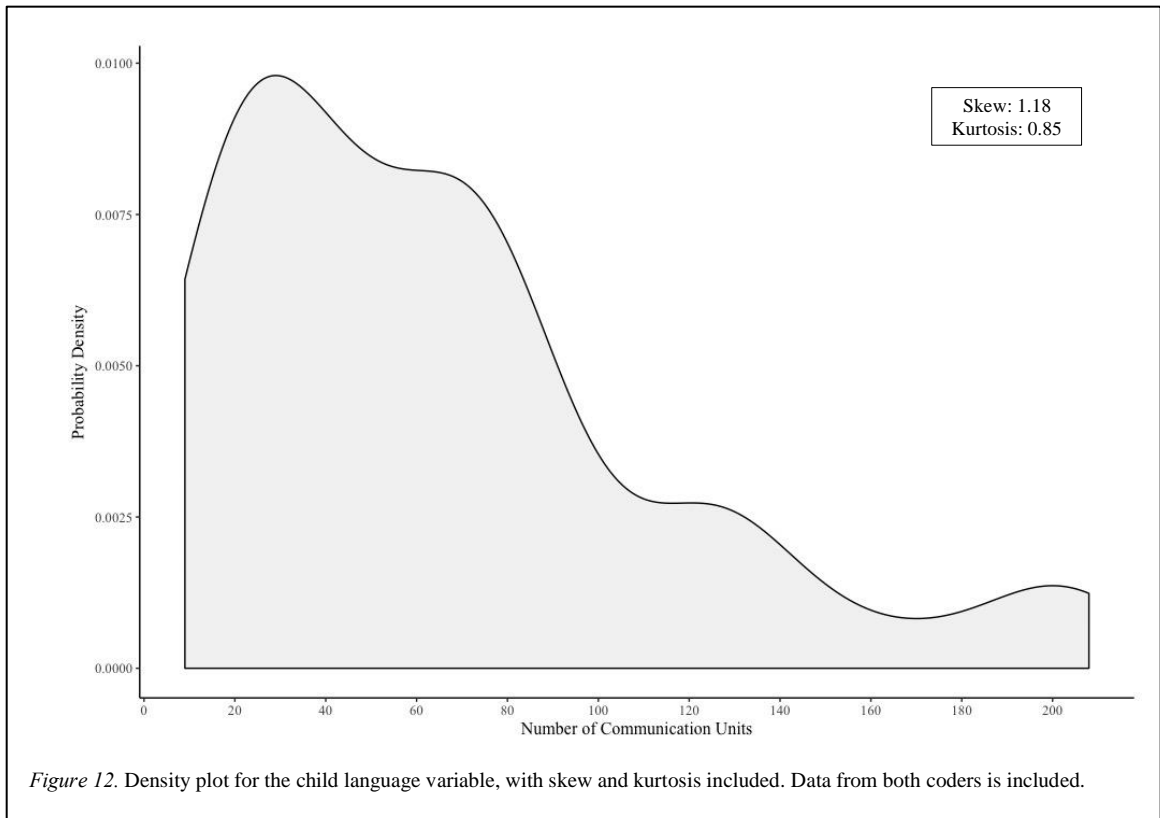


Figure 12. Density plot for the child language variable, with skew and kurtosis included. Data from both coders is included.

Summary of the Language Learning Environment Variables

Overall, the majority of the variables measured in this investigation were non-normal, such that there were extremes and considerable spread in the data, skewed distributions, and disparities between the mean and medians. One educator language variable, open-ended questions/statements, appears to most closely approximate a symmetric distribution. Given the relatively small sample size, this non-normality and extremeness was expected.

To conduct the g- and d- studies, the researcher estimated the variance components with ANOVA and further recognized that the normality assumption, typical to this statistical approach, was violated. The literature on G-Theory, however, that suggests the assumption of a normal distribution of scores, frequencies, or durations of the variables is not required and, in most applications of the approach, is violated (Brennan, 2001b; Briesch et al., 2014; Hendrickson & Yin, 2010). Additionally, using Monte Carlo simulations, researchers demonstrated that skewed distributions have minimal bearing on the estimation of the reliability coefficients (Shumate et al.,

2007). As such, the researcher details the results of the g- and d-studies with recognition that a violation of the normality assumption has a minimal impact.

Examining the Relevance of Occasion and Observer as Measurement Facets

To determine the measurement facets that were relevant to a sampling approach that seeks to characterize the preschool language learning environments of children with ASD, the researcher conducted a g-study for each of the variables in this investigation. The results are organized into the broad category variables: proximity, educator language, and child language. The estimated variance, percent of the total variance, and standard error of the variance for each source, as well as absolute (Φ) and relative ($E\rho^2$) reliability coefficients, was calculated for each variable.

Proximity

Table 5 presents the estimates, percentages, and standard errors of the variance for each source that contributed to the total variability of the proximity variable. The primary source of variance was the interaction of *persons x occasions*, such that *persons x occasions* accounted for 61% of the total variability in the duration of time educators were in proximity to the target child. In contrast, *persons* accounted for 38.6% of the total variance, and *persons x observers* accounted for less than 1% of the total variance. Consistent with these findings, the absolute and relative reliability coefficients were both estimated as 0.72, below the threshold of 0.80.

Table 5

Variance Estimations, Percentages, and Standard Errors by Source for the Proximity Variable

	Variance Estimation	Percent of Total Variance	Standard Error
<i>Persons</i>	3.18	38.6	1.84
<i>Observers</i>	<0.01	0.0	<0.01
<i>Occasions</i>	0.00*	0.0	0.24
<i>Persons x Observers</i>	0.00*	0.0	<0.01
<i>Persons x Occasions</i>	5.03	61.0	1.26
<i>Observers x Occasions</i>	<0.01	0.0	<0.01
<i>Persons x Observers x Occasions</i>	0.03	0.4	0.01
Total Variance	8.24		

Note. Percentages may not add up to 100 due to rounding. *Following guidance from Brennan, (2001) and Cronbach, Gleser, Nanda, and Rajaratnam, (1972), variance estimates that were negative were rounded to 0.00.

As a reminder, the quantification of educator proximity was necessary to better understand potential variability in occasions for the educator and child language variables. This researcher hypothesized that the presence and availability of an educator to each target child may be related to rates of educator and child behaviors. To quantify the relation between the proximity variable and each educator and child language code, the researcher used Spearman's rank order correlation coefficient (r_s). This coefficient was chosen because the data were not normally distributed and appear to have a monotonic relationship. Table 6 provides the coefficient and corresponding p -values for each estimated relation. With a p -value of 0.05, the results indicate that a significant relation existed between five out of the seven educator language variables (e.g., yes/no, choice, combined opportunities, statement, and other) and proximity. The relation was not significant for the child language variable.

Table 6

Spearman's Rho Correlations between the Proximity Variable and Each Language Variable

Variables	r_s	p
Educator Language		
Open-ended	0.17	0.12
Yes/No	0.28	<0.01**
Choice	0.26	0.02**
Imitation	0.20	0.07
Combined Opportunities*	0.36	<0.01**
Statement	0.33	<0.01**
Other	0.36	<0.01**
Child Language		
Child Verbalization	0.18	0.09

*Combined opportunities represent the combined frequency counts of open-ended, yes/no, choice, and imitation.

**Significant at $p < 0.05$.

Educator Language

Table 7 presents the variance estimates for each of the sources that contributed to the total variability of each of the educator language variables. Table 8 presents the variance percentages for each of the sources, and Table 9 presents the standard errors for each source of variance. *Persons x occasions* accounted for the largest percentage of total variability across all educator language variables (range of 53.6% to 77.5%), while *persons* accounted for 13.6% to 45.6% of the total variance. For all variables, *persons x observers* accounted for less than 1% of the total variance. The estimated reliability coefficients were between 0.39 to 0.77, which, similar to that of proximity variable, is below the threshold of presumed reliability.

Table 7

Variance Estimations by Source for the Educator Language Variables

	Specific Types of Educator Language						
	Open-Ended	Yes/No	Choice	Imitation	Combined Opportunities	Statement	Other
<i>Persons</i>	13.78	34.33	0.92	3.90	111.81	391.00	18.58
<i>Observers</i>	0.00*	0.00*	0.00*	0.00*	0.10	0.00*	0.00*
<i>Occasions</i>	0.00*	0.00*	0.00*	0.48	0.00*	0.00*	1.47
<i>Persons x Observers</i>	0.10	0.05	<0.01	0.05	0.28	5.13	0.41
<i>Persons x Occasions</i>	37.94	67.79	5.74	5.75	131.42	412.85	30.96
<i>Observers x Occasions</i>	0.07	0.22	0.00*	0.02	0.19	0.00*	0.50
<i>Person x Observer x Occasion</i>	0.37	1.33	0.08	0.42	1.36	13.19	3.79
Total Variance	52.16	103.72	6.74	10.62	245.16	822.17	55.71

*Following guidance from Brennan, (2001) and Cronbach, Gleser, Nanda, and Rajaratnam, (1972), variance estimates that were negative were rounded to 0.00.

Table 8

Variance Percentages by Source and Reliability Coefficients for the Educator Language Variables

	Specific Types of Educator Language						
	Open-Ended	Yes/No	Choice	Imitation	Combined Opportunities	Statement	Other
<i>Persons</i>	26.4	33.1	13.6	36.7	45.6	21.5	33.4
<i>Observers</i>	0.0	0.0	0.0	0.0	0.0	0	0.0
<i>Occasions</i>	0.0	0.0	0.0	4.5	0.0	0	2.6
<i>Persons x Observers</i>	0.2	0.1	0.0	0.4	0.1	0.3	0.7
<i>Persons x Occasions</i>	72.6	65.4	85.1	54.2	53.6	77.5	55.6
<i>Observers x Occasions</i>	0.1	0.2	0.0	0.2	0.1	0.0	0.9
<i>Persons x Observers x Occasions</i>	0.7	1.3	1.2	3.9	0.6	0.7	6.8
F	0.59	0.67	0.39	0.70	0.77	0.52	0.68
Er^2	0.59	0.67	0.39	0.72	0.77	0.52	0.69

Note. Percentages may not add up to 100 due to rounding.

Table 9

Variance Standard Errors by Source for the Educator Language Variables

	Specific Types of Educator Language						
	Open-Ended	Yes/No	Choice	Imitation	Combined Opportunities	Statement	Other
<i>Persons</i>	9.83	21.44	1.03	2.24	59.76	318.17	11.22
<i>Observers</i>	0.03	0.06	<0.00	0.01	0.17	0.33	0.15
<i>Occasions</i>	0.90	3.13	0.17	0.67	5.62	76.88	3.08
<i>Persons x Observers</i>	0.08	0.18	0.01	0.07	0.27	3.54	0.60
<i>Persons x Occasions</i>	9.53	17.11	1.45	1.49	33.02	354.87	8.23
<i>Observers x Occasions</i>	0.06	0.22	<0.00	0.04	0.20	0.44	0.54
<i>Person x Observer x Occasion</i>	0.09	0.33	0.02	0.10	0.34	3.30	0.95
Absolute Standard Error	3.10	4.14	1.20	1.28	5.77	18.91	2.97
Relative Standard Error	3.10	4.14	1.20	1.23	5.76	18.91	2.90

The percent variance accounted for by the interaction of *persons x occasions* for the aggregated category of combined opportunities for expressive language was the lowest among all educator language variables. This interaction, however, still accounted for more of the variability than *persons* 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).

Child Language

Table 10 presents the estimates, percentages, and standard errors of the variance for each of the sources that contributed to the total variability of the child language variable. *Persons*, or the target child themselves, accounted for the most variance (64.9%) in the number of child verbalizations, compared to the interaction of *persons x observers* (less than 1%) and the interaction of *persons x occasions* (32%). The estimation for the *persons x observers* interaction was consistent with results from the proximity and educator language variables. However, unlike

for the proximity and educator language variables, *persons* accounted for more variance than *persons x occasions* for the child language variable. Averaging across observers and occasions, participants varied systematically in the frequency of their verbalizations. The relative and absolute reliability coefficients were 0.89 and 0.88, respectively. These reliability coefficient estimations were above the 0.80 threshold and suggested that the frequencies obtained may be presumed to be reliable.

Table 10
Variance Estimations, Percentages, and Standard Errors by Source for the Child Language Variable

	Variance Estimation	Percent of Total Variance	Standard Error
<i>Persons</i>	1710.80	64.9	788.29
<i>Observers</i>	5.69	0.2	5.42
<i>Occasions</i>	51.26	1.9	83.75
<i>Persons x Observers</i>	5.18	0.2	4.13
<i>Persons x Occasions</i>	843.24	32.0	213.10
<i>Observers x Occasions</i>	0.16	0.0	1.22
<i>Persons x Observers x Occasions</i>	18.21	0.7	4.55
Total Variance	2634.54		

Note. Percentages may not add up to 100 due to rounding.

Summary and Implications for Exploring Optimization

Based on the results of the g-studies, the researcher made two conclusions regarding the conditions of the occasion and observer measurement facets as they pertain to sampling methodology. First, finding that the *persons x observers* interaction contributed little to no error across all measured variables indicates that any differences in the observed durations or

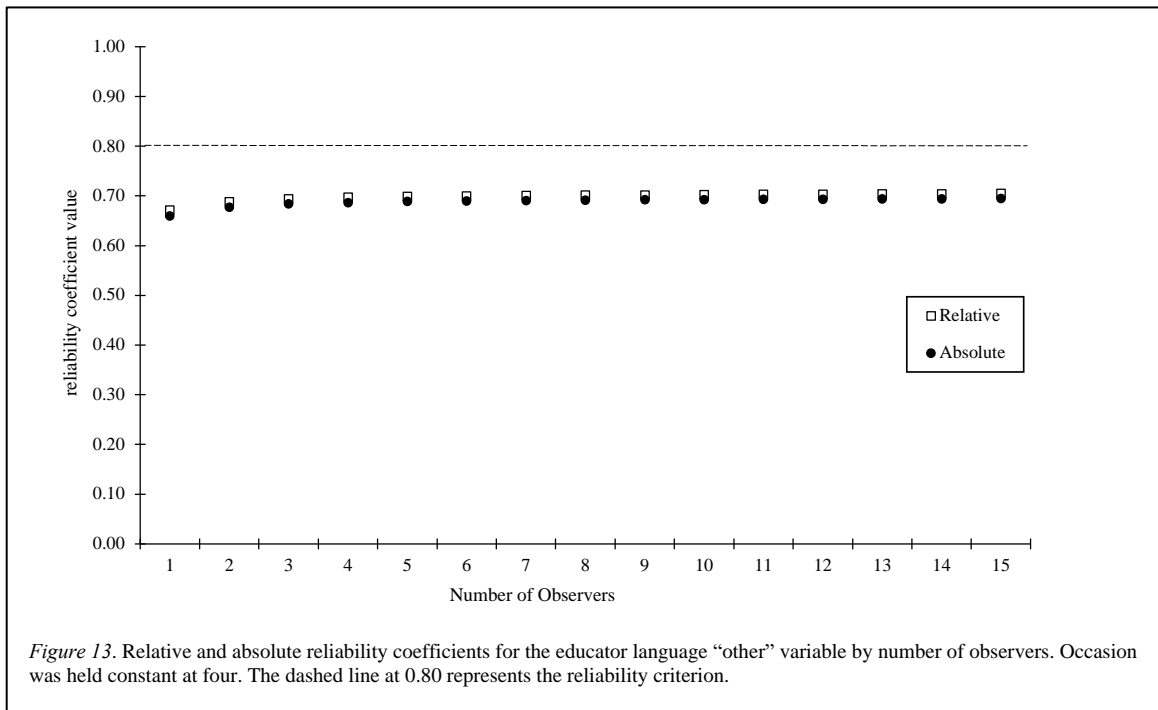
frequencies between the two coders had little to no bearing on the total variability observed. Thus, observer was likely not a relevant measurement facet in this sampling approach. Second, the interaction of *persons x occasions* contributed a substantial amount to the measurement error for all but one of the variables (i.e., child language); this contribution was more than the contribution by *persons*. This result suggests that occasion was a relevant measurement facet when characterizing the language-learning environments for children with ASD. Practically, these results further suggest that observing across four occasions was likely insufficient for obtaining reliable estimates of educator proximity and educator language variables; fewer occasions may be sufficient for a reliable estimate of child language. In conclusion, systematic manipulation of the levels of occasion as a part of d-study was the next logical and empirically supported step to optimizing the sampling approach.

Optimizing the Sampling Approach

The researcher conducted a series of planned d-studies on the same variables used in the g-studies. The purpose of these d-studies was to understand the configuration of occasion and observers that would result in a sampling approach that not only provided precise, stable estimates of the specified behavior but was also resource and cost-effective (Gleser et al., 1965; Shavelson & Webb, 2006). In planning the d-study, the researcher defined occasion as the only universe of generalization and manipulated the levels within it to determine the minimum number of occasions needed to obtain the reliability of 0.80 (Cardinet et al., 2010).

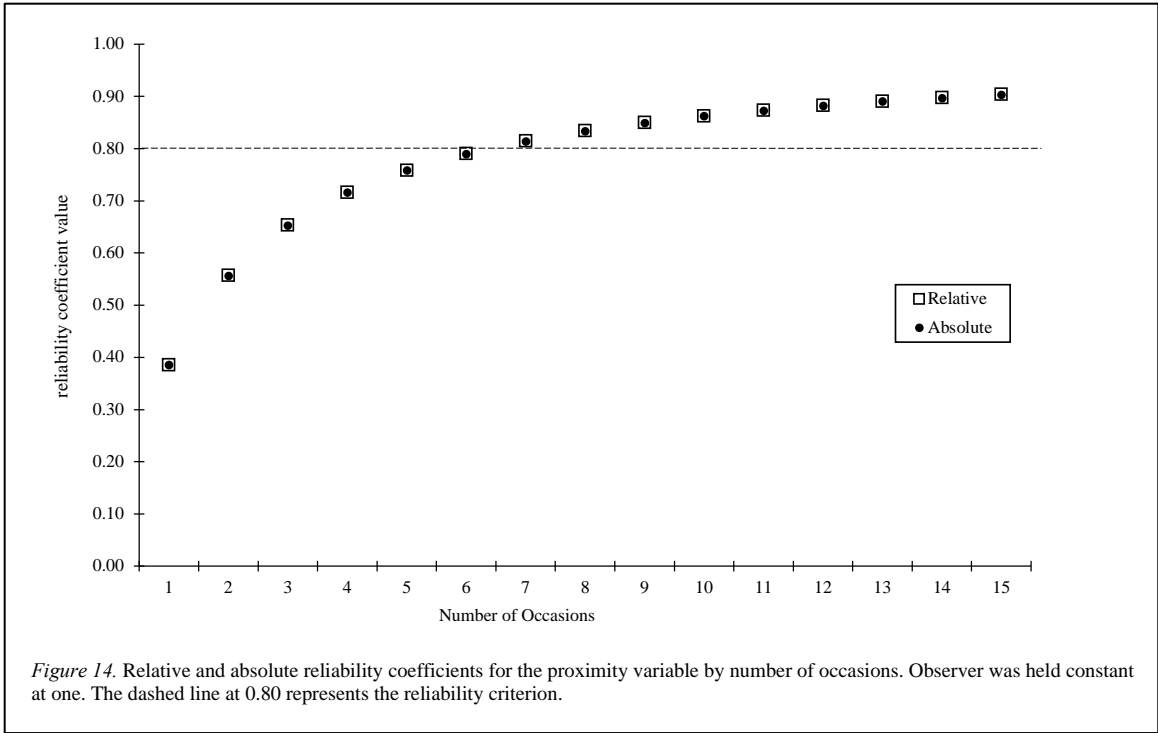
Because observer was not a relevant measurement facet, it was held constant at one throughout all d-studies. The researcher made this decision given the fact that there would be minimal impact on the reliability coefficients when adding additional observers, as well as the minimum number of observers required within an investigation to code any data and the cost-effectiveness of having only one observer. As an example of the impact on the reliability coefficients, the researcher selected the variable in which *persons x observers* contributed the highest amount of variance (i.e., the “other” code in the educator language variable at 0.7%). The

number of observers was then manipulated from one to 15, with the number of occasions held constant at four. Figure 13 displays the results graphically. The incremental increase in the reliability coefficients demonstrates that indeed, including more than one observer had little effect on the stability of the observed frequencies. In maintaining the current approach of four occasions per participant, the researcher determined that more than 10,000 coders were needed to achieve reliability coefficients of 0.80 for all variables in this investigation.



Proximity

Figure 14 presents the relative and absolute reliability coefficients for the proximity variable, as the number of occasions increased from one to 15. After seven observation sessions, the reliability coefficients were above the threshold at 0.80 for both absolute (0.81) and relative (0.81). In other words, a researcher needed to conduct at least seven sessions before stable estimates of the duration of time educators were in proximity to the target child with ASD could be obtained.



Educator Language

Figures 15 and 16 present the relative and absolute reliability coefficients, respectively, for each of the educator language variables, as the number of occasions increased from one to 15. Using the threshold of reliability of 0.80, and depending on the variable, it would take five to more than 15 occasions to produce stable estimates. More specifically, the researcher would need to observe five occasions for combined opportunities for expressive language, 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.

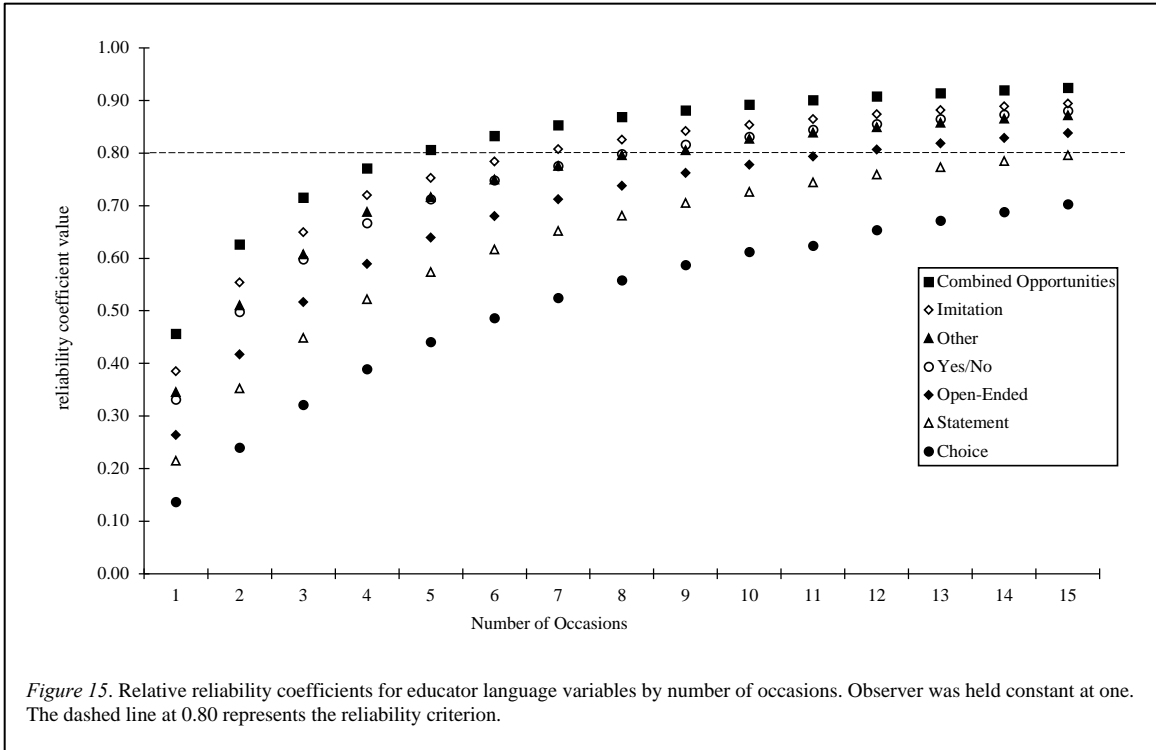


Figure 15. Relative reliability coefficients for educator language variables by number of occasions. Observer was held constant at one. The dashed line at 0.80 represents the reliability criterion.

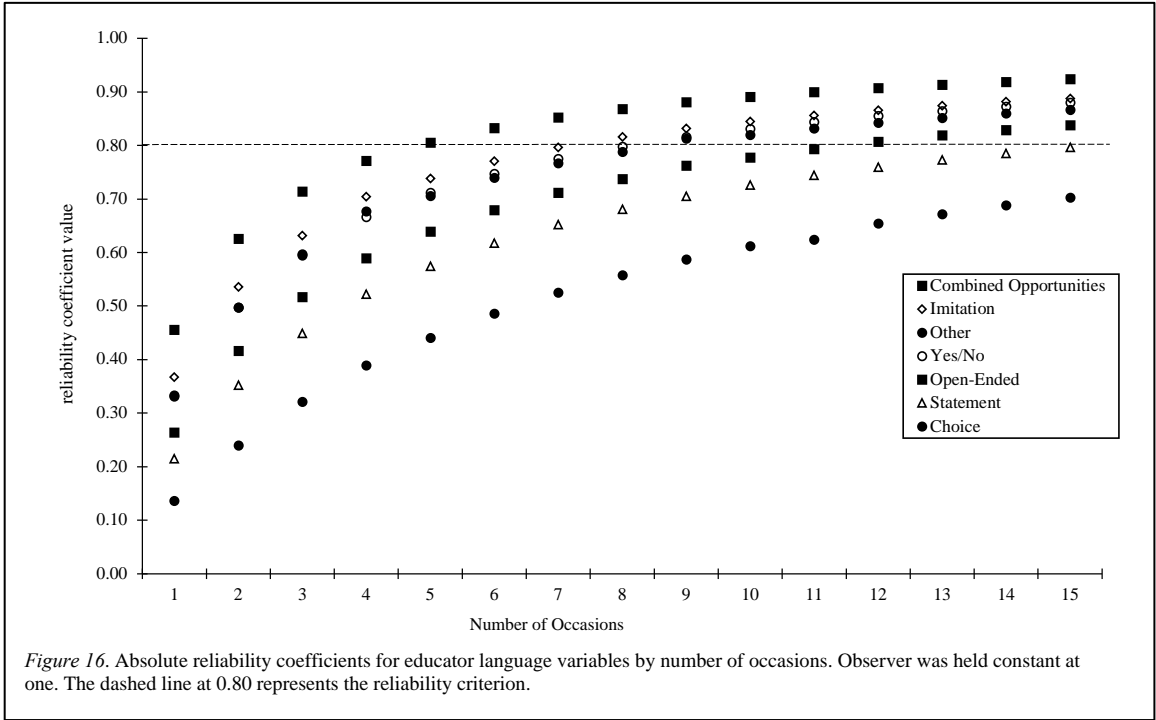
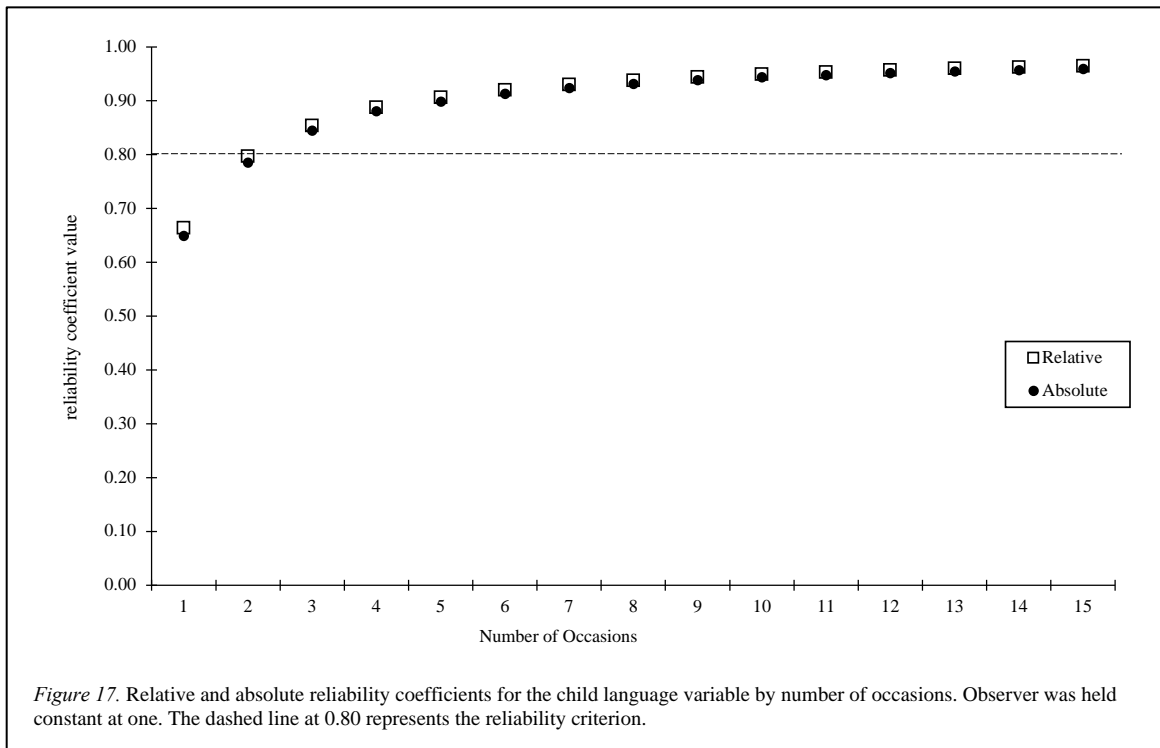


Figure 16. Absolute reliability coefficients for educator language variables by number of occasions. Observer was held constant at one. The dashed line at 0.80 represents the reliability criterion.

Child Language

Figure 17 presents the relative and absolute reliability coefficients for the child language variable, as the number of occasions increased from one occasion to 15 occasions. In meeting the threshold for reliability, the number of occasions necessary was fewer than used in the current approach. The researcher would need to observe on only three occasions for reliable estimates of the number of communication units children with ASD use during free play in their inclusive preschool classroom. At this number of occasions, the relative and absolute reliability coefficients were 0.85 and 0.84, respectively.



Chapter 5: Discussion

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 impact the inferences they make. By adopting the logic of Generalizability Theory, researchers can differentiate sources of error and quantify the extent to which observed measures precisely represent the true measure (Brennan, 2001b; Cronbach et al., 1972; Yoder et al., 2018). The current investigation aligned with this theory and had two primary aims. First, the researcher sought to understand the extent to which two aspects of the methodological approach—occasion and observer—contribute variability that may be relevant to inferences made about the language environment. Second and guided by these results, the researcher evaluated how the methodological approach could be enhanced or altered in future studies to balance statistical rigor and practical effort.

Examining the Relevance of Occasion and Observer as Measurement Facets

In the context of gathering information on language-learning environments for young children with ASD, person refers to the adult or child interaction behaviors, occasion refers to the observation session, and observer refers to the researcher who identified and recorded the educator- and child-level variables during specific occasions. 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, 2001b; McWilliam & Ware, 1994; Yoder et al., 2018). To explore the extent to which these two facets contributed to error that was relevant to the current investigation, the researcher conducted a series of g-studies. The researcher quantified the variance that each of these sources contributed to the true estimate of an interaction behavior and then evaluated the impact of each source on inferences about the language-learning environment.

Evaluating Variance Attributed to Occasion

Because the researcher used a fully-crossed design, the primary source of variance relevant to understanding the contribution of occasion was the interaction of *persons x occasions*. Across all variables that the researcher measured, *persons x occasions* provided a range of 32% to 85% of the variance. To further understand the impact of occasion precision of the observed behavior in estimating the true score, the researcher compared it to *persons*, or the source of variance that represents the object of measurement. When compared to *persons*, *persons x occasions* contributed more variance for all educator variables, resulting in reliability coefficients below the recommended threshold of 0.80 (Cardinet et al., 2010). Therefore, the combined variance indicates that there was limited precision and stability in (a) estimating the true duration educators were in proximity to children and (b) the true frequencies 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 (Brennan, 2011; Epstein, 1979; Mantzicopoulos et al., 2018; Praetorius et al., 2014; Yoder et al., 2018). Moreover, two elements identified by other researchers (Dykstra et al., 2013; Irvin et al., 2013; Sanders et al., 2016) may offer an additional explanation of this result: (a) the proximity variable and (b) the recognized measurement approach.

Proximity Variable. For children with ASD to access language learning opportunities from educators, the researcher presumed that an adult must, to some degree, be physically present and available for interaction with them. For example, it seems unlikely that a child with ASD playing with trains on one side of the classroom will receive benefit from an educator oriented away from them and providing OTRs to a group of children playing at the art table on the opposite side of the classroom. This logic is consistent with other investigations that have

explored relations between adult participation and proximity and child engagement (e.g., Sam et al., 2016; Singer et al., 2014). To the knowledge of this researcher, however, this was the first investigation to empirically explore how this variable related to rates of educator interaction behaviors for language learning.

Given the assumption that the proximity of an adult matters for language learning, the researcher operationalized this variable by first defining play areas within the classroom (see Figure 4). These play areas were then incorporated into a hierarchy to determine when an adult was in proximity to the target child (see Figure 3). The researcher hypothesized that (a) the proximity of an adult would vary from occasion to occasion and (b) proximity would be related to the observed frequencies of adult and child interaction behaviors. There is evidence from the analyses, as well as the g- and d-study to support these hypotheses. First, similar to that of the educator language variables, *persons x occasions* accounted for a considerable portion of the variance (i.e., 61%) in the g-study for the proximity variable. Pragmatically, this result suggests that across the four observation occasions, the percentage of time in which at least one educator was available for interaction with the target child was highly variable, which ranged from 20% of the session (i.e., 3.05 minutes) to 100% of the session (i.e., 15 minutes). This finding aligns with other empirical investigations which suggest that preschool children may be involved with educators from as little as 24% to over 90% of their free play session (File, 1994; Kontos, 1999; Powell et al., 2008). Furthermore, the results of the d-study indicated that at least seven observations would be required before the researcher could obtain reliable estimates of educator proximity, once again suggesting considerable variability across occasions.

This variability in proximity across occasions, however, is particularly important given the relation between an educator's proximity and their frequency of language. In fact, for all but two of the codes (i.e., open-ended and imitation), the researcher identified a significant, monotonically increasing relation (see Table 6). This finding is expected given that the coding of educator language essentially required an educator to be in proximity to the target child. That is,

observers only coded language when the adult was, at a minimum, oriented toward and directing their language to the target child (see Figure 5). On the other hand, if an adult was in a different play area and oriented away from the target child, the educator would be coded as not in proximity, *and* the educator's language would not be coded.

This approach to only coding directed language by an adult near the focal child is consistent with other literature examining language-learning environments (e.g., Qian, 2018; Sanders et al., 2016). The proximity and availability of an adult, however, have not been captured and quantified in relation to language use in previous investigations. Quantifying the proximity variable, however, is essential in future investigations for understanding variance in our measurements, making inferences, and informing our interventions. First, quantifying the educator proximity may provide a potential explanation to the contribution of occasion to the total variance, as was seen in the current investigation. Second, the measurement of this variable may also provide researchers with a plausible explanation for observed magnitudes of language interaction behaviors. For example, if findings suggest that educators are infrequently providing OTRs, the proximity variable could provide an important first step in sorting out whether the low frequency was related to educators (a) not being in proximity to the focal child and available for interaction or (b) being in proximity but not engaging in this interaction behaviors. Depending on the potential association observed, the findings may indicate different avenues of intervention, such as zoning to ensure adequate coverage of adults in play areas (McWilliam & Casey, 2008) or teaching educators to use this key interaction behavior (Simonsen et al., 2010). Pragmatically, given the role proximity may play in our analysis and interpretation, this researcher and others may consider revising the conceptual model put forth in Figure 2, by including proximity as a factor that may impact the measurement of educator-child interactions.

Recognized Measurement Approach. To characterize the language-learning environment from the perspective of the target child, the researcher coded and combined the directed language *any* educator used into a single aggregate frequency for the observation

occasion. Simply put, the data was not parsed apart by the individual educator. In this way, an expected source of variability—different educators within the classroom—was masked by the aggregation of language use across all adults into a single score.

Although the researcher aligned this method with current approaches in the field (e.g., Dykstra et al., 2013; Irvin et al., 2013; Sanders et al., 2016), it may further explain why *occasion* contributes more variability to the observed scores. With the repeated measurement of an individual, researchers can reasonably expect some degree of normal variation in scores, frequencies, or durations of a particular behavior (Gast & Ledford, 2014); across persons, though, the same expectation may or may not hold. In this study, the number of adults that interacted with the target child ranged from one to five educators across the four observation sessions. This range in and of itself introduces variability. Characteristics of these individuals, such as their role and educational level, present additional sources of variation that confound the measurement. Previous literature provides evidence that both the role and educational level of educators impact the type and frequency of talk used in preschool classrooms (Sawyer et al., 2018; M. W. Smith & Dickinson, 1994). For example, in their audio-recording of Head Start and other preschool classroom teachers, M. W. Smith and Dickinson (1994) identified positive associations between the amount of cognitively challenging talk to focal children and the educator's educational level. Likewise, Sawyer et al. (2018) found that when compared to lead teachers, assistant teachers talked to target children less frequently. Given this literature, future investigations that seek to characterize the language learning environment should consider parsing out the language used by the individual educator. By doing so, researchers can eliminate the potential masking effect of an aggregated score and limit a known source of variability at the outset. Furthermore, this measurement approach may facilitate the research in providing more precise accounts of how different adults contribute to language learning opportunities. In turn, this level of detail in the research may offer practical guidance to improve the efficacy of the overall language learning environment experienced by young children with ASD.

Impact on Narrow versus Broadly Defined Variables. *Persons x occasions* contributed a considerable amount of variance to scores for the individual educator language variables of open-ended (72.6%), yes/no (65.4%), choice (85.1%), and imitation (54.2%). It is noteworthy, however, that when the frequencies of these four variables were aggregated represent a single variable for combined opportunities for expressive language, the contribution of *persons x occasions* was less (53.6%). This shift in the contribution of *persons x occasions* to the variance when moving from highly specified variables to a broad aggregated variable is consistent with other literature. For example, Bottema-Beutel et al. (2014) had similar findings in their g-study, when examining variables as detailed as peer initiations and as comprehensive as peer interaction.

Practically, the discussion of the contribution of occasion has implications for researchers when measuring highly specified behaviors from a molecular perspective or broader dynamic groupings from a molar perspective (Baum, 2002, 2011). First, when researchers adopt a fine-grained, molecular view in their examination of language learning variables, those behaviors are often defined and analyzed in ways that link them together in close temporal proximity; for example, an adult provides an OTR, and the child responds. Given the molecular view on how specific interaction behaviors may promote language learning for children with ASD, findings from this study demonstrate that it may be necessary for researchers to increase their sampling across multiple occasions. In a related way, the molecular view of measurement may also apply when researchers are examining quality factors (Rowe & Snow, 2019), such as the degree of cognitive challenge to questions asked (e.g., Massey, 2003; Sanders et al., 2016) or examining the extent to which comments were responsive or contingent upon child behavior (Haebig et al., 2013; Qian, 2018). On the other hand, as researchers consider variables that represent broader and more dynamic groupings of behaviors from a molar view, those variables are often defined in ways that allow for time to be extended to contribute more variability to the observed behavior (e.g., any educator's use of an opportunity to respond relative to any type of response from the child). Given a molar view on how groupings of interaction behaviors may promote language

learning for children with ASD, findings from this study demonstrate that fewer observations may be needed to make inferences about the language-learning environment. This type of broad, molar view of language learning may include quantifying only the number of words adults use with children regardless of quality features such as linguistic or conceptual complexity (e.g., adult word count; Burgess et al., 2013; Dykstra et al., 2013) or the number of comments without reference to whether or not it followed the child's attentional focus (e.g., the coding of statements in the current investigation). In the end, regardless of the level of specificity chosen, it would behoove researchers to carefully consider whether and how variables themselves may contribute to the variance to *occasion* and adjust their sampling procedures accordingly.

Evaluating Variance Attributed to Observer with Traditional IRA and G-Studies

The primary source of variance that is relevant to understanding the contribution of *observer* is the interaction of *persons x observers*. In the current investigation, *persons x observers* contributed less than 1% of the variance across all variables. In studies in which observers are trained to a specified criterion of IRA (generally 80% or higher; Gast & Ledford, 2014), such as in this study, the g-study provides additional confirmation that observers are not contributing unique variance to the observed scores, as one would hope. Many studies rely solely on traditional measures of IRA when evaluating the language learning environment (e.g., Dickinson et al., 2011; Kontos et al., 2002; Qian, 2018; Sanders et al., 2016). Though the researchers 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). For example, Kontos et al. (2002) concluded that teacher involvement, over classroom quality, was a significant predictor of the complexity of child interactions within the preschool classroom. Their inference about teacher involvement was based upon a single, 2-hour observation and strong IRA. Though there *may* be some truth to their inferences about the language environment, researchers should be careful to consider evidence of good IRA during a limited sampling of the classroom as

sufficient support for drawing inferences about the language environment beyond that limited sampling. Other evidence is needed to establish the generalizability of the data beyond that sampling context, which includes examining other sources of variance (e.g., occasion, type of instructional routine).

Traditional measures of IRA are still necessary. In the present study, the researcher calculated the variance components and reliability coefficients within each g-study using the total frequency counts from each observer without attention to whether or not individual instances of behavior occurred at the same time. Though this approach provided the needed variability to estimate other sources of variability within the g-study, it is still essential to understand if observers are observing the same behaviors at the same time. Further, how IRA is calculated also has implications for the inferences made. For example, for one participant, the first observer coded a total of 81 statements, and the second observer coded 76 statements, a difference of only five. Point-by-point agreement within 3-seconds, however, was roughly 71%, with 26 disagreements recorded, considerably more than one would presume based merely on the comparison of the total frequencies between the two observers. In other words, 29% of the time within the observation occasion, the observers disagreed about whether or not a statement occurred, despite an overall count that appeared similar. House, House, and Campbell (1981, p. 54) summarize this scenario perfectly, “A single summary statistic...sacrifices a great deal of information, may obscure areas of greater or lesser uncertainty, and reduces the reader’s [or in this case, the researcher’s] evaluation to an oversimplified ‘good enough/not good enough’ discrimination.” The critical point is that though a particular score may provide information, it is the responsibility of the researcher to understand the pieces of information they have available to them as they draw conclusions and make claims.

Though the discussion by House et al. (1981) was primarily referencing general summary statistics, such as mean, median, or kappa, their logic can be extrapolated to the discussion of variance estimates and reliability coefficients within G-Theory. If a researcher is interested in

determining whether or not different observers are providing the same scores for the behavior of interest, traditional measures of IRA are both necessary and sufficient. On the other hand, if a researcher is interested in determining if an observer's score provides information that may be generalized beyond the observed sample, IRA continues to be necessary but is no longer sufficient. As such, it may be appropriate for researchers to more routinely examine how different measurement facets relate to the generalizability of the observed scores by conducting a g-study in addition to traditional methods. This dual approach would allow researchers to understand the sources of measurement error relative to the types of behaviors observed, scores generated, and sampling procedures within a context for which researchers hope to provide more generalized knowledge.

Hidden Facets

Hidden facets represent those measurement conditions wherein only one level is sampled (Brennan, 2001b). The researcher identified two hidden facets at the outset of this investigation—routine and setting—such that sampling occurred only in free play and inclusive classrooms. If readers or researchers intend to make generalizations across routines (e.g., large group, small group, music, motor room, snack) or across settings (e.g., self-contained, childcare, Head Start classrooms), a word of caution is necessary: the reliability coefficients are only as good as the conditions that have been explicitly tested. In other words, estimated coefficients do not extend to measurement facets not included in the current investigation (e.g., frequencies of behavior in free play, large group, and small group; Brennan, 2001b, 2001a). The reliability coefficients obtained are likely to provide an overestimate of reliability, particularly when researchers are planning to generalize over hidden facets, such as instructional routine or type of classroom (Cronbach et al., 1972; Webb et al., 2006). In that vein, it may be most appropriate to consider reliability coefficients as an upper limit to reliability and interpret them as such, mainly when more generalizations to contexts beyond those included in the sample are made (Brennan, 2001a; Cronbach, 1951).

Implications for Researchers

For researchers to make inferences about the true frequency, duration, or rate with which interaction behaviors occur, it is necessary to explore the multiple sources of variability within the preschool environment. Pragmatically, if the researcher had not first undertaken a g-study, the inferences made about the language learning environment for children with ASD may have been based on imprecise, unreliable data. The standard measures of reliability in the study, such as kappa and percent agreement, suggested moderate to high measurement precision (Gast & Ledford, 2014; Hallgren, 2012). In actuality, given the amount of measurement error and instability of the data as illustrated by the results of the g-study, these inferences—albeit unknowingly—would have lacked validity (Brennan, 2001b). As researchers design studies aimed at describing the preschool language learning environment and plan for generalization of their results to contexts beyond what they sampled, incorporating occasion and considering additional conditions of measurement that may impact inferences is essential. Though it may not be complete, the conceptual model that illustrated the relationship between measurement conditions and the adult-child interactional context (see Figure 2 in Chapter 2) provides a valuable starting point for addressing sources of variance when measuring language in the preschool environment.

Optimizing the Sampling Approach

Understanding how to optimize the sampling approach may offer important new avenues for researchers seeking ways to understand and optimize the efficacy of the language-learning environment for preschoolers with ASD. For example, if inferences were made that educators only used 1.99 choice questions in 15 minutes based on four brief observations, those inferences would be based on an unreliable measurement. The d-study results indicated that more than 15 observations were required to achieve a stable representation that may model the effects. In the present study, for each variable, the magnitude of the estimated variance components within the g-study informed the facets manipulated within the accompanying d-study (Li, Shavelson, Yin, &

Wiley, 2015). The researcher conducted a series of d-studies to statistically and systematically understand the combination of conditions wherein measurement error was minimized. These d-studies supported the researcher in understanding under what conditions the generalizability of the observed scores to the true scores may be maximized (Brennan, 2001b; Cronbach et al., 1972).

With observer deemed as an irrelevant facet for all variables, the researcher fixed it at one, only manipulating occasion. In general, the number of occasions needed to achieve stable estimates of educator proximity and directed language ranged from five (e.g., combined opportunities for expressive language) to more than 15 sessions (e.g., statements and choice questions). As researchers balance statistical requirements with practical cost and effort, these results highlight the importance of researchers understanding how their measurement and sampling of certain behaviors may restrict the inference they can make when faced with real-world data collection constraints (Bottema-Beutel et al., 2017). When considering the language-learning opportunities in preschool classrooms, future investigations may consider adjusting the type of opportunity for expressive language from the four prompting types in the current investigation (i.e., open-ended, yes/no, choice, imitation) to either (a) a more encompassing variable (e.g., combined opportunities for expressive language) or (b) being more selective of specific types. For example, a researcher could measure the use of open-ended versus closed-ended questions, which allows them to categorize how responses can be constrained and prompted broadly, an approach taken by Pellegrino and Scopesi (1990) and De Rivera et al. (2005). This approach further supports some specification of the OTR type beyond a single category of combined opportunities for expressive language.

Although educator variables for language and proximity required additional occasions for precise estimates, child language was stable at four occasions with absolute and relative reliability coefficients at 0.88 and 0.89, respectively. Fewer observations of child language (i.e., only three) were needed to meet technical requirements for reliability, a finding that, compared to the other

variables in this investigation, is generally positive for the practical cost. When taken at face value alone, this finding lends itself to the conclusion that the results obtained in the current investigation were reliable. It may additionally provide support for the validity of any inferences made about the frequency with which a young child with ASD verbalized during free play.

The finding of a stable, reliable estimate of child language aligns with expectations given that we presume some degree of consistency in observed frequencies of behavior within a person (Gast & Ledford, 2014). When considering the implications of this finding for real world preschool classrooms, however, this is a bit surprising and perhaps a bit disheartening, particularly in light of the degree of variability in the educator proximity and language variables. Harkening back to the conceptual model presented in Figure 1 (see Chapter 2) and previous literature, there is evidence that both adult and child behaviors influence one another in bi-directional and transitional ways (Dunst et al., 1990; Sameroff, 2009). The stability of measurement of the language behavior of the preschoolers with ASD in this study, in contrast to the lack of stability for the adult language behaviors, raises additional questions for researchers to explore.

Observed child language behavior could have been independent or unassociated with the educator language behavior. With their presumed deficits in joint attention, this thinking seems to be consistent with observations that, compared to their typically developing peers and those with general developmental delays, children with ASD respond and initiate to attempts to interact or share experience less frequently (Carpenter & Tomasello, 2000; Wetherby et al., 2004). Anecdotally, the researcher observed this interaction style of limited responding and initiating for several students. For initiations, it is plausible that, in some instances, the educators interpreted the children's language behavior as serving no apparent communicative function in the social environment (Harris & Wolchik, 1979). For example, some students in the study repetitively labeled objects in the environment or used phrases/sentences that were replications of passages found in books, songs, or movies. This behavior could be classified as delayed echolalia or

scripted language (Fine et al., 1994). When these anecdotal observations are combined with the unstable estimates of the adult language behaviors, it may be additionally plausible that the adult behaviors are not occurring close enough in time to the child's behavior (Baum, 2002, 2011). In other words, educator behavior is not functionally related to the initiation and responding behaviors of children with ASD. Taken together, future investigations should consider the extent to which sequential associations exist between adult and child language behaviors (Bakeman & Gottman, 1997) as a way to evaluate the plausibility of these hypotheses empirically.

Conceptual Limitations of G-Theory

Despite the recognized benefits of G-Theory to inform the reliability of measurement approaches related to a phenomenon of interest, it does not provide sufficient evidence of the construct validity related to the phenomenon. Reliability is generally associated with the *precision* of measurements (Kane, 1982, 1996); on the other hand, validity, or the degree to which our measurement and theory support inferences, is often conceptualized as a question of the *accuracy* of the inferences (Cone, 1977; Kane, 1982; Yoder et al., 2018). As an example, in the current study, the researcher stipulated and hypothesized that observer and occasion might be pertinent to making precise and ultimately valid conclusions about observed frequencies and durations of adult and child language interaction behaviors. This specification, however, did not and, fundamentally, should not translate into sole evidence of construct validity (Li et al., 2015). That is, G-Theory alone cannot be used to determine what variables are and are not useful or essential to consider when making inferences about a construct (Brennan, 2000). Thus, as researchers seek to examine language-learning environments, it requires careful consideration of the construct itself and the extent to which observable variables represent that construct from both a theoretical and empirical basis.

In the end, the reliability coefficients quantified and differentiated various sources of error and provide necessary estimates of precision in observed scores. What these results do not provide is information about the truth or accuracy of any inferences that researchers could make

about the variables (Brennan, 2000; Kane, 1996). As Brennan (2001) eloquently stated, “Generalizability theory—no matter how extensive—provides only scaffolding. It cannot possibly address all issues that might be of concern to a researcher” (p. 166).

Practical Limitations of the Investigation

With the recognition of the conceptual limits of conducting the g- and d-studies, there also exist practical limitations specific to this investigation. These limitations constrain conclusions relative to two important aspects of this study: (a) observed measurement error and variance components and (b) observed frequency of language behaviors of children with ASD.

Measurement Error within the Current Approach

Several details constrain conclusions made about the measurement error observed in this investigation and thus the reliability of this measurement approach. First, although the sample size was within the range of other investigations ($n = 10 - 38$; Hill et al., 2012; Mantzicopoulos et al., 2018; Praetorius et al., 2014), it was small with only 11 participants and 88 observations. For small sample sizes, the estimated variance components are not only likely to be unstable, but are also prone to have standard error estimates that are higher than the absolute value of the variance component themselves (Cronbach et al., 1972; Li et al., 2015; P. L. Smith, 1978). This latter feature was seen for at least one source of variance in all of the variables measured in the current study, suggesting instability. For example, the estimated contribution of *occasion* to the total variance for the proximity variable was -0.13; the standard error for this same source was higher, at an estimated 0.24. Though some have suggested that the total number of observations be 800 or larger (P. L. Smith, 1978), a series of simulated studies have shown that sample sizes of 50 to 300 can be robust enough for estimation of the variance components and *g*-coefficient (Atilgan, 2013). Until additional work on how sample sizes impact estimates is conducted and given that the variables themselves were highly skewed, the results of this study should be considered preliminary.

The researcher explicitly recognized two hidden facets: the type of classroom and learning context. The potential role or relevance of 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 by the researcher, but were also distinct from, linked to, or confounded the observer or occasion facets in the current study (Webb et al., 2006). Taken together, if researchers or readers intend to apply the results of the current investigation to studies using a similar methodology to language-learning environment of children with ASD (e.g., Qian, 2018; Sanders et al., 2016), these applications should proceed with caution. In addition, researchers may consider intentionally designing methodologies that include the hidden facets of the current investigation. Those methodologies could not only examine the contribution of type of classroom and learning context to measurement error but also add knowledge to the field about the extent to which interaction behaviors generalize over these settings. Toward that end, the conceptual model in Figure 2 may further guide these designs as researchers seek to address the hidden facets of this investigation.

Frequency of Child Language

Child language was the only variable with a reliability coefficient above the expected threshold of 0.80. Although these findings suggest that the estimates obtained for child verbalization may be reliable, any inferences must be qualified. First, four out of the 11 child participants refused to wear microphones for any of the study sessions. The frequency of verbalizations for these participants could not be captured on the video recording to the same degree as their counterparts who did wear their microphones. In future investigations, it may be necessary to have stricter inclusion criteria for child participants that take into account their willingness to wear a recording device.

As a means to identify and explore causal mechanisms and relationships in later experimental studies (Bijou et al., 1968; Loeb et al., 2017), it is essential for descriptive researchers to carefully consider and provide salient information about the participants and

conditions of measurement (e.g., classroom environment, routines, or settings). Descriptive information about the target students beyond basic demographics, such as cognitive skills, language skills, and autism severity that may have been helpful toward that end, however, was not gathered in this study. Given that this study serves as an important first step in developing a method for characterizing the language learning environment for any child with ASD, the researcher was less focused on specific profiles of students and more on the process of gathering information about educator-child interactions. Taken together, readers should be careful when making inferences about the extent to which children with ASD are verbalizing during free play in their inclusive classroom.

Future Directions for Researchers

Despite low-reliability coefficients and general limitations that restrict the conclusions reached, this preliminary investigation serves as a necessary platform for future investigations aimed at understanding the language learning environment of preschoolers with ASD. Considerations for future investigations will be described as they relate to (a) sampling to obtain precise and generalizable estimates, (b) evaluating preschool language environments, and (c) connecting observed language behaviors with moderating variables and outcomes.

Sampling to Obtain Precise and Generalizable Estimates

For researchers that seek to utilize the logic of G-Theory to understand the relative contribution of their procedures to measurement error, the current investigation revealed important considerations. Although occasion is frequently overlooked (Brennan, 2001a) 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 target child, seems to be supported not only by the data in this study and others (Hill et al., 2012; Mantzicopoulos et al., 2018; Praetorius et al., 2014), but also theoretically (Brennan, 2011; Epstein, 1979).

Beyond occasion, it would 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 (Booren et al., 2012; Buysse et al., 1999; Pianta et al., 2009). These structural processes may include adult-child ratio, training and educational level of staff, severity of disability, curriculum, and instructional paradigms (e.g., adult-directed, child-directed, or balanced). This increasingly inclusive universe of generalization comes at the cost of potentially larger error variances and smaller reliability coefficients (Brennan, 2000; Cronbach et al., 1972; Kane, 1982). Nevertheless, it may be more useful, and ultimately of interest, to researchers whose aim is to garner a complete picture of the preschool language learning environment in service of improving outcomes (e.g., Dykstra et al., 2013; Irvin, Boyd, & Odom, 2015; Irvin et al., 2013; Sanders et al., 2016).

Variables to Consider in the Preschool Language Learning Environment

In the early stages of the planning for this investigation, the primary aim was to develop a method for characterizing the types of opportunities to respond (OTRs) provided by educators, as these forms of interactions are frequently tied to improved outcomes (Conroy et al., 2014; Haydon et al., 2012; MacSuga-Gage & Simonsen, 2015). With the primary focus of this study on how the measurement of OTRs relates to identified sources of error, the researcher adopted a more inclusive definition of OTRs. Therefore, the observers coded *any* language directed at the target child that could prompt a response. As other researchers point out, integral in the definition of OTR, however, is an understanding that (a) the educator provides wait time for a child to respond and (b) the prompt is successful in evoking a child response (Greenwood et al., 1994; Lamella & Tincani, 2012; Wasik & Hindman, 2018). Future investigations may include a more nuanced examination of OTRs that incorporates wait time and child responding to evaluate the

degree to which those definitional changes relate to changes in how researchers may be able to obtain stable estimates within their work.

These nuances in the definition of OTRs do not negate the importance of understanding the types of OTRs provided by educators and their relation to child behavior. Some investigations have examined sequential associations between behaviors of caregivers and their child with ASD that are important to language development, such as talk (Brown & Woods, 2016), play (Bottema-Beutel et al., 2017), social communication (Adamson et al., 2001), and joint attention (Brigham et al., 2010). Associations between a preschool educator's and child's language, however, remain largely unexplored. Future investigations should examine the relation between educator and child language behaviors. A more nuanced understanding of the type and frequency with which educators are providing OTRs and the success of those OTRs in evoking a child's response may provide insights that enhance the effective delivery of intervention for children with ASD. With these types of examinations predicated on researchers' understanding the sources of measurement error within their studies, observations of phenomena may begin to reflect the actual frequencies and durations of those phenomena more closely. This examination would allow recommendations to be based on stable, reliable estimates of behaviors.

Connecting Observed Language Behaviors to Moderating Variables and Outcomes

When we have a more reliable and stable measurement of important features of the language learning environment for preschoolers with ASD, we can then examine how participant characteristics interact with that environment. Including measures that evaluate and classify cognitive development, receptive and expressive language development, and ASD severity of the students is crucial to understanding how the environment may change with specific characteristics of the child. This practice is consistent with that described in existing literature aimed at this purpose (Dykstra et al., 2013; Irvin et al., 2013, 2015; Sanders et al., 2016). For example, Irvin et al., (2013) demonstrated negative associations, such that, as the severity of the child's disability increased, adult word count decreased. This finding highlights the need for examining relations

between interaction behaviors and other variables once sources of measurement error are known and addressed through a data-informed sampling plan. Examinations of specific adult characteristics further reinforce this need; as educator burnout increased, adult word count decreased (Irvin et al., 2013). In that vein, it is not merely the child's characteristics that are critical to examine, but also the educator's characteristics, as they may be related to their implementation (or lack of implementation) of strategies aimed at promoting language development.

Accurate and reliable measurement of the language-learning environment is particularly germane for researchers and practitioners when the classification of the environment moves beyond merely stating basic descriptors to establishing and explaining the link between environmental features and child outcomes, both proximal and distal. For example, McDuffie and Yoder (2010) found a distal link, showing that parent use of comments and directives, which aligned with the focus of attention of his/her young child with ASD, was positively correlated with vocabulary 6 months later. Employing a similar methodology to establish links like this one, between the form and frequency of educator language and key child language outcomes, becomes a logical next step to follow the current investigations. Future researchers could quantify not only the moment-to-moment impact of the language learning environment but also the cumulative impact of it over time (Ford et al., 2020; Hart & Risley, 1995; Warren et al., 2007)

Conclusion

Issues in the reliability of the measurement procedures can thwart the validity of a researcher's inferences about a construct of interest. As a way to examine the contribution of these procedures to the error, employing the logic of generalizability theory can be both useful and advantageous, as illustrated in the current investigation. In this study, the researcher conducted a series of generalizability studies on measured educator and child variables, which highlighted occasion as a substantial contributor to measurement error within educator variables, though not in the child language variable. With the contribution of observer virtually negligible

for all variables, the researcher systematically only manipulated occasion in the d-studies to determine how to optimize sampling. 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. For child language, on the other hand, only three sessions were deemed to be required. As researchers must balance technical requirements against practical cost, it may be necessary to consider what and how variables are measured relative to desired inferences that are in service of accurate characterization of language learning environments for children with ASD. In the end, it is the responsibility of researchers to carefully consider their measurement and sampling of variables that have the most significant impact on language development in hopes of moving the needle forward for the roughly 20,000 children with ASD that enter kindergarten with significant expressive language delays (Anderson et al., 2007; Norrelgen et al., 2015).

References

- Adamson, L. B., McArthur, D., Markov, Y., Dunbar, B., & Bakeman, R. (2001). Autism and joint attention: Young children's responses to maternal bids. *Journal of Applied Developmental Psychology, 22*(4), 439–453. [https://doi.org/10.1016/S0193-3973\(01\)00089-2](https://doi.org/10.1016/S0193-3973(01)00089-2)
- Akamoglu, Y., & Meadan, H. (2018). Parent-implemented language and communication interventions for children with developmental delays and disabilities: A scoping review. *Review Journal of Autism and Developmental Disorders, 5*(3), 294–309. <https://doi.org/10.1007/s40489-018-0140-x>
- American Psychiatric Association. (2013). Neurodevelopmental Disorders. In *Diagnostic and statistical manual of mental disorders* (5th ed.). American Psychiatric Publishing. <https://doi.org/http://dx.doi.org/10.1016/B978-012373960-5.00026-5>
- American Speech-Language-Hearing Association. (2017). *Autism signs and symptoms*. Autism Practice Portal. http://www.asha.org/PRPSpecificTopic.aspx?folderid=8589935303§ion=Signs_and_Symptoms
- Anderson, D. K., Lord, C., Risi, S., DiLavore, P. S., Shulman, C., Thurm, A., Welch, K., & Pickles, A. (2007). Patterns of growth in verbal abilities among children with autism spectrum disorder. *Journal of Consulting and Clinical Psychology, 75*(4), 594–604. <https://doi.org/10.1037/0022-006X.75.4.594>
- Atilgan, H. (2013). Sample size estimation of g and phi coefficients in generalizability theory. *Eurasian Journal of Educational Research, 51*, 215–227.
- Bakeman, R., & Gottman, J. M. (1997). *Observing interaction: An introduction to sequential analysis*. Cambridge University Press.
- Bakeman, R., & Quera, V. (2011). *Sequential analysis and observational methods for the behavioral sciences*. Cambridge University Press.
- Barton, E. E., & Smith, B. J. (2015). Advancing high-quality preschool inclusion. *Topics in Early*

- Childhood Special Education*, 35(2), 69–78. <https://doi.org/10.1177/0271121415583048>
- 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>
- Baum, W. M. (2011). Behaviorism, private events, and the molar view of behavior. *Behavior Analyst*, 34(2), 185–200. <https://doi.org/10.1007/BF03392249>
- Berger, L. M., Hill, J., & Waldfogel, J. (2005). Maternity leave, early maternal employment and child health and development in the US. *The Economic Journal*, 115(501), F29–F47. <https://doi.org/10.1111/j.0013-0133.2005.00971.x>
- Bijou, S. W., Peterson, R. F., & Ault, M. H. (1968). A method to integrate descriptive and experimental field studies at the level of data and empirical concepts. *Journal of Applied Behavior Analysis*, 1(2), 175–191. <https://doi.org/10.1901/jaba.1968.1-175>
- Bleses, D., Makransky, G., Dale, P. S., Højen, A., & Ari, B. A. (2016). Early productive vocabulary predicts academic achievement 10 years later. *Applied Psycholinguistics*, 37(6), 1461–1476. <https://doi.org/10.1017/S0142716416000060>
- Booren, L. M., Downer, J. T., & Vitiello, V. E. (2012). Observations of children’s interactions with teachers, peers, and tasks across preschool classroom activity settings. *Early Education & Development*, 23(4), 517–538. <https://doi.org/10.1080/10409289.2010.548767>
- 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>
- Bottema-Beutel, K., Malloy, C., Lloyd, B. P., Louick, R., Joffe-Nelson, L., Watson, L. R., & Yoder, P. J. (2017). Sequential associations between caregiver talk and child play in autism spectrum disorder and typical development. *Child Development*, 89(3), 157–166. <https://doi.org/10.1111/cdev.12848>

- Brennan, R. L. (2000). (Mis)conception about generalizability theory. *Educational Measurement: Issues and Practice*, 19(1), 5–10. <https://doi.org/10.1111/j.1745-3992.2000.tb00017.x>
- Brennan, R. L. (2001a). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement*, 38(4), 295–317.
<https://www.jstor.org/stable/1435452>
- Brennan, R. L. (2001b). *Generalizability theory*. Springer.
- Brennan, R. L. (2003). Coefficients and indices in generalizability theory. In *Center for Advanced Studies in Measurement and Assessment* (Issue 1).
- 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>
- Briesch, A. M., Swaminathan, H., Welsh, M., & Chafouleas, S. M. (2014). Generalizability theory: A practical guide to study design, implementation, and interpretation. *Journal of School Psychology*, 52(1), 13–35. <https://doi.org/10.1016/j.jsp.2013.11.008>
- Brigham, N. B., Yoder, P. J., Jarzynka, M. A., & Tapp, J. (2010). The sequential relationship between parent attentional cues and sustained attention to objects in young children with autism. *Journal of Autism and Developmental Disorders*, 40(2), 200–208.
<https://doi.org/10.1007/s10803-009-0848-7>
- Brignell, A., Morgan, A. T., Woolfenden, S., Klopper, F., May, T., Sarkozy, V., & Williams, K. (2018). A systematic review and meta-analysis of the prognosis of language outcomes for individuals with autism spectrum disorder. *Autism & Developmental Language Impairments*, 3, 1–15. <https://doi.org/10.1177/2396941518767610>
- Bronfenbrenner, U. (2001). The bioecological theory of human development. In N. J. Smelser & P. B. Baltes (Eds.), *Encyclopedia of psychology* (pp. 6963–6970). Elsevier.
- Bronfenbrenner, U., & Morris, P. A. (2006). The bioecological model of human development. In N. J. Smelser & P. B. Baltes (Eds.), *Encyclopedia of psychology* (pp. 793–828). Wiley.
- Brown, J. A., & Woods, J. J. (2016). Parent-implemented communication intervention:

- Sequential analysis of triadic relationships. *Topics in Early Childhood Special Education*, 36(2), 115–124. <https://doi.org/10.1177/0271121416628200>
- 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(2), 139–153. <https://doi.org/10.1177/105381510602800205>
- 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>
- Buyse, V., Wesley, P. W., Bryant, D., & Gardner, D. (1999). Quality of early childhood programs in inclusive and noninclusive settings. *Exceptional Children*, 65(3), 301–314. <https://doi.org/10.1177/001440299906500302>
- Cardinet, J., Johnson, S., & Pini, G. (2010). *Applying generalizability theory using EduG*. Routledge.
- Carpenter, M., & Tomasello, M. (2000). Joint attention and cultural learning. In A. M. Wetherby & B. M. Prizant (Eds.), *Autism spectrum disorders: A transactional developmental perspective [Volume 9]* (pp. 31–54). Paul H Brookes Publishing.
- Casby, M. W. (2011). An examination of the relationship of sample size and mean length of utterance for children with developmental language impairment. *Child Language Teaching and Therapy*, 27(3), 286–293. <https://doi.org/10.1177/0265659010394387>
- Casenhiser, D. M., Shanker, S. G., & Stieben, J. (2013). Learning through interaction in children with autism: Preliminary data from a social-communication-based intervention. *Autism*, 17(2), 220–241. <https://doi.org/10.1177/1362361311422052>
- Chafouleas, S. M., Christ, T. J., Riley-Tillman, T. C., Briesch, A. M., & Chanese, J. A. M. (2007). Generalizability and dependability of direct behavior ratings to assess social behavior of preschoolers. *School Psychology Review*, 36(1), 63–79.

<https://doi.org/10.1080/02796015.2007.12087952>

- Christina, R., & Goodman, J. (2005). *Going to scale with high-quality early education*. RAND Corporation. <https://doi.org/10.7249/TR237>
- 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](https://doi.org/10.1016/S0005-7894(77)80077-4)
- Conn-Powers, M., Cross, A. F., Traub, E. K., & Hutter-Pishgahi, L. (2006). *The universal design of early education: Moving forward for all children*. Beyond the Journal: Young Children on the Web. https://www.iidc.indiana.edu/styles/iidc/defiles/ECC/ECC_Universal_Design_Early_Education.pdf
- Conroy, M. A., Sutherland, K. S., Vo, A. K., Carr, S., & Ogston, P. L. (2014). Early childhood teachers' use of effective instructional practices and the collateral effects on young children's behavior. *Journal of Positive Behavior Interventions*, 16(2), 81–92. <https://doi.org/10.1177/1098300713478666>
- Craig, H. K., Washington, J. A., & Thompson-Porter, C. (1998). Average c-unit lengths in the discourse of African American Children from low-income, urban homes. *Journal of Speech, Language, and Hearing Research*, 41(2), 433–444. <https://doi.org/10.1044/jslhr.4102.433>
- Cristofaro, T. N., & Tamis-LeMonda, C. S. (2012). Mother-child conversations at 36 months and at pre-kindergarten: Relations to children's school readiness. *Journal of Early Childhood Literacy*, 12(1), 68–97. <https://doi.org/10.1177/1468798411416879>
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334. <https://doi.org/10.1007/BF02310555>
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements*. 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.

<https://doi.org/10.1111/j.2044-8317.1963.tb00206.x>

- Cunningham, J. E., Zimmerman, K. N., Ledford, J. R., & Kaiser, A. P. (2019). Comparison of measurement systems for collecting teacher language data in early childhood settings. *Early Childhood Research Quarterly, 49*, 164–174. <https://doi.org/10.1016/j.ecresq.2019.06.008>
- De Rivera, C., Girolametto, L., Greenberg, J., & Weitzman, E. (2005). Children's responses to educators' questions in day care play groups. *American Journal of Speech-Language Pathology, 14*(1), 14–26. [https://doi.org/10.1044/1058-0360\(2005/004\)](https://doi.org/10.1044/1058-0360(2005/004))
- DEC/NAEYC. (2009). *Early childhood inclusion: A joint position statement of the Division for Early Childhood (DEC) and the National Association for the Education of Young Children (NAEYC)*. <https://doi.org/10.1177/1096250609347736>
- Dickinson, D. K., Porche, M. V., Dickinson, D. K., & Porche, M. V. (2011). Relation between language experiences in preschool classrooms and children's kindergarten and fourth-grade language and reading abilities. *Child Development, 82*(3), 870–886.
- Doyle, P. M., Wolery, M., Ault, M. J., & Gast, D. L. (1988). System of least prompts: A literature review of procedural parameters. *Research and Practice for Persons with Severe Disabilities, 13*(1), 28–40. <https://doi.org/10.1177/154079698801300104>
- Dunst, C. J., Lowe, L. W., & Bartholomew, P. C. (1990). Contingent social responsiveness, family ecology, and infant communicative competence. *National Student Speech Language Hearing Association Journal, 17*, 39–49.
- Dunst, C. J., Raab, M., & Trivette, C. M. (2011). Characteristics of naturalistic language intervention strategies. *Journal of Speech-Language Pathology and Applied Behavior Analysis, 5*(1), 8–16.
- 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\(2012/11-0089\)](https://doi.org/10.1044/0161-1461(2012/11-0089))
- Epstein, S. (1979). The stability of behavior: I. On predicting most of the people much of the time. *Journal of Personality and Social Psychology, 37*(7), 1097–1126. <https://doi.org/10.1037/0022-3514.37.7.1097>
- Fernald, A., Marchman, V. A., & Weisleder, A. (2013). SES differences in language processing skill and vocabulary are evident at 18 months. *Developmental Science, 16*(2), 234–248. <https://doi.org/10.1111/desc.12019>
- Fernald, A., & Weisleder, A. (2011). Early language experience is vital to developing fluency in understanding. In S. B. Neuman & D. K. Dickinson (Eds.), *Handbook of early literacy research* (Vol. 3, pp. 3–19). Guilford Press.
- File, N. (1994). Children’s play, teacher-child interactions, and teacher beliefs in integrated early childhood programs. *Early Childhood Research Quarterly, 9*(2), 223–240. [https://doi.org/10.1016/0885-2006\(94\)90007-8](https://doi.org/10.1016/0885-2006(94)90007-8)
- 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>
- Freeman, S., & Kasari, C. (2013). Parent–child interactions in autism: Characteristics of play. *Autism, 17*(2), 147–161. <https://doi.org/10.1177/1362361312469269>
- 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>
- Gast, D. L., & Ledford, J. R. (2014). *Single case research methodology: Applications in special education and behavioral sciences* (2nd Ed.). Routledge.
- Gersten, R., Baker, S., & Lloyd, J. W. (2003). Designing high-quality research in special

- education: Group experimental design. *The Journal of Special Education*, 34(1), 2–18.
- Gest, S. D., Holland-Coviello, R., Welsh, J. A., Eicher-Catt, D. L., & Gill, S. (2006). Language development subcontexts in head start classrooms: Distinctive patterns of teacher talk during free play, mealtime, and book reading. *Early Education & Development*, 17(2), 293–315. https://doi.org/10.1207/s15566935eed1702_5
- Girolametto, L., Hoaken, L., Weitzman, E., & Lieshout, R. V. (2000). Patterns of adult-child linguistic interaction in integrated day care groups. *Language, Speech, and Hearing Services in Schools*, 31(2), 155–168. <https://doi.org/10.1044/0161-1461.3102.155>
- Gleser, G. C., Cronbach, L. J., & Rajaratnam, N. (1965). Generalizability of scores influenced by multiple sources of variance. *Psychometrika*, 30(4), 395–418. <https://doi.org/10.1007/BF02289531>
- Golinkoff, R. M., Hoff, E., Rowe, M. L., Tamis-LeMonda, C. S., & Hirsh-Pasek, K. (2018). Language matters: Denying the existence of the 30-million-word gap has serious consequences. *Child Development*. <https://doi.org/10.1111/cdev.13128>
- Greenwood, C. R., Delquadri, J., & Hall, R. V. (1984). Opportunity to respond and student academic performance. In W. L. Heward, T. E. Heron, Trapp-Porter, & D. S. Hill (Eds.), *Focus on behavior analysis in education* (pp. 58–88). Charles Merrill.
- Greenwood, C. R., Hart, B., Walker, D., Risley, T., Gardner III, R., Sainato, D. M., Cooper, J. O., Heron, T. E., Heward, W. L., Eshleman, J. W., & Grossi, T. A. (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).
- Gutiérrez, K. D., & Penuel, W. R. (2014). Relevance to practice as a criterion for rigor. *Educational Researcher*, 43(1), 19–23. <https://doi.org/10.3102/0013189X13520289>
- Haebig, E., McDuffie, A., & Ellis Weismer, S. (2013). The contribution of two categories of

- parent verbal responsiveness to later language for toddlers and preschoolers on the autism spectrum. *American Journal of Speech-Language Pathology*, 22(1), 57.
[https://doi.org/10.1044/1058-0360\(2012/11-0004\)](https://doi.org/10.1044/1058-0360(2012/11-0004))
- Hall, R. V., Delquadri, J., Greenwood, C. R., & Thurston, L. (1982). The importance of opportunity to respond in children's academic success. In E. Edgar, N. Haring, J. Jenkins, & C. Pious (Eds.), *Mentally handicapped children: Education and training* (pp. 107–140). University Park Press.
- Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, 8(1), 23–34.
<https://doi.org/10.1016/j.biotechadv.2011.08.021.Secreted>
- Hammer, C. S., Farkas, G., & Maczuga, S. (2010). The language and literacy development of Head Start children: A study using the Family and Child Experiences Survey database. *Language Speech and Hearing Services in Schools*, 41(1), 70. [https://doi.org/10.1044/0161-1461\(2009/08-0050\)](https://doi.org/10.1044/0161-1461(2009/08-0050))
- 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/http://dx.doi.org/10.1111/jir.12283>
- Harris, S. L., & Wolchik, S. A. (1979). Suppression of self-stimulation: three alternative strategies. *Journal of Applied Behavior Analysis*, 12(2), 131–136.
<https://doi.org/10.1901/jaba.1979.12-185>
- Hart, B., & Risley, T. R. (1975). Incidental teaching of language in the preschool. *Journal of Applied Behavior Analysis*, 8(4), 411–420. <https://doi.org/10.1901/jaba.1975.8-411>
- Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Paul H Brookes Publishing.
- Haydon, T., MacSuga-Gage, A. S., Simonsen, B., & Hawkins, R. (2012). Opportunities to respond: A key component of effective instruction. *Beyond Behavior*, 22(1), 23–31.

<https://doi.org/10.1177/1476127005050030>

Hendrickson, A., & Yin, P. (2010). Generalizability Theory. In G. R. Hancock & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences*. Routledge.

Hepting, N. H., & Goldstein, H. (1996). What's natural about naturalistic language intervention? *Journal of Early Intervention, 20*(3), 249–264.

<https://doi.org/10.1177/105381519602000308>

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>

Hoff, E. (2005). *Language development*. Wadsworth/Thomson Learning.

Hoff, E. (2006). How social contexts support and shape language development. *Developmental Review, 26*(1), 55–88. <https://doi.org/10.1016/j.dr.2005.11.002>

Hollich, G. J., Hirsh-Pasek, K., Golinkoff, R. M., Brand, R. J., Brown, E., Chung, H. L., Hennon, E., & Rocroi, C. (2000). Breaking the language barrier: An emergentist coalition model for the origins of word learning. *Monographs of the Society for Research in Child Development, 65*(3), i–vi, 1–123. <https://doi.org/10.1111/1540-5834.00090>

House, A. E., House, B. J., & Campbell, M. B. (1981). Measures of interobserver agreement: Calculation formulas and distribution effects. *Journal of Behavioral Assessment, 3*(1), 37–57. <https://doi.org/10.1007/BF01321350>

Hudry, K., Aldred, C., Wigham, S., Green, J., Leadbitter, K., Temple, K., Barlow, K., & McConachie, H. (2013). Predictors of parent–child interaction style in dyads with autism. *Research in Developmental Disabilities, 34*(10), 3400–3410.

<https://doi.org/https://doi.org/10.1016/j.ridd.2013.07.015>

Irvin, D. W., Boyd, B. A., & Odom, S. L. (2015). Child and setting characteristics affecting the adult talk directed at preschoolers with autism spectrum disorder in the inclusive classroom. *Autism, 19*(2), 223–234. <https://doi.org/10.1177/1362361313517398>

- 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>
- Kaiser, A. P., Hester, P. P., & McDuffie, A. S. (2001). Supporting communication in young children with developmental disabilities. *Mental Retardation and Developmental Disabilities Research Reviews*, 7(2), 143–150. <https://doi.org/10.1002/mrdd.1020>
- Kane, M. (1982). A sampling model for validity. *Applied Psychological Measurement*, 6(2), 125–160. <https://doi.org/10.1177/014662168200600201>
- Kane, M. (1996). The precision of measurements. *Applied Measurement in Education*, 9(4), 355–379. https://doi.org/10.1207/s15324818ame0904_4
- Kane, M. (2002). Inferences about variance components and reliability-generalizability coefficients in the absence of random sampling. *Journal of Educational Measurement*, 39(2), 165–181. <https://www.jstor.org/stable/1435254>
- Kontos, S. (1999). Preschool teachers' talk, roles, and activity settings during free play. *Early Childhood Research Quarterly*, 14(3), 363–382. [https://doi.org/10.1016/S0885-2006\(99\)00016-2](https://doi.org/10.1016/S0885-2006(99)00016-2)
- Kontos, S., Burchinal, M., Howes, C., Wisseh, S., & Galinsky, E. (2002). An eco-behavioral approach to examining the contextual effects of early childhood classrooms. *Early Childhood Research Quarterly*, 17(2), 239–258. [https://doi.org/10.1016/S0885-2006\(02\)00147-3](https://doi.org/10.1016/S0885-2006(02)00147-3)
- Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2013). Single-case intervention research design standards. *Remedial and Special Education*, 34(1), 26–38. <https://doi.org/10.1177/0741932512452794>
- Kublin, K. S., Wetherby, A. M., Crais, E. R., & Prizant, B. M. (1998). Prelinguistic dynamic assessment: A transactional perspective. In A. M. Wetherby, S. F. Warren, & J. Reichle

- (Eds.), *Transitions in prelinguistic communication* (Volume 7, pp. 285–312). Paul H Brookes Publishing.
- Kuhl, P. K. (2010). Brain mechanisms in early language acquisition. *Neuron*, *67*(5), 713–727.
<https://doi.org/10.1016/j.neuron.2010.08.038>
- Kuhl, P. K., Tsao, F.-M., & Liu, H.-M. (2003). Foreign-language experience in infancy: Effects of short-term exposure and social interaction on phonetic learning. *Proceedings of the National Academy of Sciences*, *100*(15), 9096–9101.
<https://doi.org/10.1073/pnas.1532872100>
- Lamella, L., & Tincani, M. (2012). Brief wait time to increase response opportunity and correct responding of children with autism spectrum disorder who display challenging behavior. *Journal of Developmental and Physical Disabilities*, *24*(6), 559–573.
<https://doi.org/10.1007/s10882-012-9289-x>
- Landa, R. J. (2007). Early communication development and intervention for children with autism. *Mental Retardation and Developmental Disabilities Research Reviews*, *13*, 16–25.
<https://doi.org/10.1002/mrdd>
- Li, M., Shavelson, R. J., Yin, Y., & Wiley, E. (2015). Generalizability theory. In *The encyclopedia of clinical psychology* (pp. 1322–1341). John Wiley & Sons, Inc.
<https://doi.org/10.1002/9781118625392.wbecp352>
- Light, J., & McNaughton, D. (2014). Communicative competence for individuals who require augmentative and alternative communication: A new definition for a new era of communication? *AAC: Augmentative and Alternative Communication*, *30*(1), 1–18.
<https://doi.org/10.3109/07434618.2014.885080>
- Loban, W. (1966). *The language of elementary school children*.
- Loban, W. (1976). *Language development: Kindergarten through grade twelve*.
- Loeb, S., Dynarski, S., McFarland, D., Morris, P., Reardon, S., & Reber, S. (2017). Descriptive analysis in education: A guide for researchers. *U.S. Department of Education, Institute of*

- Education Sciences. National Center for Education Evaluation and Regional Assistance, March*, 1–40. <https://doi.org/10.1094/PDIS.2003.87.5.550>
- Lord, F., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.
- MacSuga-Gage, A. S., & Simonsen, B. (2015). Examining the effects of teacher-directed opportunities to respond on student outcomes: A systematic review of the literature. *Education and Treatment of Children*, 38(2), 211–239. <https://doi.org/10.1353/etc.2015.0009>
- 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>
- Massey, S. L. (2003). Teacher-child conversation in the preschool classroom. *Early Childhood Education Journal*, 31(4), 227–231. <https://doi.org/10.1023/B:ECEJ.0000024113.69141.23>
- McDuffie, A., & Yoder, P. (2010). Types of parent verbal responsiveness that predict language in young children with autism spectrum disorder. *Journal of Speech, Language, and Hearing Research*, 53(4), 1026–1039. [https://doi.org/10.1044/1092-4388\(2009/09-0023\)](https://doi.org/10.1044/1092-4388(2009/09-0023))
- McWilliam, R. A., & Casey, A. (2008). *Engagement of every child in the preschool classroom*. Paul H. Brookes Publishing.
- 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>
- Meadan, H., Ostrosky, M. M., Zaghawan, H. Y., & Yu, S. (2009). Promoting the social and communicative behavior of young children with autism spectrum disorders. *Topics in Early Childhood Special Education*, 29(2), 90–104. <https://doi.org/10.1177/0271121409337950>
- Miller, J. (1981). *Assessing language production in children*. Allyn and Bacon.
- Morra Pellegrino, M. L., & Scopesi, A. (1990). Structure and function of baby talk in a day-care

- centre. *Journal of Child Language*, 17(1), 101–114.
<https://doi.org/10.1017/S030500090001312X>
- 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>
- Norrelgen, F., Fernell, E., Eriksson, M., Hedvall, Å., Persson, C., Sjölin, M., Gillberg, C., & Kjellmer, L. (2015). Children with autism spectrum disorders who do not develop phrase speech in the preschool years. *Autism*, 19(8), 934–943.
<https://doi.org/10.1177/1362361314556782>
- Owens, R. E. (2016). *Language Development: An Introduction* (9th ed.). Pearson.
- Paul, R. (2007). *Language disorders from infancy through adolescence* (3rd ed.). Mosby.
- Paul, R. (2008). Interventions to improve communication in autism. *Child and Adolescent Psychiatric Clinics of North America*, 17(4), 835–856.
<https://doi.org/10.1016/j.chc.2008.06.011>
- Pavelko, S. L., & Owens Jr., R. E. (2017). Sampling utterances and grammatical analysis revised (SUGAR): New normative values for language sample analysis measures. *Language, Speech & Hearing Services in Schools*, 48(3), 197–215.
https://doi.org/10.1044/2017_LSHSS-17-0022
- 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>
- Praetorius, A. K., Pauli, C., Reusser, K., Rakoczy, K., & Klieme, E. (2014). One lesson is all you

- need? Stability of instructional quality across lessons. *Learning and Instruction*, 31, 2–12.
<https://doi.org/10.1016/j.learninstruc.2013.12.002>
- Pruden, S. M., Hirsh-Pasek, K., & Golinkoff, R. M. (2006). The social dimension in language development. In P. Marshall & N. Fox (Eds.), *The development of social engagement: Neurobiological perspectives* (pp. 118–152). Oxford University Press.
https://doi.org/10.1007/SpringerReference_223340
- 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>
- R Core Team. (2015). *RStudio: Integrated development for R*. R Studio, Inc.
<http://www.rstudio.com/>
- Rimm-Kaufman, S. E., Voorhees, M. D., Snell, M. E., & La Paro, K. M. (2003). Improving the sensitivity and responsivity of preservice teachers toward young children with disabilities. *Topics in Early Childhood Sepcial Education*, 23(3), 151–163.
<https://doi.org/10.1177/02711214030230030501>
- Rodriguez, E. T., Tamis-LeMonda, C. S., Spellmann, M. E., Pan, B. A., Raikes, H., Lugo-Gil, J., & Luze, G. (2009). The formative role of home literacy experiences across the first three years of life in children from low-income families. *Journal of Applied Developmental Psychology*, 30(6), 677–694. <https://doi.org/10.1016/j.appdev.2009.01.003>
- Romeo, R. R., Leonard, J. A., Robinson, S. T., West, M. R., Mackey, A. P., Rowe, M. L., & Gabrieli, J. D. E. (2018). Beyond the 30-Million-Word Gap: Children’s conversational exposure is associated with language-related brain function. *Psychological Science*, 29(5), 700–710. <https://doi.org/10.1177/0956797617742725>
- Rowe, M. L. (2008). Child-directed speech: Relation to socioeconomic status, knowledge of child development and child vocabulary skill. *Journal of Child Language*, 35(01), 185–205.
<https://doi.org/10.1017/S0305000907008343>

- Rowe, M. L., Leech, K. A., & Cabrera, N. (2017). Going beyond input quantity: Wh-questions matter for toddlers' language and cognitive development. *Cognitive Science*, *41*, 162–179. <https://doi.org/10.1111/cogs.12349>
- Rowe, M. L., & Snow, C. E. (2019). Analyzing input quality along three dimensions: Interactive, linguistic, and conceptual. *Journal of Child Language*, 1–17. <https://doi.org/10.1017/S0305000919000655>
- Rowland, C. F., Pine, J. M., Lieven, E. V. M., & Theakston, A. L. (2003). Determinants of acquisition order in wh-questions: Re-evaluating the role of caregiver speech. *Journal of Child Language*, *30*(3), 609–635. <https://doi.org/10.1017/S0305000903005695>
- Sam, A. M., Reszka, S. S., Boyd, B. A., Pan, Y., Hume, K., & Odom, S. L. (2016). The association between adult participation and the engagement of preschoolers with ASD. *Autism Research and Treatment*, *2016*, 1–10. <https://doi.org/10.1155/2016/6029837>
- Sameroff, A. J. (2009). The transactional model. In A. J. Sameroff (Ed.), *The transactional model of development: How children and contexts shape each other*. (pp. 3–21). American Psychological Association. <https://doi.org/10.1037/11877-001>
- 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>
- Scarborough, H. S. (2001). Connecting early language and literacy to later reading (dis)abilities: Evidence, theory, and practice. In S. B. Neuman & D. K. Dickinson (Eds.), *Handbook of early literacy* (Vol. 1, pp. 97–110). Guilford Press.
- 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.
- Shneidman, L. A., Arroyo, M. E., Levine, S. C., & Goldin-Meadow, S. (2013). What counts as effective input for word learning? *Journal of Child Language*, *40*(3), 672–686.
<https://doi.org/10.1017/S0305000912000141>
- Shonkoff, J. P., & Phillips, D. A. (2000). *From Neurons to Neighborhoods*. National Academy Press. <https://doi.org/10.17226/9824>
- Shumate, S. R., Surles, J., Johnson, R. L., & Penny, J. (2007). The effects of the number of scale points and non-normality on the generalizability coefficient: A Monte Carlo study. *Applied Measurement in Education*, *20*(4), 357–376. <https://doi.org/10.1080/08957340701429645>
- Siller, M., & Sigman, M. (2008). Modeling longitudinal change in the language abilities of children with autism: Parent behaviors and child characteristics as predictors of change. *Developmental Psychology*, *44*(6), 1691–1704. <https://doi.org/10.1037/a0013771>
- Simonsen, B., Myers, D., & DeLuca, C. (2010). Teaching teachers to use prompts, opportunities to respond, and specific praise. *Teacher Education and Special Education*, *33*(4), 300–318. <https://doi.org/10.1177/0888406409359905>
- Singer, E., Nederend, M., Penninx, L., Tajik, M., & Boom, J. (2014). The teacher's role in supporting young children's level of play engagement. *Early Child Development and Care*, *184*(8), 1233–1249. <https://doi.org/10.1080/03004430.2013.862530>
- Smith, M. W., & Dickinson, D. K. (1994). Describing oral language opportunities and environments in head start and other preschool classrooms. *Early Childhood Research Quarterly*, *9*(3–4), 345–366. [https://doi.org/10.1016/0885-2006\(94\)90014-0](https://doi.org/10.1016/0885-2006(94)90014-0)
- Smith, P. L. (1978). Sampling errors of variance components in small sample multifacet generalizability studies. *Journal of Educational Statistics*, *3*(4), 319.
<https://doi.org/10.2307/1164776>
- Spiker, D., Boyce, G. C., & Boyce, L. K. (2002). *Parent-child interactions when young children*

- have disabilities* (Vol. 25, pp. 35–70). Academic Press.
- [https://doi.org/https://doi.org/10.1016/S0074-7750\(02\)80005-2](https://doi.org/https://doi.org/10.1016/S0074-7750(02)80005-2)
- Suen, H. K., & Ary, D. (1989). Reliability: The generalizability approach. In *Analyzing quantitative behavioral observation data* (pp. 131–156). Lawrence Erlbaum Associates, Inc.
- Swiss Society for Research in Education Working Group. (2012). *EduG User Guide* (6.1).
- Tapp, J., Wehby, J., & Ellis, D. (1995). A multiple option observation system for experimental studies: MOOSES. *Behavior Research Methods, Instruments, & Computers*, 27(1), 25–31.
- <https://doi.org/10.3758/BF03203616>
- Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press.
- Tomasello, M., & Farrar, M. J. (1986). Joint attention and early language. *Child Development*, 57(6), 1454–1463. <http://www.jstor.org/stable/1130423>
- U.S. Department of Education. (2018). *Number of children ages 3 through 5 served under IDEA, Part B, by disability and state: 2017-2018*. <http://www2.ed.gov/programs/osepidea/618-data/static-tables/index.html>
- Venter, A., Lord, C., & Schopler, E. (1992). A follow-up study of high-functioning autistic children. *Journal of Child Psychology and Psychiatry*, 33(3), 489–597.
- <https://doi.org/10.1111/j.1469-7610.1992.tb00887.x>
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.
- Walker, S., & Berthelsen, D. (2008). Children with autistic spectrum disorder in early childhood education programs: A social constructivist perspective on inclusion. *International Journal of Early Childhood*, 40(1), 33–51. <https://doi.org/10.1007/BF03168362>
- Warren, S. F. (2015). Right from birth: Eliminating the talk gap in young children. In *LENA Foundation* (Issue May).
- Warren, S. F., Fey, M. E., & Yoder, P. J. (2007). Differential treatment intensity research: A

- missing link to creating optimally effective communication interventions. *Mental Retardation and Developmental Disabilities Research Reviews*, 13(1), 70–77.
<https://doi.org/10.1002/mrdd.20139>
- Warren, S. F., Gilkerson, J., Richards, J. A., Oller, D. K., Xu, D., Yapanel, U., & Gray, S. (2010). What automated vocal analysis reveals about the vocal production and language learning environment of young children with autism. *Journal of Autism and Developmental Disorders*, 40(5), 555–569. <https://doi.org/10.1007/s10803-009-0902-5>
- Warren, S. F., & Walker, D. (2005). Handbook of research methods in developmental science. In D. M. Teti (Ed.), *Handbook of research methods in developmental science* (pp. 249–270). Blackwell Publishing Ltd. <https://doi.org/10.1111/b.9780631222618.2004.00015.x>
- Wasik, B. A., & Hindman, A. H. (2018). Why wait? The importance of wait time in developing young students' language and vocabulary skills. *Reading Teacher*, 72(3), 369–378.
<https://doi.org/10.1002/trtr.1730>
- Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2006). Reliability coefficients and generalizability theory. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 81–124). [https://doi.org/10.1016/S0169-7161\(06\)26004-8](https://doi.org/10.1016/S0169-7161(06)26004-8)
- Weisleder, A., & Fernald, A. (2013). Talking to children matters: Early language experience strengthens processing and builds vocabulary. *Psychological Science*, 24(11), 2143–2152.
<https://doi.org/10.1177/0956797613488145>
- Wetherby, A. M., Prizant, B. M., & Hutchinson, T. A. (1998). Communicative, social/affective, and symbolic profiles of young children with autism and pervasive developmental disorders. *American Journal of Speech-Language Pathology*, 7, 79–91.
<http://www.annualreviews.org/doi/10.1146/annurev.psych.56.091103.070159>
- 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>

What Works Clearinghouse. (2008). Evidence standards for reviewing studies. In *What Works Clearinghouse* (Issue May).

Wittmer, D. S., & Honig, A. S. (1991). Convergent or divergent? Teacher questions to three-year-old children in day care. *Early Child Development and Care*, 68(1), 141–147.

<https://doi.org/10.1080/0300443910680113>

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