AUTHOR QUERY FORM

Journal title: TEC

Article Number: 528052

Dear Author/Editor,

Greetings, and thank you for publishing with SAGE. Your article has been copyedited, and we have a few queries for you. Please respond to these queries when you submit your changes to the Production Editor.

Thank you for your time and effort.

Please assist us by clarifying the following queries:

No	Query
1	Please check that (a) all authors are listed in the proper order; (b) clarify which part of each author's name is his or her surname; (c) verify that all author names are correctly spelled/punctuated and are presented in a manner consistent with any prior publications; and (d) check that all author information, such as affiliations and contact information, appears accurately.
2	Please review the entire document for typographical errors, mathematical errors, and any other necessary corrections; check headings, tables, and figures.
3	Please check whether the affiliation and the corresponding author details are correct.
4	Please provide complete reference details for "McCormick and Nellis 2004" or allow us to delete the citation.
5	Please check whether the Research Questions are numbered and set correctly.
6	Please check whether the edits made in Tables 1 to 4 are correct.
7	Please provide complete reference details for "Abidin 1995" or allow us to delete the citation.
8	Please provide the expansion of "IOA."
9	Note that the statement "The positions stated here are not . . . or the Vanderbilt Kennedy Center" has been set as an Authors' Note. Please check whether this is correct.
10	Please confirm whether the given conflict of interest and funding statements are accurate: "The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article."
11	Please confirm whether the given funding statement is accurate: "The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported in part by NIDCD Grant RO15DC7660, OSEP Grant H325D100034A, and the Vanderbilt Kennedy Center."
12	Please provide publisher details with location for the reference "Webb and Shavelson 2005."
13	Please confirm that you have sufficiently reviewed your proof and queries, and that you understand this is your FINAL opportunity to review your article before publication.

Hammill Institute on Disabilities

Topics in Early Childhood Special Education
1–9

DOI: 10.1177/0271121414528052
tecse.sagepub.com
SAGE

Measuring Representative Communication in Young Children With Developmental Delay

Micheal Sandbank, MEd[1] and Paul Yoder, PhD[1] [AQ: 1][AQ: 2]

Abstract

Generalizability and decision studies provide a mathematical framework for quantifying the stability of a given number of measurements. This approach is especially relevant to the task of obtaining a representative measure of communicative behavior in young children and supports an alternative to the debate regarding which type of assessment yields the most representative scores. The current article provides a report of a generalizability and decision study on 63 toddlers with developmental delay who were treated for 6 months using an intervention that targeted communication and vocabulary goals. Two variables—rate of intentional communication acts and rate of different words—were measured across three assessment contexts at four communication sampling periods. Results verified that measurement stability increased with time and development for both variables, regardless of the type of assessment procedure used.

Keywords

generalizability studies, decision studies, children, developmental delay, communication

The concept of representativeness as it applies to the measurement of young children's behavior is prerequisite to systematic research of early child development. Researchers desire that a "test" or "sample of the behavior indicating an ability or characteristic" is representative of the "way the child usually behaves" (Neisworth & Bagnato, 2004). The "way the child usually behaves" is very difficult to quantify at an individual level (i.e., within an individual child). However, by using group design logic, we can ask whether a group of children are similarly ranked on an ability, regardless of measurement context (i.e., stable across measurement context). If they are, then we can say that one session is "representative" of the others. It is this concept of representativeness that we are speaking to in this article.

One way professionals have attempted to address the goal of measuring representative performance is to argue that particular types of measurement are more likely to yield representative scores than others. One problem with this approach is that there is no consensus on which type is more likely to yield the most representative scores. For example, Neisworth and Bagnato (2004) suggested that the most representative context for assessing young children is a less structured interaction that occurs with a familiar person in a familiar context. Those who take this perspective point to research showing that children with autism, children with Down syndrome, and typically developing children, have been shown to preferentially direct social behavior toward a caregiver rather than a stranger during play interaction (Dissanayake & Crossley, 1996; Mundy & Sigman, 1989). However, higher scores do not mean an assessment is more representative or stable across contexts. Alternatively, arguments can be made that procedures carried out by a skillful professional might elicit key behaviors, yielding optimally valid scores. Yet another perspective argues that structured procedures provide a standard number of opportunities for performance of the behaviors in question, thus variation in scores will better reflect child abilities rather than the number of evocative opportunities for key behaviors. Measurement error resultant from variation in the familiarity or skill of the adult interaction partner, the nature of the environment and materials, and the number of opportunities for performance of the key behavior could all reduce the probability of obtaining scores that are stable across contexts. These different perspectives illustrate the probable fact that scientists will have difficulty agreeing on the representativeness of a single measurement context.

This article will discuss an alternative way to think about representative measurement. Instead of debating which type of measurement context will yield the most representative

[1]Vanderbilt University, Nashville, TN, USA

Corresponding Author:
Micheal Sandbank, Department of Special Education, Vanderbilt University, 110 Magnolia Circle, Office 414, Nashville, TN 37203, USA.
Email: micheal.p.gmaz@vanderbilt.edu [AQ: 3]

scores, we argue that the practice of (a) collecting multiple types of relevant measurement samples and (b) averaging the scores from these samples will produce more representative scores than those from any single type of session. Although this approach has been recommended, the need for it has not been empirically demonstrated, to our knowledge. In addition, it is likely that the need for this expensive approach to assessment is higher for emerging skills in children with developmental delay (DD). Before we flesh out the logic for this prediction, we must present the theoretical and mathematical framework for this approach.

Classical Test Theory (CTT) suggests that the task of obtaining a representative measure of an ability or characteristic from a behavior sample is inhibited by measurement error, which is resultant from characteristics of the measurement process (Algina & Penfield, 2009). Various attributes of measurement contexts (such as interaction partners, test administrators, and test materials) can introduce variance into participant scores that are not related to between-participant differences. These attributes of measurement, known as facets, are potential sources of measurement error and low reliability (Shavelson & Webb, 1991). Because validity cannot exceed the square of reliability, reliability places a ceiling on the upper bound of validity (Nunnally, 1978). Thus, variable scores with low reliability cannot be as valid as those with higher reliability.

While much attention is given in current research to interrater reliability, the extent to which participants are similarly ranked by scores from different raters or coders, little attention is given to reliability within other aspects of measurement, such as the extent to which participants are similarly ranked by scores from different measurement contexts (i.e., across-context stability). In the current article, we investigated the stability of scores across "communication sampling contexts." We used this term to both specify the abilities we wish to measure (communication) and to broaden the meaning of the term "measurement context" to one that we could address in the current article. In this case, the facet that we refer to as communication sampling context includes several measurement characteristics that can influence observed scores, such as interaction partners, test activities, and test materials. Variability in the influential aspects of communication sampling context can have a profound impact on observed scores, and the resultant error can obscure important between-participant differences (increasing the likelihood of a Type II error). Thus, one goal of the current article is to highlight methods for examining and minimizing the potential error introduced by communication sampling context. These methods arise from Generalizability theory (G theory; Cronbach, Gleser, Nanda, & Rajaratnam, 1972).

G theory posits that the mathematical framework of CTT can be used to estimate the amount of variance contributed by different aspects of measurement. According to G theory, the mean squares (MS) for the various factors in the design provided by a traditional ANOVA can be used to separate "true" person variance (i.e., what we wish to measure) from that which is resultant from each measurement facet (error variation; Webb & Shavelson, 2005). The partitioned variance can then be used to calculate a generalizability (G) coefficient, a type of intraclass correlation (ICC), indicating the level of measurement stability achieved. This is referred to as a generalizability (G) study.

Theoretically, an average of multiple scores across variations of a single facet would improve the estimate of a participant's "true score" by averaging the random error, yielding scores with greater reliability. When the goal is to obtain a representative score of a child's generalized communication skill, the value of across-context score averaging is clear. An average that comprises scores from multiple relevant settings is likely to be more representative than a single score from a single context. Still, questions exist as to the number of measurement contexts across which one must average to obtain a reliable estimate of the characteristic or ability.

Decision (D) studies use the variance estimates provided by G studies to calculate the optimal number of samples across which it is necessary to average to achieve measurement stability (Shavelson, Webb, & Rowley, 1989). The results of D study assume that variations of a single facet are equivalently valid and only provide information regarding the number of estimates needed. For example, the results of a D study may suggest that stability can be achieved by averaging the score estimates across four communication sampling contexts, but they do not imply that any single type of communication sampling context might yield optimally representative score estimates.

Developmental theory suggests that the level of stability obtained across measurement contexts might vary as a function of a child's developmental stage. Infant use of new skills is extremely variable across contexts, while more mastered skills are reliably executed across contexts (McCormick & Nellis, 2004**[AQ: 4]**). Consequently, a population of children beginning to acquire a skill will have lower stability from a single measurement than a population of children that has demonstrated growth in that skill. Given this prediction and the assertions of G theory, it would seem that the number of sessions across which one would need to average to obtain stable scores would be greater when a group of children are first learning communication skills, rather than later, when they are more proficient.

The present investigation attempts to address questions related to the above-stated issues as they apply to the measurement of communication competence in initially nonverbal toddlers with DD. We used three types of communication samples to represent variations of the facet "communication sampling context": (a) structured interaction with an unfamiliar examiner, (b) unstructured interaction with an unfamiliar

examiner, and (c) unstructured interaction with a parent. These were conducted at each of four time points throughout the study (at entry, and then 3 months, 6 months, and 9 months following entry). Two variables were measured: (a) rate of intentional communication acts and (b) rate of different words. G and D studies were conducted to estimate the number of communication samples needed to achieve a target level of across-context stability as time and development progressed. The research question was as follows[AQ: 5]:

> **Research Question 1:** Does the number of communication sampling contexts needed to produce a stable estimate of communication ability in young children with DD decrease with development?"

We expected that, as time and language acquisition progressed, scores from a single communication sampling context would exhibit increasing stability, and consequently less contexts would be needed to obtain a representative average score. This prediction was made because we could reasonably expect (and test) that participants would exhibit development in intentional communication and lexical diversity over the course of 9 months.

Method

Design

We used a longitudinal correlational design to conduct a post hoc analysis on data from an existing study, which investigated the effects of Milieu Communication Teaching (MCT; Fey, Warren, Fairchild, Sokol, & Yoder, 2006) as a communication and language intervention for young children with DD (Fey, Warren, Yoder, & Bredin-Oja, 2013). For the purposes of the intervention study, all participants were randomized to one of two treatment groups receiving MCT. The low-intensity group (LI) received one 60-min session per week of MCT for the length of the study, while the high-intensity group (HI) received five 1-hr sessions per week of the same treatment package. As will be shown in the preliminary analyses, group membership did not affect stability estimates. Thus, details of the treatments are not covered in this report. The fact that all children received communication treatment was relevant to the expectation that, as a group, children would exhibit development in intentional communication and lexical diversity skills.

Participants

The study was conducted at Vanderbilt University in Nashville, Tennessee, and the University of Kansas Medical Center in Kansas City, Kansas. Prior to the study, advertisements recruited children that were aged 18 to 27 months, from English-speaking homes, with general DD, significant delay in the acquisition of words, and no diagnosis of autism. Inclusion criteria for the larger study of initially low-verbal children with DD were as follows: (a) 20 or fewer spoken or signed content words in the expressive lexicon, as reported by the primary caregiver on the MacArthur–Bates Communicative Disorders Inventory Words and Gestures (MB-CDI; Feldman et al., 2000); (b) a Bayley Cognitive Composite standard score (Bayley-III; Bayley, 2006) between 55 and 75; (c) a score of 2.75 or lower on the Screening Tool for Autism in Two-year-olds (STAT; Stone, Coonrod, & Oosley, 2000), indicating low risk of autism; (d) normal hearing in at least one ear, as determined by a hearing screening; and (e) motor skills sufficient to sit unsupported and engage in play with an interventionist. A total of 251 children were assessed, 76 were identified and enrolled, and 63 completed all scheduled assessment sessions. Thus, a total of 63 toddlers were included. Caregivers were biological parents, except for three cases in which the caregiver was an adoptive parent. The Parenting Stress Index/Short Form, Third Edition (PSI/SF; Abidin, 1995[AQ: 7]) indicated that eight parents exhibited clinically significant stress (i.e., raw score > 90). Participant and parent characteristics at study onset are presented in Table 1.

Table 1. Participant and Parent Characteristics.[AQ: 6].

Measure	*M* (*SD*)
Age in months	22.36 (3.05)
% male	38
Bayley-III Cognitive Composite standard score	65.40 (6.74)
Bayley-III age equivalency	12.54 (2.09)
No. of toys played with in DPA	8.84 (2.31)
MCDI expressive vocabulary	1.06 (1.44)
MCDI receptive vocabulary	59.51 (65.31)
Primary caregiver education level[a]	7.01 (1.21)
PSI/SF total raw score[b]	69.44 (16.20)

Note. Bayley-III = Bayley Scales of Infant and Toddler Development, Third Edition. DPA = Developmental Play Assessment; MCDI = MacArthur Communicative Development Inventory; PSI/SF = Parenting Stress Index/Short Form.
[a]A scale score of 7 means 3 to 4 years of college education. [b]PSI raw score mean is based on the scores from 53 families.

Measures

Dependent variables were coded from three separate communication samples that were intended to reflect three assessment formats: (a) structured interaction with an unfamiliar examiner, (b) unstructured interaction with an unfamiliar examiner, and (c) unstructured interaction with a familiar examiner. We used a standardized communication assessment, the Communication and Symbolic Behavior Scales (CSBS; Wetherby & Prizant, 2002), which was

administered by a trained speech therapist, to serve as the structured interaction with an unfamiliar examiner. A lab-created assessment administered by a trained speech therapist, which we termed the examiner–child semi-structured play session (ECSS), reflected an unstructured assessment context with an unfamiliar examiner. Finally, a parent–child free play (PCFP) assessment, during which parents were provided materials and left to play with their child as they typically might, served as the unstructured interaction with the familiar partner.

CSBS. The CSBS is a standardized tool that is used to evaluate the language and social communication abilities of children between the developmental ages of 6 months and 2 years (Wetherby & Prizant, 2002). The assessment structure prescribes a series of clinician-led interactions that are designed to elicit communication from a small child. The exact sequence of the interaction contexts was 5 to 10 min of warm-up, 10 to 20 min of "communicative temptations," and 5 min of book sharing. During warm-up, the child was given a chance to acclimate to the professional and the environment. Neither structured language elicitation occurred during warm-up nor were child behaviors coded. The communicative temptations context involved the systematic presentation of several high-interest stimuli that were designed to entice child-initiated communication acts. Stimuli were presented in the following order: (a) wind-up toy, (b) balloon, (c) bubbles, (d) peek-a-boo, (e) puppet, (f) blocks in a clear box, (g) cereal in a jar, and (h) toys in a clear bag. Each stimulus was presented for 1 to 2 min, and communicative behaviors were met with expanding prompts. The final 5 min of the session were devoted to book sharing, during which the examiner introduced one of three age-appropriate books and allowed the child to explore the book without prompting. The same three books were used during every CSBS assessment. During the book-sharing portion of the assessment, examiners refrained from asking the child to label pictures or turn pages, and instead followed the child's lead by acknowledging and commenting on what the child looked at, pointed toward, or verbally labeled.

ECSS. In the ECSS, the examiner introduced one of three toy sets at a time and switched to a new set if a child exhibited disinterest or a desire to play with different toys. For example, if the child repeatedly threw toys off the table, or exhibited distress and needed to be redirected, the examiner would remove all toys from the table and offer a choice of the remaining two sets. The same three sets of toys were available at each assessment period. Each set of toys was stored in a clear plastic bin, so that children could see and request new sets. Included toy sets were miniature settings designed for groups of small toy people. These toy sets featured a schoolhouse, a set of carnival rides (i.e., Ferris wheel, swing, rocket ship ride), and a playground. During this communication sample procedure, examiners were instructed to limit their scaffolding of children's communication and play behaviors, and were not permitted to directly prompt communicative attempts. For example, examiners were not permitted to use time delays to prompt for behavior regulators, physically prompt specific gestures, make explicit requests for communication (e.g., "Tell me"; "Show me"), or withhold toys to prompt the child to exhibit an intentional communication act. Instead, examiners responded naturally to any child communication behaviors by interpreting, acknowledging, or expanding on the topic of communication. Examiners were permitted to model appropriate use and higher levels of play with toys. Examiners were also permitted to ask questions that were relevant to child play behaviors (e.g., "Where's he going?"; "What's she doing?"; "What is that?"). Examiners were also allowed to introduce new toys to an unengaged child by labeling the new toy and handing it to the child. The total duration of each ECSS assessment was 15 min.

PCFP. The PCFP assessment refers to a loosely structured communication sample, which involves a child participant and a caregiver (almost always the mother). Caregivers were provided with two toy sets and given the opportunity to "play" independently with their children for 10 min. Included toy sets featured a small barn with corresponding animal and farmer toys, and a small house with corresponding furniture and people toys. Researchers then returned and provided three age-appropriate books for parents and children to "look at." Parents were not instructed to read the books to their children. The book-sharing portion of the sample lasted 5 min. The toy sets and books available to parents during this assessment were the same across all four time periods.

Dependent variables. Two variables, rate of intentional communication acts and rate of different words, were coded from all three communication samples every 3 months for 9 months (total four time periods, 12 communication samples), using a highly detailed coding manual. A copy of the coding manual is available upon request from the second author.

Intentional communication acts were defined as any of the following: (a) words, (b) conventional signs, (c) conventional gestures coupled with attention toward an adult (e.g., distal point, "shh" gesture, wave, reach, etc.), (d) unconventional gestures that intrinsically show coordinated attention to object and person (i.e., upturned palm, giving an object to an adult, showing an object to an adult), or (e) nonword vocalizations accompanied by additional evidence of coordinated attention to object and person. Attention to an adult was defined as gaze to an adult's face, immediate and precise answering of an adult question, or exact verbal

imitation of an adult. Coordinated attention to object and person was defined as showing attention to an object and person within 3 s (either sequentially or simultaneously). This definition excluded attention shifts between person and object that were caused by adults (e.g., instances in which the examiner held toys next to their face and the child then gazed at the examiner's face).

To be coded as words, vocalizations were required to (a) be found in an unabridged English dictionary; (b) contain a vowel nucleus that was functionally equivalent to that of the perceived word; (c) contain one or more consonants of the perceived word, or child-like substitutions of those consonants (acceptable substitutions were prespecified); and (d) be used in a semantically and pragmatically conventional manner, as determined by the coder. Variations of the root word (e.g., books) were not coded separately from the root (e.g., book).

Interobserver Reliability

Communication samples were coded by teams of trained coders using a video-coding software (PROCODER; Tapp & Walden, 1993). Use of this software allowed coders to splice communication sample videos into short segments, highlight segments that contained relevant behaviors, and review these segments in real and slowed time. Each team consisted of one primary coder and one reliability coder. Prior to study onset, teams coded a series of training communication samples until they reached a minimum level of gross agreement (.85) for three consecutive samples. During training, coding teams resolved discrepancies on a point-by-point basis, with guidance from an extremely detailed 50-page coding manual.

During the study, one of every five communication samples was randomly selected and independently coded for reliability by the second coder. The primary coder was not informed about which sessions would be coded for reliability purposes. While G theory can be used to estimate stability achieved across different coders, that procedure requires that two separate coders score each participant in all three contexts. As only a fifth of all communication samples were coded for reliability, we were not able to separately estimate error variance due to coders from error variance due to context in this generalizability investigation. Consequently, we have documented rater score stability by computing the absolute ICC for reliability estimates, using the SPSS Version 20 reliability scale procedure. The average ICCs, estimating interobserver reliability across all four time points, were .90 for CSBS agreement samples, .89 for ECSS agreement samples, and .92 for PCFP agreement samples. The high ICCs for each communication sampling context suggest that raters introduced minimal error variance, if any at all, to our observed scores. All scores used in the generalizability study were taken from the primary coder.

Generalizability Study

Central to G theory and G studies is the concept of the universe of all admissible observations. This universe is characterized by the sources of variance in the observed scores: the facet of measurement (in this case, communication sampling context) and the facet of differentiation (i.e., what we want to measure variation among; in this case, person). Given the MS provided by the ANOVA for each facet and their corresponding interactions, we can estimate variance components for Person, Context, and the Person × Context interaction. Using these variance component estimates, we may then calculate the absolute level of stability (the absolute G coefficient) achieved by averaging across one or more communication sampling contexts, for each successive time period. Formulas detailing the method for calculating the absolute G coefficient (Cardinet, Johnson, & Pini, 2012), as well as an example of this calculation, are provided in Table 2.

EduG

We used the statistical program EduG (Swiss Society for Research in Education Working Group, 2012) to compute the G coefficients for each dependent variable across person and context. EduG is a free downloadable software created exclusively for conducting generalizability analyses with ease. For each time period, individual's observed scores from each of the three communication samples composed the 189 observations (63 persons × 3 contexts) in a crossed design (Person × Context). The program calculated the variance components needed to conduct G and D studies, treating the facets of person and context as random, because we wished to generalize results to the entire population of observational measurement contexts, for the entire population of young children with DD.

Results

Preliminary Analyses

To ensure that group membership did not affect stability estimates, we conducted a mixed-factor ANOVA with time and context as two within-subjects factors and group membership as a between-subjects factor, and examined the interaction of the group and context factors. After correcting the degrees of freedom for within-subjects comparisons for deviance from sphericity (Greenhouse–Geisser), neither the Group × Context interaction nor the Group × Context × Time interaction terms were significant, for either dependent variable. These results led us to pool participants across groups for further analysis.

As different measurement periods were analyzed as proxies for change in development, it was necessary to test

Table 2. Calculation of Absolute G Coefficient for the Rate of Different Words at Time 1 With Three Assessment Contexts.

Given:
$MS_p = 2.232$, $n_p = 63$
$MS_c = 1.1714$, $n_c = 3$
$MS_{p\times c} = 0.706$

Variance component	Formula	Calculation
Instrumentation variance	$\sigma_c = \frac{MS_c - MS_{p\times c}}{n_p}$	$\sigma_c = \frac{1.174 - 0.706}{63} = 0.00743$
Relative variance	$\sigma_{rel} = \frac{MS_{p\times c}}{n_c}$	$\sigma_{rel} = \frac{0.706}{3} = 0.23533$
Absolute variance	$\sigma_{abs} = \sigma_{rel} + \frac{\sigma_c}{n_c}$	$\sigma_{abs} = 0.23533 + \frac{0.00743}{3} = 0.23781$
True person variance	$\sigma_p = \frac{MS_p - MS_{p\times c}}{n_c}$	$\sigma_p = \frac{2.232 - 0.706}{3} = 0.50867$
Absolute G coefficient	$G_{abs} = \frac{\sigma_p}{\sigma_p + \sigma_{abs.err}}$	$G_{abs} = \frac{0.50867}{0.50867 + 0.23781} = 0.68142$

Note. MS_c = mean squares for context facet; $MS_{p\times c}$ = mean squares for interaction; n_p = number of persons; n_c = number of contexts; MS_p = mean squares for person facet.

Table 3. Means and Standard Deviations for Both Outcomes Within Each Communication Sampling Context.

		RICN	RDW
Measure	Time	*M* (*SD*)	*M* (*SD*)
CSBS	1	18.64 (10.73)	0.84 (2.01)
	2	35 (29.79)	3.67 (7.69)
	3	46 (31.70)	6.14 (10.65)
	4	57.86 (34.78)	10.68 (15.1)
ECSS	1	7.23 (8.29)	0.45 (1.34)
	2	13.17 (13.51)	2.52 (5.09)
	3	20.14 (18.82)	3.80 (6.64)
	4	28.81 (23.51)	7.81 (11.37)
PCFP	1	10.23 (10.07)	0.91 (1.95)
	2	22.95 (20.89)	4.02 (7.28)
	3	33.91 (31.25)	8.02 (13.71)
	4	43.86 (34.01)	11.52 (17.18)

Note. RICN = rate of intentional communication; RDW = rate of different words; CSBS = Communication and Symbolic Behavior Scales; ECSS = examiner–child semi-structured play session; PCFP = parent–child free play.

whether participants exhibited growth on both dependent variables. Because the data did not satisfy assumptions of normality, we tested within-subject growth in each context by comparing Time 1 and Time 4 scores, and assessed the significance of each mean difference using *p* values taken from bootstrapped distributions. For the rate of intentional communication, results indicate highly significant growth for scores from the CSBS, $d = 1.27$, $p = .001$; from the ECSS, $d = 1.13$, $p = .001$; and from the PCFP, $d = 1.10$, $p = .002$. For the rate of different words, results indicate highly significant growth for scores from the CSBS, $d = 0.80$, $p = .001$; from the ECSS, $d = 0.87$, $p = .001$; and from the PCFP, $d = 0.61$, $p = .003$. Means and standard deviations across measurement periods are presented for each dependent variable in Table 3. Thus, growth occurred on the variables of interest, regardless of communication sampling context.

Research Question 2: Does the number of communication sampling contexts needed to produce a stable estimate of communication ability decrease with time?

We conducted generalizability and decision studies using data from the assessments of the two dependent variables (rate of intentional communication acts and rate of different words) across three communication samples at each of four time periods. Results of the generalizability studies indicated the levels of reliability achieved for each time period (1–4) with communication samples from all three communication sampling contexts. In addition, the generalizability studies provided stability estimates for scores obtained from any single communication sampling context. Results of the decision studies indicated the projected stability level that could be achieved by averaging scores from additional communication sampling contexts. As indicated in the introduction, these assume equivalent validity of contexts and simply provide the number of score estimations (from any relevant context) across which it is necessary to average to achieve a robust level of scores

Table 4. Component Variance Percentages for Rate of Intentional Communication.

			Time 1		Time 2		Time 3		Time 4	
Variable	Source	*df*	MS	% Var	MS	% Var	MS	% Var	MS	% Var
RICN	P	63	0.71	45.40	4.02	61.9	7.30	66.50	9.64	72.40
	C	2	2.81	10.50	13.61	10.8	19.60	9.00	23.44	8.80
	P × C	126	0.17	44.10	0.51	27.3	0.80	24.50	0.77	18.80
RDW	P	63	2.23	41.40	43.46	77.1	111.87	72.10	228.83	83.90
	C	2	1.71	1.30	20.87	1.60	144.22	4.50	126.21	2.10
	P × C	126	0.71	57.40	3.67	21.3	10.93	1.30	12.01	13.90

Note. Component variance percentages are estimated using scores from the first measurement at each time period. MS = mean squares; % Var = percentage of variance attributed by source; RICN = rate of intentional communication; P = person, C = context; P × C = person by context; RDW = rate of different words.

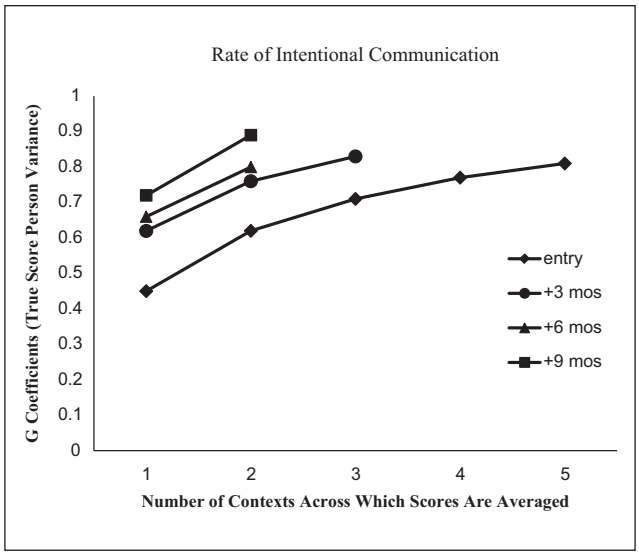

Figure 1. Generalizability coefficients for the rate of intentional communication across collected and projected communication samples for each assessment period.

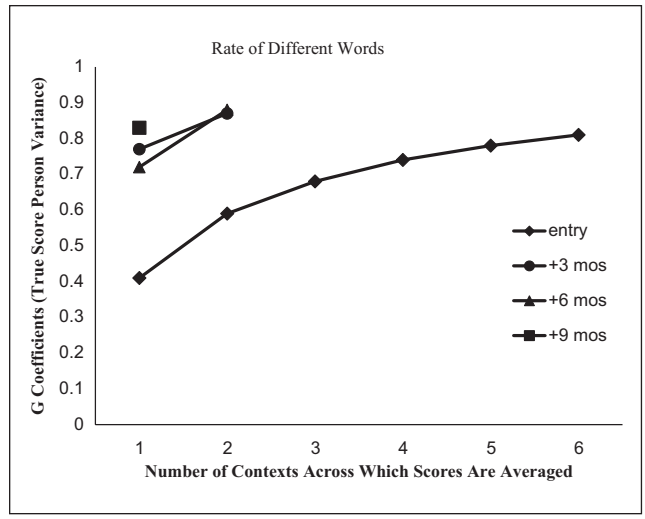

Figure 2. Generalizability coefficients for the rate of different words across collected and projected communication samples for each assessment period.

stability (.8). Table 4 presents MS and variance components for each outcome. Figures 1 and 2 present achieved and projected stability estimates at each time point for the rate of intentional communication and the rate of different words, respectively.

Rate of intentional communication. Figure 1 displays the results for the rate of intentional communication. The number of communication samples needed to obtain a G coefficient of 0.8 decreased with time and development, such that scores needed to be averaged across five communication samples at Time 1, across three at Time 2, across two at Time 3, and across two at Time 4, to achieve a stability level of at least .8. Results confirm the hypothesis that the number of communication samples needed to produce a stable estimate of the rate of intentional communication decreased with time and development.

Rate of different words. Figure 2 displays the results for the average number of different words used per communication sample. Projections from the decision study suggest that scores taken from an additional three communication samples (of any type) were necessary to achieve the criterion stability of 0.8 at Time 1. This number decreased over time. At 3 months and 6 months after entry, only two communication samples were necessary to achieve a G coefficient of 0.8. Nine months after entry, only a single communication sample was needed to achieve criterion stability. Results are consistent with the hypothesis that the number of communication samples needed to produce a stable estimate of the rate of different words decrease with time and development.

Discussion

In this study, we demonstrated that growth occurred over time on the dependent variables and then estimated the

number of communication samples across which one would need to average scores to achieve a criterion level of stability, for each time period. Results showed that (a) scores of the dependent variables increased over time and (b) this occurred across all three contexts, illustrating that this increase was not bound to a specific context. In addition, results indicated that the number of needed communication sample scores across which it was necessary to average to obtain a stable estimate reduced as time transpired. We used this combination of information to infer that as children develop, one needs to assess communication in fewer communication sampling contexts to derive a stable estimate of their generalized ability to communicate. These results support the notion that stability of performance across contexts increases with skill proficiency.

Strengths

The current study had several strengths. First, the authors limited an obvious source of error-rater disagreement by verifying that interobserver reliability (IOA[AQ: 8]) was high for all three communication sampling contexts. Second, a relatively large sample size was used, which, in turn, narrows the confidence interval around the estimates of number of sessions needed to produce a stable estimate of communication variation among participants. Third, the use of a longitudinal design with four measurement periods allows replication of the pattern of increasing stability with development of the communication skills in the same group of children. Fourth, the selection of preschoolers with DD increased the relevance of the findings to practitioners and researchers serving preschoolers with DD.

Limitations

Despite several strengths, the current investigation did have some limitations. First, we were unable to separately estimate the effects of development from those of repeated test exposure on measurement stability. Second, because this was a post hoc analysis, other potentially relevant conditions of communication sampling were excluded from our investigation. Third, the contexts included in this study serve as a sample of variations of a single measurement type: direct observation communication samples of young children. Because we have only sampled from the population of direct observation communication samples, we are limited in our ability to generalize results to other types of communication assessment for children, such as parent report.

Implications for Researchers

That skill proficiency can differentially affect measurement stability in young children should be of interest to early childhood special education researchers because their investigations often involve the measurement of new and emerging skills. Investigators that fail to minimize error introduced by measurement facets beyond rater are less likely to detect important relationships. Just as investigators often use a preemptive power analysis to determine appropriate sample size, investigators should consider doing preemptive decision studies using data from previous studies to determine the number of assessment variations that will be necessary to average across to obtain a representative measure. The ease of this estimation, especially with the help of free modern computing software, compels researchers to use this process to understand the influence of the various facets that effect behavioral measurement.

While it is appropriate to hypothesize, based on these results, that the stability of any type of communication assessment will increase with age and skill acquisition, it is not appropriate to directly generalize the recommended number of communication sampling contexts to a form of communication assessment other than direct observation communication samples similar to ones sampled for children similar to those studied. The specific number of assessments reported in the decision studies is less informative than the general pattern that is conveyed. This pattern suggests that more measurements are needed to achieve a reliable estimate of ability at the beginning of skill acquisition, while fewer measurements are needed as the skill becomes more fluent.

G theory provides a method for examining the amount of error variance introduced by different measurement system characteristics (G studies), as well as a method for identifying the number of assessments needed to minimize that error (D studies). Carefully designed G studies allow investigators to examine sources of measurement error with great flexibility. Future research could specifically investigate the extent to which interaction partner, interaction, style, materials each introduce measurement error by administering several assessments in which only one measurement characteristic varies. Graham, Sandbank, and Lane (2014) demonstrated this approach in an investigation of score stability of writing samples. Two types of administrators (researchers and teachers) administered two types of writing prompts each to a sample of students. Each writing sample was scored by two raters, yielding a total of 8 scores for each participant (2 administrators × 2 sample types × 2 raters). The careful manipulation of each facet allowed investigators to separately examine error variance introduced by raters, administrators, and writing sample type.

Implications for Providers

Given the results of this investigation, providers might consider administering multiple assessments across several relevant contexts when attempting to profile a child's

developmental skill. This is especially important for cases in which these skills are newly emerging. Practitioners often favor measures that yield higher scores, believing that these measures represent a child's "true ability" and attribute lower scores to uncontrolled contextual variables. Instead, providers should consider giving equal weighting to scores from multiple assessments, whether high or low. Both high and low scores can serve as a starting place for constructing goals to expand the use of emerging skills across multiple contexts. The communication sampling contexts in which a child has his highest scores provide insight into the contextual variables (e.g., familiarity and structure) that might be scaffolding the child's use of newly developing behavior. The communication sampling contexts in which a child has his lowest scores provide insight into the contexts in which far-transfer and mastery might be tested. Thus, all of these scores should be considered together to form a holistic understanding of a child's developmental level.

Acknowledgment

We would like to acknowledge the contributions of Marc Fey and Steve Warren.

Authors' Note

The positions stated here are not necessarily the positions of National Institutes of Health (NIH), Office of Special Education Programs (OSEP), or the Vanderbilt Kennedy Center.[AQ: 9]

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.[AQ: 10]

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported in part by NIDCD Grant RO15DC7660, OSEP Grant H325D100034A, and the Vanderbilt Kennedy Center.[AQ: 11]

References

Algina, J., & Penfield, R. (2009). Classical Test Theory. In R. Millsap & A. Maydeu-Olivares (Eds.), *The SAGE handbook of quantitative methods in psychology* (pp. 93–122.). London, England: SAGE.

Bayley, N. (2006). *Bayley Scales of Infant and Toddler Development* (3rd ed.). San Antonio, TX: The Psychological Corporation.

Cardinet, J., Johnson, S., & Pini, G. (2012). *Applying generalizability theory using EduG*. New York, NY: Routledge Academic.

Cronbach, L., Gleser, G., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability of scores and profiles*. New York, NY: Wiley.

Dissanayake, C., & Crossley, S. A. (1996). Proximity and sociable behaviors in autism: Evidence for attachment. *Journal of Child Psychology and Psychiatry*, *37*, 149–156.

Feldman, H., Dollaghan, C., Campbell, T., Kurs-Lasky, M., Janosky, J., & Paradise, J. (2000). Measurement properties of the MacArthur Communicative Development Inventories at ages one and two years. *Child Development*, *71*, 310–322.

Fey, M., Warren, S., Fairchild, M., Sokol, S., & Yoder, P. (2006). Early effects of responsivity education/prelinguistic milieu teaching for children with developmental delays and their parents. *Journal of Speech, Language, and Hearing Research*, *49*, 526–547.

Fey, M., Warren, S., Yoder, P., & Bredin-Oja, S. (2013). Is more better? Milieu Communication Teaching in toddlers with intellectual disabilities. *Journal of Speech, Language, and Hearing Research*, *56*, 679–693.

Graham, S., Sandbank, M., & Lane, K. (2014). *Examining the stability of writing assessments for students with writing difficulties*. Manuscript in preparation.

Mundy, P., & Sigman, M. (1989). The theoretical implications of joint-attention deficits in autism. *Development and Psychopathology*, *1*, 173–183.

Neisworth, J. T., & Bagnato, S. J. (2004). The mismeasure of young children: The authentic assessment alternative. *Infants & Young Children*, *17*, 198–212.

Nunnally, J. C. (1978). *Psychometric theory*. New York, NY: McGraw-Hill.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: SAGE.

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist*, *44*, 922–932.

Stone, W., Coonrad, E., & Oosley, O. (2000). Brief report: Screening Tool for Autism in Two-year olds (STAT): Development and preliminary data. *Journal of Autism and Developmental Disorders*, *30*, 607–612.

Swiss Society for Research in Education Working Group. (2012). *EduG* (Version 6.1) [Computer software]. Unpublished instrument. Retrieved from http://www.irdp.ch/edumetrie/englishprogram.htm

Tapp, J., & Walden, T. A. (1993). PROCODER: A professional tape control, coding, and analysis system for behavioral research using videotape. *Behavior Research Methods, Instruments and Computers*, *25*, 53–56.

Webb, N. M., & Shavelson, R. J. (2005). Generalizability theory: Overview. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (Vol. 2, pp. 717–719).[AQ: 12]

Wetherby, A. M., & Prizant, B. M. (2002). *Communication and Symbolic Behavior Scales: Developmental profile*. Baltimore, MD: Paul H. Brookes.[AQ: 13]