AMERICAN JOURNAL ON INTELLECTUAL AND DEVELOPMENTAL DISABILITIES
2014, Vol. 119, No. 6, 589–605

DOI: 10.1352/1944-7558-119.6.589

Generalizability and Decision Studies to Inform Observational and Experimental Research in Classroom Settings

Kristen Bottema-Beutel, Blair Lloyd, Erik W. Carter, and Jennifer M. Asmus

Abstract

Attaining reliable estimates of observational measures can be challenging in school and classroom settings, as behavior can be influenced by multiple contextual factors. Generalizability (G) studies can enable researchers to estimate the reliability of observational data, and decision (D) studies can inform how many observation sessions are necessary to achieve a criterion level of reliability. We conducted G and D studies using observational data from a randomized control trial focusing on social and academic participation of students with severe disabilities in inclusive secondary classrooms. Results highlight the importance of anchoring observational decisions to reliability estimates from existing or pilot data sets. We outline steps for conducting G and D studies and address options when reliability estimates are lower than desired.

Key Words: *observational measurement; severe disabilities; social interaction; secondary classrooms; inclusion*

As calls to increase inclusion within general education classrooms intensify (Jackson, Ryndak, & Wehmeyer, 2008/2009), the social and academic participation of students with severe disabilities within these classrooms continues to be of central interest to practitioners and researchers alike. The extent to which students with autism and/or intellectual disabilities interact with peers and are engaged in relevant learning activities represent important indicators of the effectiveness of educational interventions designed to support inclusive classroom participation (Carter, Sisco, Brown, Brickham, & Al-Khabbaz, 2008). Researchers often conceptualize inclusion in terms of students' social and/or academic behavior within the classroom and, thus, design measurement procedures to record behavior in these everyday settings (e.g., Carter et al., 2008; Chung, Carter, & Sisco, 2012; Linnenbrink-Garcia, Rogat, & Koskey, 2011; Pasco, Gordon, Howlin, & Charman, 2008). However, the social and academic behavior of students can vary substantially from day to day due to a variety of contextual factors, including the instructional practices used by teachers, student grouping arrangements, and classroom activities (Carter et al., 2008). Reliably measuring social interactions can be especially complex, as social opportunities are highly influenced by context. This variability can be challenging for experimental researchers interested in understanding the extent of behavior change associated with their interventions. Researchers strive to minimize or account for variation due to factors other than the construct of interest to accurately detect intervention effects. Variation of this type is defined as measurement error, and includes any aspect of the measurement system that introduces variation into scores *not* due to the construct of interest.

Reliability in measurement has long been an important aspect of group-design research (Cronbach, Rajaratnam, & Gleser, 1963). In this article, we specifically focus on reliability in observational measurement, which is the consistency of observed scores for a given person under a given set of conditions. The general principles we describe regarding reliability, however, may be applied to any measurement procedure. Researchers commonly report statistics quantifying the extent to

DOI: 10.1352/1944-7558-119.6.589

which independent observers agree with one another in their application of the measurement system, as maximizing interobserver agreement reduces the error produced in a measurement system and increases the reliability of scores. However, reliability also is affected by sources of variation other than the observers. For example, in educational settings, instructional format and student groupings often vary across observation sessions (McWilliam & Ware, 1994). Low reliability increases the probability of accepting the null hypothesis when it is false, or concluding an intervention is not effective when it is actually having an impact (i.e., Type II error; Cohen, Cohen, West, & Aiken, 2003). Therefore, researchers interested in examining intervention effectiveness should attend to whether their observational measurement system is able to produce reliable estimates that minimize measurement error and allow for the detection of intervention effects.

Educational researchers also want to know how outcome measures of interest are affected in the contexts most salient to students. For example, within schools, social interactions and academic engagement occur in classroom contexts where students spend the majority of their day. Classroom observations allow for direct measurement of how peer interactions and classroom engagement unfolds in settings and with communication partners who represent a typical school day experience and, thus, have stronger ecological validity (i.e., the extent to which a measurement context resembles naturally occurring contexts; Brooks & Baumeister, 1977). Yoder and Symons (2010) describe a tension between the ecological validity and the structuredness (i.e., degree to which influential variables are held constant across sessions or participants) of observational measurement procedures. Observations in classrooms are typically unstructured such that contextual variables associated with the student outcome of interest (e.g., social interactions) are not held constant by researchers and instead are allowed to vary naturally across sessions. This results in natural variation simply due to differences in contextual factors, such as opportunity, rather than changes in the construct of interest. Although classrooms are an ecologically valid observational setting for special education research, the contextual differences from one observation session to the next makes it unlikely that a single measurement opportunity (e.g., one class period or a single 10-min observation period) will yield a reliable index of social or academic behavior (Yoder & Symons, 2010). In practical terms, this means researchers will need to plan for multiple sessions and compute an average score across these sessions to obtain reliable estimates of the outcome measure. Thus, determining the requisite number of observation sessions that will provide reliable estimates becomes an essential empirical question. In this article, we describe generalizability (G) and decision (D) studies as a valuable approach for answering this question within special education research.

G Theory

G theory was originally introduced in the 1950s as a statistical approach used to evaluate the reliability or stability of behavioral measurements (both observational and nonobservational measurements alike; Gleser, Cronbach, & Rajaratnam, 1965; Shavelson & Webb, 2006). G theory offers a procedure for partitioning error variance into contributions from the construct being measured and those from different features of the measurement system, termed *measurement facets*. In G theory, measurement facets are analogous to factors in analysis of variance (ANOVA) terminology (Cardinet, Johnson, & Pini, 2010). In most educational studies, and in the study we present, the focus of measurement is a person (Brennan, 1992). A *true score* is the best representation of an individual's score on a given construct (e.g., frequency of peer interactions as a measure of social participation). Theoretically, researchers can estimate a true score by averaging scores across all possible observation sessions within the bounds of admissible sessions. Because it is not feasible or efficient to obtain all possible measures of observable indicators of a person's score on a construct, researchers make an estimate by sampling a subset of scores during selected observations (Gleser et al., 1965). Among-participant variance across these within-person averages is the variance in true score, which is what researchers intend to measure. When using group designs, researchers typically are interested in explaining among-participant variance in true scores using factors or correlates unrelated to measurement facets (e.g., an intervention effect).

When measurement facets account for large amounts of variance among persons, such variance is referred to as "measurement error" or

DOI: 10.1352/1944-7558-119.6.589

"error variance." Estimating different sources of error variance allows researchers to better understand how the features of their measurement system contribute to the deviation in observed scores from the true score. They may then use information about the sources of error variance to make decisions about how to decrease the amount of error associated with different measurement facets in future studies (e.g., increasing the number of observers or number of sessions used to compute scores).

As mentioned previously, facets are characteristics of the measurement system identified by researchers as possible sources of error variance (e.g., observation sessions, observers, test items). If there is more than one facet, researchers must not only consider the main effects for each facet, but also the interactions between measurement facets and the object of measurement (e.g., sessions by persons). The interactions among measurement facets (e.g., sessions by observers) and main effects of measurement facets (e.g., sessions, observers) must be retained in the model to interpret the key interactions of interest. Researchers' characterization of each facet constitutes the universe of admissible sessions. For educational researchers interested in the social interactions of students with disabilities in inclusive classrooms, all classrooms attended by students with disabilities during a given time frame (e.g., a semester) would be the universe of admissible observations for a session facet.

G theory provides a set of mathematical computations for partitioning error among measurement facets. We will introduce these equations first by discussing single-facet measurement systems, in which session is the only facet (for an example of a two-facet G and D study, see Bruckner, Yoder, & McWilliam, 2006). Equation 1 describes how a single score may be conceptualized as the grand mean in the population plus contributions from person and session:

$$\text{Score: } X_{ps} = \mu + \nu_p + \nu_s + \nu_{ps,e} \qquad (1)$$

where μ is the grand mean in the population and universe, ν_p is the effect of person (the object of measurement), ν_s is the effect of session (the measurement facet), and $\nu_{ps,e}$ is the effect of the person × session interaction, which is conflated with the residual component, including any effect not represented by the other components in the model.

The observed score variance may be broken down into independent variance components, which are usually analyzed as random effects. The variance of the score given in Equation 1, across the population of persons and the conditions in the universe of admissible sessions, may be quantified as:

$$\text{Variance: } \sigma^2(X_{ps}) = \sigma^2_p + \sigma^2_s + \sigma^2_{ps,e} \qquad (2)$$

In Equation 2, each subscript denotes the component contributing to the overall variance. That is, σ^2_p represents the variance due to person, σ^2_s represents the variance due to session, and $\sigma^2_{ps,e}$ represents the variance due to the person × session interaction, which is conflated with variance due to the residual component. These variance components are represented in the Venn diagram in Figure 1. Note that the equations are for single person-session combinations, as opposed to average scores across sessions (Brennan, 1992). Averages are considered in D studies, which will be described in a later section.

A second example of a single-facet study is when multiple observers score each participant. Equations for conceptualizing single scores and variance are the same as those shown above, except the subscripts represent observer instead of session (Equations 3 and 4):

$$\text{Score: } X_{ps} = \mu + \nu_p + \nu_o + \nu_{po,e} \qquad (3)$$

$$\text{Variance: } \sigma^2(X_{ps}) = \sigma^2_p + \sigma^2_o + \sigma^2_{po,e} \qquad (4)$$

There is an important distinction between using multiple observers in the context of G and D studies versus in the assessment of interobserver reliability. Because a *g* coefficient is a type of intraclass correlation coefficient (ICC), both applications generate an ICC to provide an index of reliability. G and D studies, however, are conducted with the purpose of determining how many observers are needed to *average across* to compute scores for each participant, whereas interobserver

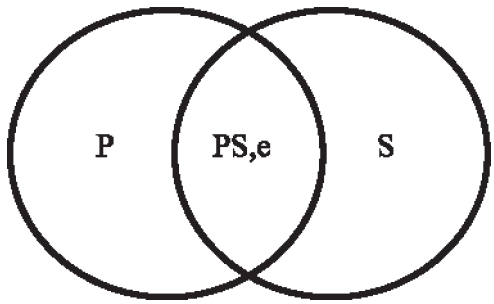

Figure 1. Variance components diagram, where P and S represent Persons and Sessions, respectively.

DOI: 10.1352/1944-7558-119.6.589

reliability procedures are used simply to provide an index of observer reliability, with no implications for how the variable of interest is computed. Additionally, interobserver reliability is usually measured only for a subset of the data, whereas G and D studies provide information on how many observers should score *all* of the participants in the study (with the intention of averaging multiple observers' scores for each participant).

G Studies

The variance components just described may be estimated by conducting a G study. To conduct a G study, researchers gather data from a sample of n_p persons, who are observed in n_s sessions (which may be crossed with n_o observers if observer is also a facet of interest). Different values of n represent the number of levels for each facet. If a study involves three observation sessions per person, the session facet is said to have three levels. If session is the only facet, then the study design is represented as $p \times s$ (read as "person by session"). If observer is included as an additional facet, the study design is represented as $p \times s \times o$. The variance components may be estimated from mean squares derived from a within-subjects ANOVA (see Table 1). Unlike traditional uses of ANOVA, which calculate means and variances to perform tests of statistical significance, G studies do not involve significance tests (Brennan, 1992). A G study generates a g coefficient, which is the proportion of the expected observed score variance due to the best estimate of the true score. In this case, the best estimate of the true score for a given person is the average of scores across the three sessions. If observer were also a facet, the best estimate would be the average of scores across sessions and observers. The value of the g coefficient is an index of the extent to which one can generalize from the collected data to the true score (Gleser et al., 1965). The equation for generating a g coefficient from a single-facet study is as follows:

$$g = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_s^2/n_s + \sigma_{ps,e}^2/n_s} \quad (5)$$

All g coefficients are between 0 and 1, with values closer to 1 indicating higher reliability and thus lower probability of Type II error (e.g., concluding there is no intervention effect when there is an intervention effect). A coefficient of .80 has been used as a criterion value to consider scores reliable (Cardinet et al., 2010). Consensus on this criterion, however, is yet to be reached and some researchers have identified coefficients of .70 and .60 as indicating acceptable reliability (e.g., Berk, 1979; Mitchell, 1979).

D Studies

A D study uses information gathered from a G study to determine the reliability of scores under a variety of study conditions with different levels of each facet. D studies allow researchers to determine the most efficient means to achieve a desired criterion of reliability (e.g., .80) that will allow for an accurate generalization of scores to the universe of scores. Therefore, a G study provides a means of estimating the stability of measurements already made, whereas a D study allows researchers to use information about error contributions to optimize measurement procedures for future studies carried out under similar conditions (Cardinet et al., 2010). Limiting the number of observations can have important implications for conserving personnel resources and monies within large-scale studies. For example, a two-facet D study might indicate that a sufficiently high g coefficient may be achieved with either (a) two

Table 1
Variance Component Equations

Source of variation	Variance component	Equation for deriving variance component from mean squares
Persons (p)	σ^2p	$(MS_p - MS_{ps,e})/ns$
Session (s)	σ^2s	$(MS_s - MS_{ps,e})/np$
Person × session, error (ps,e)	σ^2ps,e	$MS_{ps,e}$

Note. Adapted from Bruckner, Yoder, and McWilliam (2006) and Shavelson and Webb (2006).
MS= mean square; MS_p = mean square person; MS_s = Mean square session; $MS_{ps,e}$ = mean square person × session + error; np= number of persons; ns= number of sessions.

DOI: 10.1352/1944-7558-119.6.589

observers and one observation session, or (b) one observer and three observation sessions. Researchers may then decide which option is more feasible given budgetary and personnel restrictions and other aspects of the study procedures. To generate *g* coefficient values for a given number of sessions, the variance components derived from the ANOVA are used and the desired number of sessions is substituted into Equation 5. For example, Equation 6 would be used to determine the *g* coefficient for four sessions:

$$g = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_s^2/4 + \sigma_{ps,e}^2/4} \quad (6)$$

When conducting a D study, researchers indicate the universe of generalization, which may differ from the universe of admissible observation sessions discussed previously (Brennan, 1992; Cardinet et al., 2010). For example, admissible sessions may be randomly selected from a universe of classroom sessions. However, researchers may wish to generalize only to the particular set of sessions represented in the G study (considered a fixed or finite inference), or to all classroom sessions that are similar to the measurement context (considered a random or infinite inference; e.g., classrooms in which students with disabilities are included). D studies, therefore, may be used to generalize to a random sampling of students who would be observed in a random sampling of classroom sessions. In the following example, G and D study designs involve generalization to an infinite population and universe.

Study Purpose

Although this approach was developed as early as the 1950s (Gleser et al., 1965), few examples exist of special education researchers using G theory when planning studies involving observational measures (e.g., Bruckner et al., 2006; McWilliam & Ware, 1994; Tindal, Yovanoff, & Gellar, 2010). The purpose of our article is to (a) report findings from a series of G and D studies focusing on observations of the social and academic outcomes of students with severe disabilities in inclusive classrooms, and (b) illustrate how other special education researchers might plan for future observational assessments using these approaches. In addition to describing the steps necessary to conduct G and D studies on pilot and/or prior data sets, we address considerations when initial estimates of reliability are lower or more variable than expected.

We conducted G studies to estimate the reliability or stability of observed scores across classroom observation sessions collected during the first phase of an ongoing intervention project designed to increase peer interactions and academic engagement for secondary students with severe disabilities. We conducted D studies to identify the parameters under which a conservative criterion level of reliability (i.e., .80) may be achieved in future studies focusing on peer interactions and academic engagement in classroom settings. The original data sets included observation sessions lasting for the entire duration of high school class periods (e.g., 45–90 min). Because these sessions were not structured by researchers, contextual variables related to instructional formats varied within and across classroom observation sessions. In Study 1, we conducted G and D studies using all available observation time. After identifying significant variability in *g* coefficients across dependent measures, we then attempted to increase reliability by decreasing variability due to changes in instructional formats across sessions. In Study 2, we limited our data set to the most common instructional formats. That is, dependent measures were recalculated within specified instructional formats and additional G and D studies were conducted using this data subset.

Study 1

Method

We conducted separate G and D studies using data collected at two time points: toward the start of the school semester (pre-assessment) and toward the end of the school semester (post-assessment). We anticipated potential differences in *g* coefficients from pre- to post-assessments on our outcomes measures, as the nature and extent of interaction opportunities and academic engagement is likely to change substantially over the course of a school semester. For example, the variability in social interactions and academic engagement may decrease as the semester progresses and behavior patterns become more stable within a particular classroom setting. Therefore, we calculated *g* coefficients and conducted D studies separately for each measurement time point.

©AAIDD
DOI: 10.1352/1944-7558-119.6.589

Participants and setting. Our analysis included a subset of participants enrolled in a randomized control trial testing the efficacy of two peer-mediated intervention approaches (i.e., peer support arrangements and peer networks). To participate in the study, students (a) were between the ages of 14 and 22 (grades 9–12), (b) had a primary or secondary label of intellectual disability or autism, (c) were enrolled in at least one general education class with support from a paraprofessional or special educator, and (d) were currently or had been previously eligible for the state's alternate assessment and/or had a moderate-to-severe cognitive impairment (as noted in school records). Our pre-assessment data included 63 high school students with severe disabilities. Our post-assessment data included a subset of 36 students who participated in a peer network or who were assigned to the control (i.e., business as usual) condition. Students who participated in a peer support arrangement were excluded because the assessment protocol used for this group was slightly modified for post-assessment observations and they received intervention in the classroom in which we observed. We anticipated our analyses would be less interpretable if a subset of observational data were collected under different conditions. Thus, we included all study participants in the pre-assessment analyses, but only included participants for whom the post-assessment protocol was identical to the pre-assessment protocol in the post-assessment analyses. Participant demographics are presented in the first two columns of Table 2.

We collected all observational data in 12 public high schools in two states. Observations took place in inclusive general education classrooms in which the participating student with severe disabilities was enrolled and supported by a paraprofessional or special educator. The majority of participants (68% and 72% included in the pre- and post-assessment, respectively) were included in elective classes (e.g., Chorus, Art, Photography, Health), whereas the remaining students were included in core academic classes (e.g., English, History, Algebra). The vast majority of participants (94% and 92% included in the pre- and post-assessment, respectively) used speech to communicate, whereas the remaining students used alternative methods (e.g., manual signs, gestures, aided systems). Although the two states revealed distinct proportions of participants in academic versus elective classes (55% and 27%), we found no differences in the instructional formats observed within classrooms across these states.

Table 2
Participant Information for Study 1 (Main Analysis) and Study 2 (Constrained Analysis)

	Study 1 Main analysis		Study 2 Constrained analysis	
Student demographics	Pre-assessment	Post-assessment	Pre-assessment	Post-assessment
Number of students observed	63	36	41	23
Mean age of students	16.6	16.6	16.7	16.6
Gender				
Male	69.8%	75.0%	78.0%	87.0%
Race/ethnicity				
European-American	66.7%	66.7%	65.9%	69.6%
African-American	15.9%	13.9%	19.5%	13.0%
Asian-American	6.3%	5.6%	4.9%	8.7%
Native- or Alaskan-American	1.6%	–	2.4%	–
Other	4.8%	5.6%	4.9%	8.7%
Not reported	4.8%	8.3%	2.4%	–
Primary special education disability				
Autism spectrum disorder	47.6%	50.0%	53.7%	56.5%
Cognitive/Intellectual disability	49.2%	50.0%	43.9%	43.5%
Other health impairment	1.6%	–	2.4%	–
Not reported	1.6%	–	–	–

DOI: 10.1352/1944-7558-119.6.589

Observational coding procedures. We conducted all observation sessions in secondary general education classrooms in which participants were included. We selected one classroom for each student, which served as the context for three observation sessions at pre-assessment and three observation sessions at post-assessment. Observers included graduate students and a university researcher who completed an observer training sequence consisting of (a) scoring 90% or higher on two written quizzes of coding procedures, (b) achieving 90% agreement or higher with expert observers on two training videos, and (c) achieving 80% agreement or higher with an expert observer in a live classroom setting. During observation sessions, observers positioned themselves to be unobtrusive to ongoing instructional activities and peer interactions, and remained in the classroom for the entire class period. Observation session durations averaged 47 min (range, 17–94) for pre-assessment data and 49 min (range, 13–90) for post-assessment data. Variation in session duration was due to differences in class schedules (e.g., block versus traditional scheduling, shortened classes due to testing) and differences in how long the participating student was present in the class (e.g., arriving late to class or leaving early).

We collected live data on tablet computers, using a Lily collector interface and the Multi-Option Observation System for Experimental Studies (MOOSES; Tapp, Wehby, & Ellis, 1995), a software program that facilitates the simultaneous collection of both event- and duration-based measures with notation in real time. We collected data on the following social and contextual variables, which we chose and operationalized based on previous observational research conducted in secondary classrooms (e.g., Carter, Cushing, Clark, & Kennedy, 2005; Carter et al., 2008).

Peer interactions. We recorded peer interactions among students with and without disabilities using frequency counts and subsequently converted data to interactions per minute. We defined peer interactions as verbal (e.g., "Here's your pencil," "Good morning John!") or nonverbal (e.g., raising hand to initiate a high five, waving) communicative behaviors between participating students with severe disabilities and their classmates. Interactions could be either social- or task-related in topic. We counted each conversational turn separately without regard to length of utterance. We coded each communicative behavior according to its source (student with disabilities or peer without disabilities) and purpose (initiation or response). Behaviors such as reading aloud to oneself, echolalia, and conversations with teachers or paraprofessionals were not coded as interactions, as the interventions focused specifically on promoting communicative interactions with peers as an indication of inclusion. We coded peer interactions as *initiations* if they were preceded by at least 5 s without an interaction or if they reflected a change in topic from social- to task-related (or vice versa). We coded all other peer interactions as *responses*. We combined initiations and responses separately for students with disabilities and for their peers without disabilities to create aggregate variables of *peer interactions by student with disability* and *peer interactions by peers without disabilities*. It is important to note there was not necessarily a single peer without disabilities contributing to these variables, as any peer who initiated or responded to the target student with disabilities was counted in the tally of peer interactions. The mean number of peers who interacted with target students with disabilities in a given session was 1.13 (range, 0–5) at pre-assessment and 1.69 (range, 0–7) at post-assessment. These means and ranges did not change for participants included in Study 2.

Academic engagement. We measured academic engagement of students with disabilities as a duration variable and converted it to proportion of total class time. We coded students with disabilities as *engaged in consistent activities* when they were actively involved in or attending to instruction or classroom activities that were aligned with those provided by the general education teacher to the majority of the class. We coded students with disabilities as *engaged in inconsistent activities* when they were actively involved in or attending to instruction or activities that were not aligned with those provided to the majority of the class, but were assigned by a paraprofessional or teacher. We coded students as *not engaged* when they were not actively attending to any activities or materials related to instruction or when no instruction was provided.

Instructional format. We coded the instructional format as a duration variable and converted it to proportion of total class time. Formats included *large group*, which occurred when seven or more students in addition to the focus student were receiving instruction from a single educator; *small group*, which occurred when the focus student worked in a group with between two

©AAIDD
DOI: 10.1352/1944-7558-119.6.589

and six other classmates; *independent work*, which occurred when the focus student primarily worked independently on assignments; *one-to-one peer*, which occurred when the focus student primarily worked with only one other peer; *one-to-one adult*, which occurred when the focus student primarily worked with an adult; and *no instruction*, which occurred when the focus student was not assigned any tasks, or was in a prolonged state of transition between tasks.

Interobserver agreement. A single observer coded three observational sessions for each participant. A second, independent observer also coded one of these three sessions (33% of sessions per participant). We calculated percentages of interobserver agreement to determine the extent to which the two observers agreed. We calculated agreement on peer interactions as the number of agreements (i.e., interaction codes within a 5-s window of agreement) divided by the number of agreements plus disagreements and multiplied by 100%. We calculated agreement on academic engagement as the number of seconds both observers coded each engagement variable as *on* plus the number of seconds both observers coded the engagement variable as *off* divided by the total number of seconds in the session and multiplied by 100%. We calculated kappa coefficients for duration measures only. For pre-assessment data, mean percentages of agreement were 82.3% across peer interaction measures and 99.2% across academic engagement measures. The mean kappa across academic engagement measures was .91. For post-assessment data, mean percentages of agreement were 92.2% across peer interaction measures and 99.8% across academic engagement measures. The mean kappa across academic engagement measures was .93. Because both observers did not code all sessions in our dataset, observer and session facets were not *fully crossed* (see Brennan, 2001, for further discussion) and we did not use the second observer's scores to compute average scores across sessions. In this commonly used framework, observer cannot be considered a facet of measurement in the G study.

EduG software and procedures. To conduct our analysis, we used EduG, a free online G calculator (EduG, 2012). Sums of squares were calculated for each dependent variable using within-subjects ANOVA, with session as the repeated factor. In the EduG software, we indicated person (P) as our object of measurement, with session (S) as the measurement facet. We indicated subjects were randomly sampled from the population of persons with severe disabilities, and three sessions were sampled from the universe of admissible sessions. For the D study component, the EduG software estimated *g* coefficients (using Equation 6) for numbers of observation sessions other than the three observations we conducted at pre- and post-assessment.

Results

Results of Study 1 are displayed in the first two columns of Table 3, which depict the number of

Table 3
Number of Sessions Required to Achieve Reliability Criterion for Each Variable by Assessment Period and Analysis Type

	Number of sessions required to achieve $g \geq .8$			
	Study 1 Main analysis		Study 2 Constrained analysis	
Dependent variables	Pre-assessment	Post-assessment	Pre-assessment	Post-assessment
Initiations by SWD	6	3	6	2
Responses by SWD	2	5	6	2
Total peer interactions by SWD*	2	4	5	2
Initiations by PWOD	9	5	5	3
Responses by PWOD	2	4	7	2
Total peer interactions by PWOD*	3	3	5	2
Engaged consistent	2	3	2	4
Engaged inconsistent	3	11	3	6
Not engaged	3	2	2	2

Note. SWD = students with disability; PWOD = peer without disability.
*aggregated variable (initiations and responses combined).

DOI: 10.1352/1944-7558-119.6.589

sessions required to achieve a criterion *g*-coefficient of .80. Figures 2 and 3 depict how *g*-coefficients are projected to change based on the number of observation sessions conducted at pre-assessment and post-assessment time points, respectively. For pre-assessment data, the number of sessions required to achieve a *g*-coefficient of .80 was high for both initiations by students with disabilities and initiations by peers without disabilities relative to other dependent variables. At pre-assessment, the number of observation sessions necessary to obtain reliable estimates of initiations was between six and nine, whereas all other interaction and engagement variables required two to three sessions to reach the criterion *g*. At post-assessment, the number of sessions required to achieve a *g*-coefficient of .80 decreased for both initiation variables. That is, measures of initiations were more reliable at post-assessment than at pre-assessment. For other variables, however, the number of sessions required to obtain reliable estimates increased from pre- to post-assessment, including responses and engaged-inconsistent, which required no fewer than 11 sessions to attain sufficient reliability. Finally, across both pre- and post-assessment data, aggregated peer interactions by students with disabilities and aggregated peer interactions by peers without disabilities required fewer sessions to achieve the criterion *g* (i.e., were more reliable) than separate initiation and response variables.

Study 2

Method

As in Study 1, we conducted separate G and D studies for pre- and post-assessment data. To

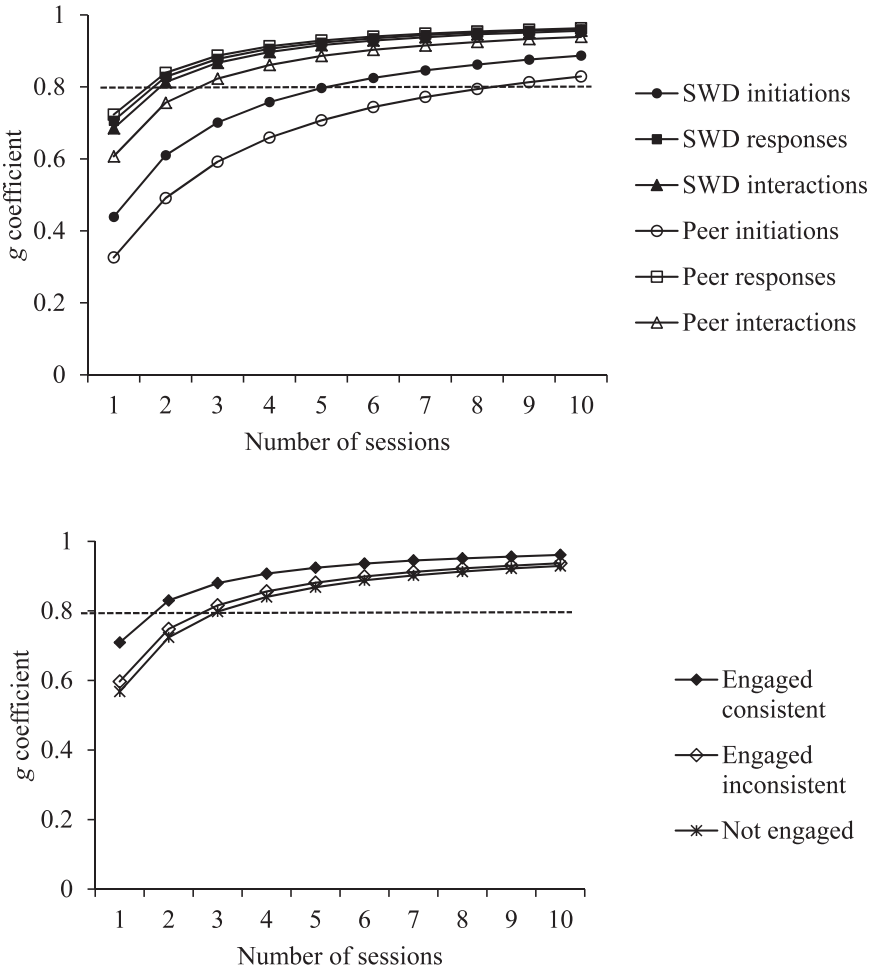

Figure 2. *g* coefficients for interaction measures (top panel) and engagement measures (bottom panel) by number of sessions at pre-assessment (dotted line represents reliability criterion). SWD=students with disabilities.

DOI: 10.1352/1944-7558-119.6.589

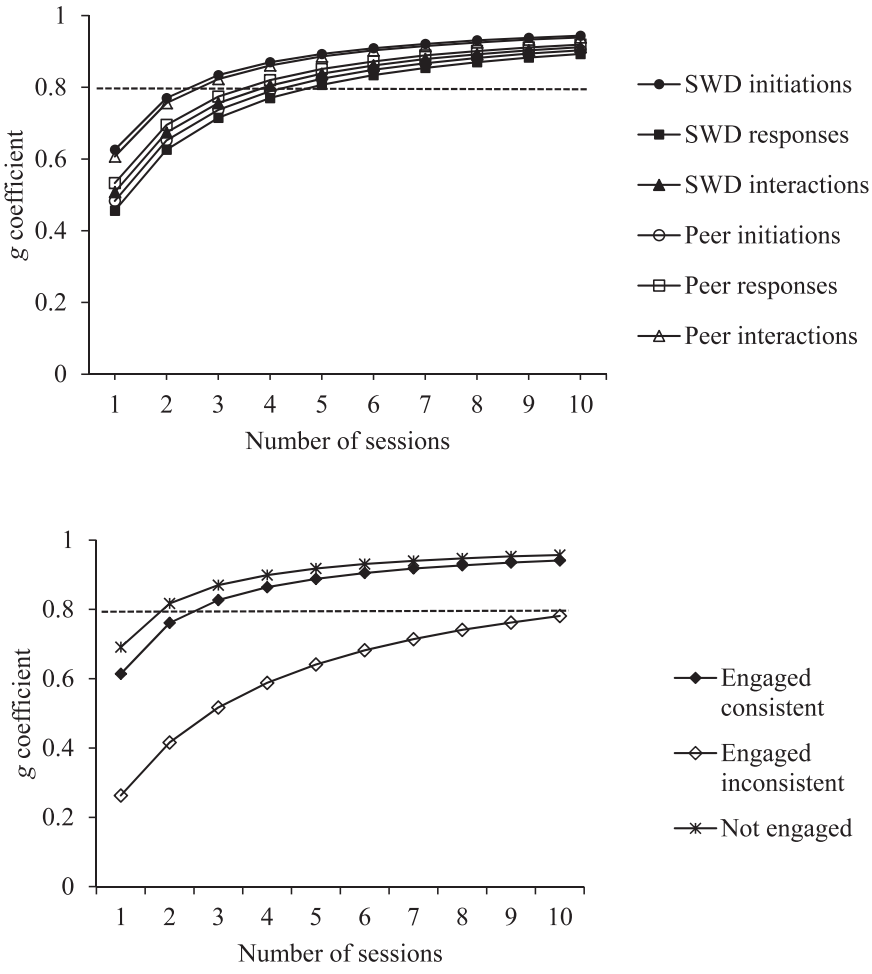

Figure 3. *g* coefficients for interaction measures (top panel) and engagement measures (bottom panel) by number of sessions at post-assessment (dotted line represents reliability criterion). SWD = students with disabilities.

improve the stability of each variable, our second set of analyses focused on sessions constrained by instructional format. We hypothesized that by using data from a subset of instructional arrangements, the stability of the outcome measures may increase. That is, we attempted to remove a portion of variation due to changes in instructional formats (e.g., large group, small group, independent work) within and across observation sessions. To do this, we inspected our data to determine which instructional arrangements met a criterion level of occurrence for each session (i.e., 10 min). We found that combining large-group instruction and independent work arrangements would allow us to keep the largest number of participants in our data sets. In other words, we only included participants who had three sessions of data in which at least 10 min of the session were characterized by either large group instruction or independent work. We recalculated rates of peer interactions and proportions of engagement using the multigroup option in the MOOSES software. Next, we recalculated sums of squares using our reduced data set, and reentered these data into the EduG software to obtain new *g* coefficients and D study output.

Participants and setting. Study 2 analyses included a subset of participants from Study 1. Pre-assessment data included 41 secondary students with severe disabilities across 11 schools, and post-assessment data included 23 students across 10 schools. We identified these participants because their data included a minimum of 10 min of session time in one or both of the two most common instructional formats (i.e., large group and independent work). We anticipated similar

DOI: 10.1352/1944-7558-119.6.589

opportunities for peer interactions across these two formats, as both typically involved students seated at individual desks in proximity to their peers. In addition, the expectation during both large group and independent work formats was that students were not working with their peers but either attending to the teacher's lecture or working on their own. Thus, although we did not expect these instructional formats to be associated with the highest rates of peer interactions, we anticipated these formats to be associated with similar opportunities for peer interactions. Demographic information for participants included in Study 2 is presented in the second set of columns of Table 2. Similar to Study 1, the vast majority of participants (98% and 96% included in the pre- and post-assessment, respectively) used speech to communicate, whereas the remaining students used alternative methods (e.g., manual signs, gestures, aided systems). Over half of participants (56% and 61% included in the pre- and post-assessment, respectively) were included in elective classes, whereas the remaining students were included in core academic classes.

Observational coding procedures and interobserver agreement. Observational procedures were identical to those described for Study 1, except session durations were shorter as a result of sampling data collected during a subset of instructional formats. The mean session duration for the pre-assessment data was 39 min (range, 10–81). The mean session duration for the post-assessment data was 42 min (range 10–88). Interobserver agreement was calculated according to the same procedures used in Study 1. For pre-assessment data, mean percentages of agreement were 83.8% across interaction measures and 99.4% across academic engagement measures. The mean kappa across engagement measures was .93. For post-assessment data, mean percentages of agreement were 90.7% across interaction measures and 99.8% across engagement measures. The mean kappa across academic engagement measures was .96.

Results

Results of Study 2 are displayed in the second set of columns in Table 3, which depict the number of sessions required to achieve criterion *g* coefficients for observational data constrained by instructional format. Figures 4 and 5 depict how *g*-coefficients are projected to change based on the number of observation sessions conducted at pre- and post-assessment time points, respectively. Compared to the post-assessment data from Study 1, the post-assessment data from Study 2 show that constraining the observational data based on instructional format improved reliability. That is, with the exception of engaged-inconsistent, constraining the data to a subset of instructional contexts substantially improved the likelihood of capturing reliable estimates of these variables using between two and four observation sessions. When comparing pre-assessment data from Study 1 and Study 2, however, this pattern of improved stability was only observed for two variables (i.e., initiations by peers without disabilities and not engaged). In comparison to the mixed results between time points of Study 1, the results of Study 2 also reveal increased stability at post-assessment in comparison to pre-assessment (across all interaction measures).

Discussion

We conducted G and D studies on an existing set of classroom data to (a) generate *g* coefficients indicating the stability of a set of dependent variables measuring peer interaction and academic engagement, (b) determine the number of sessions required to achieve a *g* coefficient of .80 for each dependent variable, and (c) evaluate whether constraining the observational data by instructional arrangement would increase *g* coefficients and, thus, result in a fewer number of sessions required to obtain reliable estimates. We found that *g* coefficients and the resulting number of sessions required to meet criterion varied among dependent variables within an assessment time point, as well as between time points (pre- and post-assessment) for each variable. That is, the number of sessions required to produce reliable estimates and, thus, protect against Type II error varied depending on the outcome measure and time point in the semester. Within the context of intervention research, the implication is that the likelihood of correctly identifying an intervention effect can vary based on the dependent variable and, potentially, the time during the school year when data are collected.

Results of Study 1 showed that peer interaction variables for both peers and students with disabilities were less stable when measured as separate initiation and response components. This outcome suggests a tradeoff between the

DOI: 10.1352/1944-7558-119.6.589

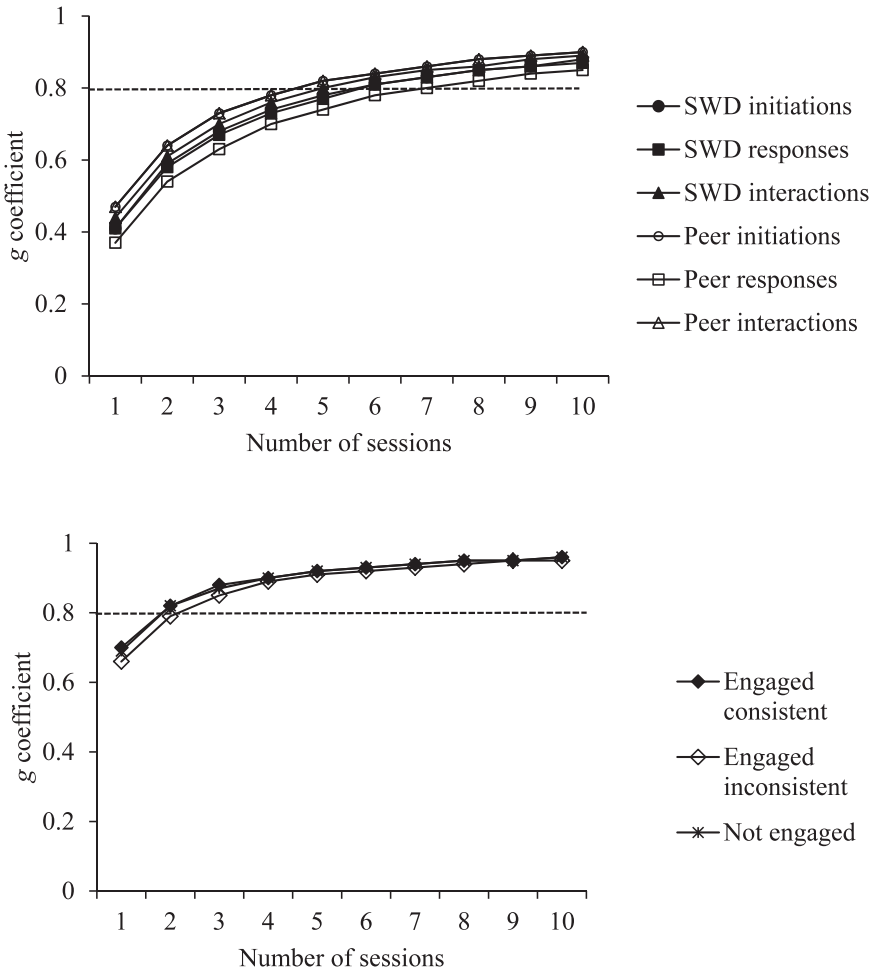

Figure 4. *g* coefficients for interaction measures (top panel) and engagement measures (bottom panel) by number of sessions at pre-assessment, constrained by instructional arrangement (dotted line represents reliability criterion). SWD = students with disabilities.

specificity and reliability of an observational variable such that highly specified variables (e.g., peer initiations) require more sessions to obtain stable estimates than variables defined more broadly (e.g., peer interactions). That is, the advantages of identifying highly specified dependent variables must be considered against the increased risk of a Type II error. In the context of the current study, for example, suppose the peer-mediated intervention did indeed increase interactions between students with disabilities and their peers, but that the aggregate peer interaction variable was the only variable with sufficient reliability to reveal this intervention effect. If this broader interaction variable was not included in the analyses, we would incorrectly assume the peer-mediated intervention had no effect on peer interactions. A second implication is that not all operationally defined behaviors intended to represent a social construct can be measured reliably. Researchers should consider either modifying or removing dependent variables revealing very low *g* coefficients that do not logically fit within a broader aggregate variable to avoid wasting valuable time and training efforts.

Results of Study 1 also revealed that some variables became more stable from pre- to post-assessment, whereas other variables became less stable. For example, *g* coefficients for initiation variables increased from pre- to post-assessment, whereas *g* coefficients for response variables decreased from pre- to post-assessment. The implication of these findings is that reliability can fluctuate across dependent measures and time points. Thus, gaining insight into these possible ranges of reliability and planning the number of sessions

©AAIDD
DOI: 10.1352/1944-7558-119.6.589

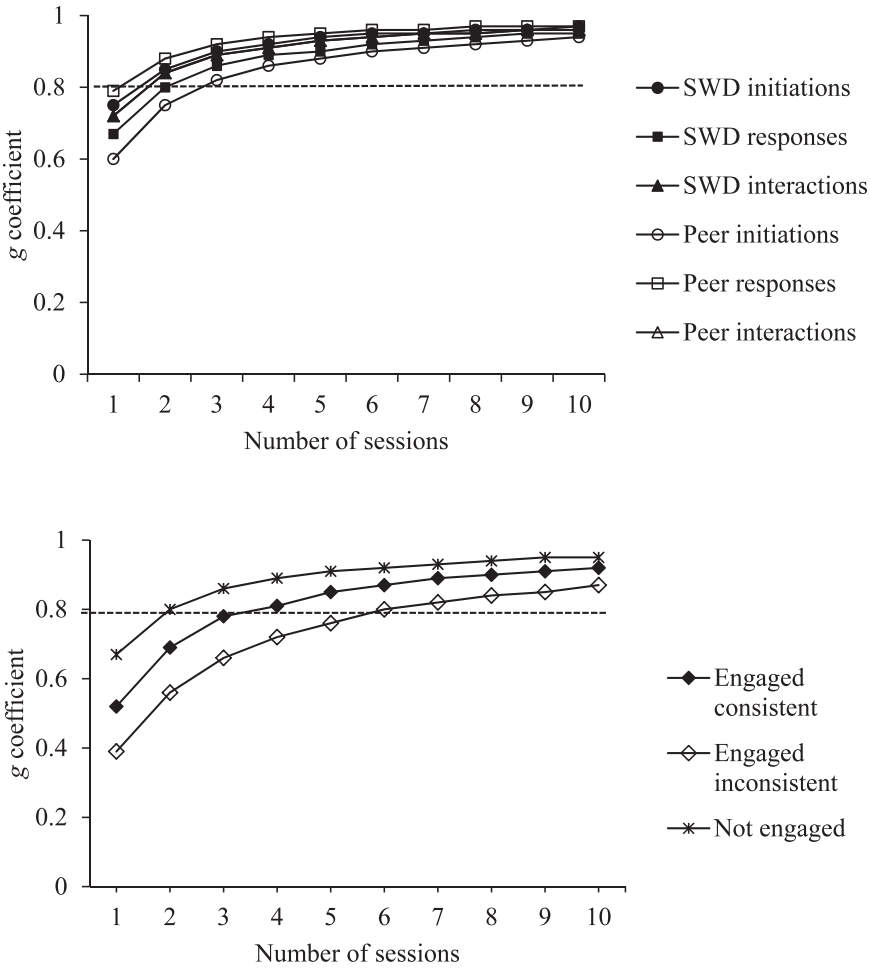

Figure 5. *g* coefficients for interaction measures (top panel) and engagement measures (bottom panel) by number of sessions at post-assessment, constrained by instructional arrangement (dotted line represents reliability criterion). SWD = students with disabilities.

accordingly is better than the alternative of using single observation sessions or arbitrarily identifying *n* sessions and hoping it will be sufficient.

Constraining the observational data by instructional format successfully increased stability across all peer interaction variables within the post-assessment data (compared to the post-assessment data that were not constrained). In addition, the G and D analyses using constrained data revealed a clearer pattern of increased stability from pre- to post-assessment (compared to the mixed results of Study 1). One interpretation of this outcome is that interaction patterns became more stable over time within these particular instructional formats. This outcome differs from results of a recent study showing that, for young children participating in a developmental language intervention, outcome measures at the post-assessment were less stable than those at the pre-assessment (Yoder & Sandbank, under review). The authors suggested newly emerging behavior was being measured at the post-assessment, which was less stable because it had not been fully integrated into the child's behavioral repertoire. In the current study, our focus was on adolescents, some of whom participated in a peer-mediated intervention designed to increase social connections. Both the age range (i.e., 15–22) and the nature of the intervention would suggest development of new behavioral repertoires is not at play in this case. Instead, contextual variations in classrooms, as well as the natural development of peer relationships in particular classrooms over the course of a semester, may be more salient to variations in pre- and post-assessment measures. At the beginning of the semester, interaction patterns between students

DOI: 10.1352/1944-7558-119.6.589

within large-group and independent work instructional arrangements may be less stable. In other words, students have yet to establish typical patterns of interactions or relationships with nearby classmates. By the end of the semester, however, students may have established a more stable interactive routine, with more consistent relationships and modes of interaction and/or engagement. In practical terms, researchers should plan to conduct enough observation sessions at both pre- and post-assessment periods so stable estimates can be derived at both time points.

It is important to note that the findings presented here do not invalidate the observational data collected as part of the large-scale intervention study or prior research projects in which similar dependent measures were collected in classrooms. In fact, when intervention effects are still found with observational variables producing low *g* coefficients, the effect is likely to be a robust one. The primary advantage of conducting G and D studies, however, is the ability to design measurement procedures for future studies that will maximize the likelihood of correctly identifying intervention effects. In this way, the approach used here is similar to a power analysis, which indicates the optimal number of participants to detect statistically significant effects given an expected effect size. Power analyses and D studies are similar in that they are conducted using existing data in preparation for future studies.

In general, these analyses reveal school-based observations have the potential to yield highly variable scores across sessions, and likely require multiple observation sessions per student to achieve stable estimates of social- and/or academic-related outcome measures. Results of Studies 1 and 2 suggest the number of sessions required to obtain stable estimates of these variables may exceed what can be reasonably expected for a research team apart from extensive grant-funded resources. Therefore, researchers should take steps to understand the interactive context in which they will be observing prior to conducting observations and plan measurement procedures that will enable reliable estimation of the construct of interest within the practical boundaries of a research project.

Recommendations

Researchers could conduct G and D studies in preparation for a future research project by either collecting pilot data specifically for this purpose, or by analyzing existing data collected using measurement procedures similar to those they will be using in the future study (as was done here). Results of these G and D studies can guide researchers in designing their measurement procedures to maximize the likelihood of correctly identifying intervention effects. Depending on the measurement facets included in the model (e.g., number of observation sessions, observers), projected *g* coefficients enable researchers to weigh the costs (e.g., time, resources) and benefits (i.e., increase in reliability) of averaging across multiple sessions and/or observers. Though we were unable to calculate the variance in scores due to observer in the current study (only a portion of observation sessions were coded by more than one observer), similar procedures also may be used to incorporate the number of observers in this cost-benefit analysis. For some constructs measured through observational measurement systems, variance due to session outweighs variance due to observer (McWilliam & Ware, 1994), making it especially important to evaluate the contribution of session to measurement error. However, if researchers are interested in determining variance associated with observer, they should collect data in which multiple observers score each session.

In the event that *g* coefficients are low enough to require a number of observational sessions or observers that is not feasible given available resources, our findings suggest three strategies for improving the stability of outcome measures. First, researchers could revisit their dependent variables by aggregating across subcategories of variables or perhaps redefining them. When variables of similar constructs are aggregated, findings from our analyses suggest the stability of the estimate may improve. Although some specificity of measurement is lost, aggregating variables may allow researchers to better estimate constructs at a broader level and maximize the probability of detecting intervention effects. When selecting observational measures, researchers must strike a balance between construct specificity and sufficient stability across measurement opportunities.

Second, researchers could constrain their measurement context along some naturally occurring variable thought to influence the construct of interest. Because our data set included information on instructional format, and because

DOI: 10.1352/1944-7558-119.6.589

instructional format was a variable that could influence both peer interactions and academic engagement, we conducted an additional set of G and D studies using only data collected during two of the seven instructional formats (Study 2). Though the two instructional formats were selected primarily based on convenience (i.e., available data), the constrained analysis revealed increased stability across measures at the post-assessment and a clearer pattern of increased stability from pre- to post-assessment that was not apparent in Study 1. Other examples of contextual variables that potentially influence social and/or academic behaviors include proximity to peers and/or adults, physical classroom arrangements, or types of academic and/or social activities. (Alternatively, researchers could constrain their data sets based on individual student profile variables thought to influence the construct of interest [e.g., diagnosis, mode of communication]. Our analyses did not address variation in individual student profiles, which also may have impacted the reliability estimates.)

We recommend researchers conducting group-design intervention research in classroom settings consider interactive contexts that are both stable *and* likely to elicit the construct of interest. We chose large group and independent work arrangements for two reasons. First, these arrangements represented the most common instructional formats in our data set, which allowed us to include the greatest number of participants. Second, we expected the opportunity for peer interactions and academic engagement would be similar for both arrangements, as both large group and independent work were associated with similar physical seating arrangements as well as social and/or academic expectations by teachers. A potential limitation of selecting these arrangements, however, was that these were not the instructional formats hypothesized to elicit peer interactions. If we were to select naturally occurring contexts based on conditions expected to be more conducive to peer interactions, we may have chosen small-group instruction as our measurement context. Our data set, however, did not include a sufficient number of participants (i.e., 10; Bakeman & Quera, 2011) who were engaged in small-group instruction for a sufficient length of time (i.e., 10 min; McWilliam & Ware, 1994) to produce interpretable results.

Finally, an alternative approach to improving estimates of observational variables would be to purposefully introduce structure into the measurement context as an effort to minimize variability due to changing contextual factors across sessions (Yoder & Symons, 2010). This approach is based on the same logic as constraining the measurement context along naturally occurring contextual variables. For observational measurements in classrooms, researchers could work with educators to introduce an activity or series of similar activities, which may serve as the assessment context. For example, observations could be done during a 10-min small-group worksheet or workbook exercise aligned with classroom instruction. When incorporating structure into the observational context, we recommend researchers consider the extent to which the structured activity (a) elicits the construct of interest, (b) reduces the influence of contextual or "nuisance" variables not part of the dependent variable, and (c) preserves the authenticity of the environment such that there is not a large discrepancy between sessions in which structure is introduced and typical classroom sessions. Introducing enough structure into the assessment will require advance preparation and collaboration with school sites, but also may reduce the number of observation sessions required to yield reliable data.

Limitations

Several limitations to our study should be considered when interpreting these findings. First, we were unable to calculate the variance in scores due to rater, as sessions were not fully crossed with raters in our existing data set. Our reasonably high estimates of interobserver agreement across outcome measures, however, suggest the variance due to rater was relatively minimal and likely outweighed by variance due to observation session. Second, our constrained analysis (Study 2) was focused on a subset of instructional formats that was most represented in our data set and expected to provide similar opportunities for peer interaction and academic engagement. These instructional formats were not, however, likely to be most conducive to peer interactions. Had our data set included sufficient observation time spent in small-group instruction, we would have selected this format for constraining the data and may have identified a different pattern of stability across measures. Third, our analyses did not address variation due to individual student profile variables. That is, constraining our data set according to some student-level variable expected

DOI: 10.1352/1944-7558-119.6.589

to influence the variance in peer interactions or academic engagement may have further increased the stability of the outcome measures.

Conclusion

The current demonstrations of G and D studies using an extant data set can be useful for group design researchers employing similar observational measurement procedures in classroom settings. Although G theory has been applied to the improvement of measurement procedures for over 50 years, its application within special education research has been limited. Whereas conventional wisdom suggests obtaining more data is always better, G and D studies provide researchers tools to evaluate cost-benefit scenarios of additional data collection specific to their own measures and contexts of interest.

References

Bakeman, R., & Quera, V. (2011). *Sequential analysis and observational methods for the behavioral sciences*. Cambridge, UK: Cambridge University Press.

Berk, R. A. (1979). Generalizability of behavioral observations: A clarification of interobserver agreement and interobserver reliability. *American Journal of Mental Deficiency, 83*, 460–472.

Brennan, R. L. (1992). Generalizability theory. *Educational Measurement: Issues and Practice, 11*, 27–34. http://dx.doi.org/10.1111/j.1745-3992.1992.tb00260.x

Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer.

Brooks, P., & Baumeister, A. (1977). A plea for consideration of ecological validity in the experimental psychology of mental retardation. *American Journal of Mental Deficiency, 81*, 406–416.

Bruckner, C. T., Yoder, P. J., & McWilliam, R. A. (2006). Generalizability and decision studies: An example using conversational language samples. *Journal of Early Intervention, 28*, 139–153. http://dx.doi.org/10.1177/105381510602800205

Cardinet, J., Johnson, S., & Pini, G. (2010). *Applying generalizability theory using EduG*. New York, NY: Routledge.

Carter, E. W., Cushing. L. S., Clark, N. M., & Kennedy, C. H. (2005). Effects of peer support interventions on students' access to the general curriculum and social interactions. *Research and Practice for Persons with Severe Disabilities, 30*, 15–25. http://dx.doi.org/10.2511/rpsd.30.1.15

Carter, E. W., Sisco, L. G., Brown, L., Brickham, D., & Al-Khabbaz, Z. A. (2008). Peer interactions and academic engagement of youth with developmental disabilities in inclusive middle and high school classrooms. *American Journal on Mental Retardation, 113*, 479–494. http://dx.doi.org/10.1352/2008.113:479-494

Chung, Y., Carter, E. W., & Sisco, L. G. (2012). Social interactions of students with disabilities who use augmentative and alternative communication in inclusive classrooms. *American Journal on Intellectual and Developmental Disabilities, 117*, 349–367. http://dx.doi.org/10.1352/1944-7558-117.5.349

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regressional correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Cronbach, L. J., Rajaratnam, N. R., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *The British Journal of Statistical Psychology, 16*, 137–163. http://dx.doi.org/10.1111/j.2044-8317.1963.tb00206.x

EduG. (2012). *EduG version 6.1-e, generalizability study*. Société Suisse pour la Recherche en Éducation, Groupe de travail Edumétrie - Qualité de l'évaluation en éducation; software prepared by Maurice Dalois and Léo Laroche, Educan Inc, Longueuil, Qc.

Gleser, G. C., Cronbach, L. J., & Rajaratnam, N. (1965). Generalizability of scores influenced by multiple sources of variance. *Psychometrica, 30*, 395–418.

Jackson, L. B., Ryndak, D. L., & Wehmeyer, M. L. (2008/2009). The dynamic relationship between context, curriculum, and student learning: A case for inclusive education as a research-based practice. *Research and Practice for Persons With Severe Disabilities, 33/34*, 175–195.

Linnenbrink-Garcia, L., Rogat, T. K., & Koskey, K. L. K. (2011). Affect and engagement during small group instruction. *Contemporary Educational Psychology, 36*, 13–24. http://dx.doi.org/10.1016/j.cedpsych.2010.09.001

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*, 34–47. http://dx.doi.org/10.1177/105381519401800104

DOI: 10.1352/1944-7558-119.6.589

Mitchell, S. K. (1979). Interobserver agreement, reliability, and generalizability of data collected in observational studies. *Psychological Bulletin, 86*, 376–390. http://dx.doi.org/10.1037/0033-2909.86.2.376

Pasco, G., Gordon, R. K., Howlin, P., & Charman, T. (2008). The classroom observation schedule to measure intentional communication (COSMIC): An observational measure of the intentional communication of children with autism in an unstructured classroom setting. *Journal of Autism and Developmental Disorders, 38*, 1807–1818. http://dx.doi.org/10.1007/s10803-008-0569-3

Shavelson, R. J., & Webb, N. (2006). Generalizability theory. In J. L. Green, G. Camilli, & P. B. Elmore (Eds.), *Complementary methods for research in education* (pp. 309–322). Washington, DC: American Educational Research Association.

Tapp, J. T., Wehby, J. H., & Ellis, D. (1995). MOOSES: A multi-option observation system for experimental studies. *Behavior Research Methods, Instruments, and Computers, 27*, 25–31.

Tindal, G., Yovanoff, P., & Gellar, J. P. (2010). Generalizability theory applied to reading assessments for students with cognitive disabilities. *The Journal of Special Education, 44*, 3–17. http://dx.doi.org/10.1177/0022466908323008

Yoder, P. J. & Sandbank, M. S. (under review). *Measuring representative communication in 3-year-olds with intellectual disabilities*. Manuscript submitted for publication.

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

Received 3/29/2013, accepted 3/25/2014.

This research was presented as a poster at the 2013 American Educational Research Association Annual Meeting and at the 2013 Annual Gatlinburg Conference. Partial support for this article came from the Institute of Education Sciences, U.S. Department of Education, through Grant R324A100391 to Vanderbilt University and the University of Wisconsin—Madison. The first author received support from the Institute of Education Sciences, U.S. Department of Education, through Grant R324B080005 to Vanderbilt University. We would like to thank the students and families that participated in this research.

Authors:
Kristen Bottema-Beutel, Boston College; **Blair Lloyd** and **Erik W. Carter,** Vanderbilt University; and **Jennifer M. Asmus,** University of Wisconsin-Madison.

Address correspondence concerning this article to Kristen Bottema-Beutel, Lynch School of Education, Boston College, Department of Teacher Education, Special Education, and Curriculum & Instruction, 140 Commonwealth Avenue, Chestnut Hill, MA 02467 USA (email: Kristen.bottema-beutel@bc.edu).