Table of Contents

INTRODUCTION

Research in applied behavior analysis (ABA) and early intervention/early childhood special education (EI/ECSE) often occurs in a clinic or researcher-controlled setting. However, findings in a researcher-controlled environment lack ecological validity and results may not generalize to other settings such as the home environment. Children with challenging behavior often exhibit behavior across settings including the home environment where the caregiver is not trained in ABA and evidence-based practices. Research in ABA and EI/ECSE has increasingly moved from researcher-controlled settings to naturalistic settings to increase the generalizability and ecological validity of findings (Carr et al., 2004; Wolery et al., 2005).

In naturalistic settings such as the home, behavior can vary substantially from any given moment due to the contextual variables such as child's knowledge of the expectations, availability of reinforcement, and the child's physical and mental state. The variability in behavior due to contextual variables reduces the likelihood that any one observational session is a stable estimate of the child's true rate of challenging behavior (Yoder et al., 2019). However, research targeted on parent-implemented interventions often occur in the home environment alone (Barton & Fettig, 2013), thus obtaining stable and representative estimates of child challenging behavior in unstructured, naturalistic environments (e.g., home) is an important goal for researchers in the fields of ABA and EI/ECSE.

Observational Measurement via Direct Observation

Direct observation of child challenging behavior is commonly used in ABA and EI/ECSE research to demonstrate the effectiveness of interventions. Measurement via direct observation involves estimating the true value of the target behavior by applying either a discontinuous (i.e., time-sampling) or continuous (i.e., count of frequency or duration) recording system (Yoder et al., 2019). In discontinuous recording, the observation session is divided into intervals (commonly 5 to 30s in EI/ECSE research; Lane & Ledford, 2014), and then observers record the presence or absence of the behavior during the interval. Rules are created prior to commencing the study and may involve the coder marking the occurrence of the behavior at the end of the interval (momentary time sampling), during the entire interval (whole interval sampling), or at any point during the interval (partial interval sampling). Continuous recording does not rely on behavior sampling, but rather, observers count either the occurrence of observed behaviors (e.g., number of times a child throws an object) or the duration of an observed behavior (e.g., the length of time a child cries). With continuous count recording, observers record how many behaviors occur (event sampling) or both how many behaviors occur and the timing of each behavior (timed event sampling). When selecting a measurement system, researchers consider the dimension of interest (e.g., count or duration), available resources (e.g., software, number of coders), and data reliability.

Timed event sampling is considered the most accurate measurement system by experts in observational measurement (Cunningham et al., 2019; Ledford et al., 2018; Yoder et al., 2019), while partial interval sampling is cited as the most commonly utilized system in educational and ABA research (Cooper et al., 2007; Yoder et al., 2018). Thus, we will focus on the relative strengths and weaknesses of these two approaches. Although accurate, timed event sampling is the most resource-intense method as it requires observational software and re-watching segments of the video to accurately mark the onset or offset of key behaviors (Ayres & Ledford, 2014). Partial interval sampling can be more efficient as coders can mark for multiple behaviors within an interval (Ayres & Ledford, 2014); however, research has documented a high likelihood of measurement error associated when utilizing partial interval sampling to estimate count with both long intervals (Ledford, et al., 2015; Mann et al., 1991) and high rate behaviors (Mann et al., 1991). To account for the systematic error associate with utilizing partial interval sampling, observational measurement experts recommend using Poisson-corrected estimates and short intervals to increase the accuracy of the estimates (Yoder et al., 2018); however, other experts point out partial interval sampling depends on both the prevalence and incidence of the behavior leading to reduced construct validity even with Poisson-corrected estimates (Pustejovsky & Swan, 2015). The most common interval length when utilizing partial interval sampling is 10 s (Fiske & Delmolino, 2012) and partial interval sampling with 10 s intervals has been demonstrated to be sensitive to detecting changes across conditions, however the sensitivity decreases as the rate of behaviors increase (Rapp et al., 2008). Despite the apparent disadvantages to using partial interval sampling, recent reviews reported partial interval sampling to be the most common method in challenging behavior research (Maggin et al., 2017, Lloyd et al., 2016) and 10 s was a common interval length for measuring the impact of parent-implemented interventions (Duda et al., 2008; Dunlap et al., 2006; Fiske & Delmolino, 2012).

Measurement of Reliability

Current standards in both group and single-case design suggest the internal validity of a research study is contingent on the reliability of the data (What Works Clearinghouse Single Case Research Design Standards, Kratochwill et al., 2010, 2013). Conventionally, reliability of observational data is calculated by comparing the scores of two observers on the same observation. With partial interval and timed event sampling, the timing of key events is recorded allowing for the calculation of point-by-point agreement (i.e., comparing the timing and recorded behavior with a priori agreement time-window). Percent agreement is commonly used, but it is influenced by the rate of the observed behavior. The statistic may be artificially inflated when the behavior occurs often (Lei et al., 2007) or when interval sampling is utilized (Rapp et al., 2011). Low reliability increases the likelihood of a Type II error or concluding an intervention did not work when it did. On the other hand, when reliability data are artificially high it may lead researchers to be more confident in their data and identify an effect when one is not present (i.e., Type I error). To reduce the probability of a Type I error, occurrence and non-occurrence agreement should be reported when utilizing interval sampling (Bailey & Burch, 2017; Harris & Lahey, 1978; Yoder et al., 2019).

The difference between any observed score and the true score is referred to as measurement error. Interobserver agreement is only one of the possible sources of measurement error in relation to direct observation of behavior. Other sources include coder, item, method, setting, and dimension (Cone, 1977; Yoder et al., 2016). Generalizability (G) theory can be used to improve the reliability or stability of observational variables and identify the primary sources of measurement error (Cronbach et al., 1972; Gage et al., 2018; Yoder et al., 2016; Yoder et al., 2019). Reliability and stability in G studies are often used interchangeable as they reference the scores stability across observers, contexts, or time (Yoder et al., 2019). In G theory, the true score is conceptualized as the average of all valid observations, and any one study is just a sample. The goal in group design research is to identify the variance across individual participants, thus it is important to separate variance due to person (i.e., differentiated facet) from variance due to other sources of measurement error (i.e. measurement facets). G theory uses analysis of variance (ANOVA) logic to estimate the amount of variability contributed by the person (i.e., individual participants) compared to the amount of variability contributed by the measurement facets (Yoder et al., 2019). The calculated mean squares for the various facets can be used to partition true person variance from variance due to measurement facets. The partitioned variance can then be used to calculate a generalizability (g) coefficient which indicates how accurately the observed scores can be generalized to the person’s behavior in all possible situations (Cronbach, 1972; Yoder et al., 2016). In other words, the *g* coefficient is a type of intraclass correlation that can be used to indicate a criterion level of measurement stability has been achieved.

G studies are characterized by the number of measurement facets and the number of levels of each facet with level being the number of units of each facet. Once the facets of interest are identified, researchers can partition measurement error to each facet allowing the identification of the measurement facet that is contributing to the most measurement error. This information can be used to improve the stability of an estimate in future studies.

Decision Studies

Following a G study, a Decision (D) study is conducted using the variance estimates from the G study to predict the stability of estimates in future research. The goal of a D study is to answer “What if” questions regarding variation in measurement design. In ABA and EI/ECSE research, two important questions are:

1. If using timed event sampling, how can we increase the stability of our estimate?
2. If using partial interval sampling, how can we increase the stability of our estimate?

In a D study, a researcher uses the variance estimates to identify changes to the included measurement facets that would lead to optimization of design characteristics to minimize measurement error. In other words, the information from a D study will allow future researchers to compare and augment various measurement designs to reach a desired level of reliability (Shavelson & Webb, 1991; Yoder et al., 2016). In a D study, you determine a priori the minimum reliability criterion and calculate coefficients for different combinations of facets. Benchmarks for adequate and robust reliability and dependability range from .60 (Berk, 1979) to .80 (Cardinet et al., 2011). D studies can be used to improve future studies by estimating the number of observations, the duration of observations, or the number of raters needed to compute a stable estimate of a generalized person characteristic (Yoder et al., 2019). A D study will allow us to determine whether adding minutes of the observation, adding raters, or adding occasions will lead to improved stability of the true score estimate. This information can be used to inform future studies when challenging behavior is a variable of interest. Longer observation sessions can lead to increased participant burden and may result in attrition. Knowing the minimum length of sessions needed for each behavior sampling method can inform researchers on what method to select and how to potentially reduce participant burden. The information allows researchers to select the least resource-intense method that yields sufficient stability (Yoder et al., 2019).

Measuring Validity

Validity refers to whether a test measures what it is designed to measure. A score can be reliable (i.e., have a high *g* coefficient) and still lack validity. However, a score cannot be valid without acceptable reliability. One means to establish validity of a test is to compare observed scores with a previously validated test that measures the same construct. This is referred to as concurrent validity. When observed scores correlate with scores on the validated test, the scores are said to have high concurrently validity.

Previous Examples

To our knowledge, no previous G studies have been conducted on child challenging behavior. However, previous G studies on other direct observational measures of child behavior (e.g., engagement, communication) can inform our decisions for the current study. McWilliam and Ware (1994) measured nine categories of engagement using a 10s partial interval sampling method. Forty-seven children were observed across four 15-min sessions, each coded by three independent coders in a fully crossed design. McWilliam and Ware found that 50% of the variance in observed scores was due to the interaction between sessions and children. In a D study, McWilliam and Ware found that longer sessions resulted in the ability to average across fewer sessions, but not by much leading them to decide more sessions is better than longer sessions. In 2019, Bottema-Beutel and colleagues measured two states of joint engagement in 20 children ages 7 to 17 months. Half of the participants had a sibling with an Autism diagnosis, while the other half had a sibling without any diagnoses. Children were observed across two 15-min sessions, each coded by two coders. Bottema-Beutel and colleagues found that the number of sessions and coders needed varied both by the participant group (sibling with ASD vs sibling without ASD) and by the dependent variable (higher-order vs. lower-order engagement). The studies on engagement inform future researchers that sometimes longer sessions are helpful in obtaining stable estimates of child behavior, but in general averaging scores across several sessions is better than scores from a single long session. Additionally, other characteristics of the variables may impact the optimal study-design.

Bruckner and colleagues (2006) demonstrated the value of G and D studies in examining the conversational level of 24 preschoolers with grammatical and phonological impairments. Buckner and colleagues collected two, 20-min unstructured language samples with an adult examiner. Language samples were transcribed and coded for mean length of utterance, number of different word roots, total utterances, and intelligible utterances. Buckner and colleagues found one session coded by one rater was sufficiently reliable for all measures with the exception of intelligible utterances. To achieve sufficient reliability on intelligible utterances, scores across five sessions and one rater would need to be averaged. Bruckner and colleagues concluded measuring intelligible utterances in unstructured contexts may not be feasible for future research. In 2016, Yoder and colleagues utilized G theory to established criterion-related validity of a measure of speech comprehensibility. Yoder and colleagues measured speech comprehensibility in 10 elementary-aged children with Down syndrome. Children were observed across eight 5-min sessions, each rated by 4 independent, untrained coders. Yoder and colleagues found averaging across four untrained observers on four 5-min segments resulted in stable ratings. Yoder and colleagues then compared the stable scores to an orthography-based measure. High correlation between the two measures provided evidence that the orthography-based measure was valid. The studies on communication demonstrate that the variable of interest, the measurement context, and the level of coder training are important considerations for how many sessions and coders are needed to achieve sufficient reliability.

RESEARCH QUESTIONS

The goal of the current study was to identify which study-design characteristics are needed to obtain a stable and valid measure of challenging behavior in the home environment. Specifically, we wanted to answer the following research questions:

1. Which aspect of the measure (e.g., length of observation, behavior sampling method, number of different coders, number of occasions) accounted for the largest percentage of error variance in child challenging behavior (i.e., produces the most measurement error)?

2. What is the minimum length of observation needed to achieve minimal reliability (g coefficient) and dependability (phi coefficient)?
3. Varying the number of coders, length of sessions, and number of occasions, what study-designs produce stable estimates of child challenging behavior using timed event sampling?
4. Varying the number of coders, length of sessions, and number of occasions, what study-designs produce stable estimates of child challenging behavior using 10s partial interval sampling?
5. Using the optimal scores from the D study, do observed scores on child challenging behavior correlate with parent-reported challenging behavior? In other words, does our measure have criterion-validity evidence.

METHOD

Participants

We conducted a post hoc analysis of data from a longitudinal randomized controlled trial (RCT) examining the effectiveness of a technology-based parent support intervention. After obtaining IRB, caregivers of children ages 2 to 6 with high rates of challenging behavior were recruited to participate in the larger RCT. As part of the RCT, families submitted videos to Box of a home-based routine during which their child engaged in high rates of challenging behavior. Caregivers were asked to submit at least five, 15-min videos prior to receiving the intervention. See Barton et al. (in preparation) for RCT method and results. Inclusion criteria for the RCT included: (a) at least child aged 2 to 6 years diagnosed with a disability or demonstrating a delay in social-emotional development measured via the Ages & Stages Questionnaires: Social-Emotional, 2nd Edition (Squires, Bricker, & Twombly, 2015) and (b) caregiver-reported challenging behavior at least 3 times per week. Families were excluded from the study if the reported challenging behavior required immediate, intense intervention (e.g., self-injury). To be included in the current study, participants must have submitted at least four 15-min videos prior to receiving the intervention. Fifteen minutes was selected to segment 5 min sections for analysis of video length in the model. In the initial RCT, a total of 57 families were recruited. Of those families, 33 met criteria for the current study. Caregivers were primarily white (*n* = 27; 82%), female (*n* = 33; 100%), and college educated (*n* = 26; 79%). Caregivers also self-identified as Asian (*n* = 3; 9%), Hispanic or Latino (*n* = 3; 9%), and Black ((*n* = 2; 6%). Additionally, the majority of participants had at least one sibling in the home (*n* = 26; 79%). Child age and diagnosis information forthcoming.

Observational Raters

Raters were two first-year masters students enrolled in a special education program with no prior experience coding. Rater 1 was a 22-year-old, white female with a bachelor's of education in Cross Categorical Special Education and 3 years of experience working with children. Rater 2 was a 24-year-old, white female with a bachelor's of arts in Psychology and 3 years of experience working with children.

Observational Coding Procedures

Four 15-min occasions were collected for each child participant. Caregivers were instructed to record a challenging routine and interact with their child in a typical fashion. Parents self-recorded the routine with a researcher-provided iPad and uploaded the videos to Box. Caregivers had two weeks to submit videos. Each occasion (i.e., 15-min video) was coded by two independent coders via ProcoderDV observational software. If a video was longer than 15 min, coders coded the first 15 mins. To estimate child challenging behavior, we utilized two sampling methods on each occasion: timed event and 10s partial interval. When using timed event sampling, the coders marked the onset or beginning of each instance of challenging behavior observed. An instance began and ended after a 3s pause with no behaviors present. When using partial interval sampling, the coders marked the presence or absence of an onset of challenging behavior for each 10s interval. If challenging behavior began (i.e., onsets) at any moment during the interval, coders marked challenging behavior as present. In other words, latency from the challenging behavior in the previous intervals was not used for decisions. If a behavior began in one interval and continued into another, the behavior was marked in the first interval it was present. These sampling methods were selected as they are considered the most valid for estimating frequency of behavior (Cunningham et al., 2019; Ledford et al., 2018; Yoder et al., 2019) and the most commonly used in the field (Cooper et al., 2007; Yoder et al., 2018). Although recommendations for partial interval sampling suggest a shorter interval (e.g., 5 s) may be more valid and sensitive, a 10 s interval was selected given that it is used most commonly and to ensure our findings can inform current research practices (Fiske & Delmolino, 2012; Yoder et al., 2018).

Challenging behavior. Challenging behavior was defined as behavior that interferes with the child's meaningful engagement in the environment or social interactions. Common behavior topographies include physical aggression, verbal aggression, tantrumming, and elopement. Non-examples included repetitive or self-stimulatory behaviors that do not meet other criteria below and behaviors defined by their absence (e.g., noncompliance). See Table 1 for operational definitions, examples, and non-examples. Behaviors were not considered challenging when appropriate for the context or the child's ability (e.g., development, learning). For example, a child appropriately kicking a ball in the context of play was not coded as a challenging behavior; a child kicking a truck after given the vocal direction "Please clean up." was coded as a challenging behavior.

Training. First-year masters students enrolled in a special education program with no prior experience coding were trained following three separate steps. Coders met with the first author to review observational measurement and overall goals of the study. Next, coders received an extensive coding manual with a description of each behavior sampling method as well as definitions, examples, and non-examples of challenging behavior (See Table 1). Coders then met with the first author to review the coding manual and together coded one video from the RCT, which was not included in the current study. Training began with partial interval sampling for all coders. After the initial meeting, coders independently coded non-study videos using partial interval sampling. Following each coded video, the first author calculated point-by-point agreement between the first author's code and the coder's code separately for each behavior sampling method. The first author met with each coder after each video and conduct a discrepancy discussion on all disagreements. This process continued until 80% point-by-point agreement was reached with the first author on 3 videos. Once criterion-level training was reached for partial interval, the process was repeated with timed event sampling. Once trained on both sampling methods, each coder separately and independently coded each 15-min session twice, once with partial interval and once with timed event in a separate pass of coding. Videos were assigned to coders so that the second pass was at least a week after the first pass. We selected this fully crossed design to analyze the variance due to the main effects and interactions of each measurement facet.

Data extraction. Following the coding of each session, the first author and a first-year master's student each moved the data from the individual code files to one excel spreadsheet for analysis. Each researcher extracted the data for 15 min, 10 min, and 5 min segments for each occasion, rater, and sampling method. The start times for the 10- and 5-min segments were randomly selected using a random clock generator (https://www.random.org/clock-times/?mode=advanced) at the participant level so that each occasion began at the same time for -min segments and each occasion began at the same time for 10-min segments for each participant. After each researcher extracted the data, the first-author compared the two excel files. All discrepancies were corrected by checking the individual code file. Inter-rater reliability for data extraction was 96.84. Once confirmed, data were converted to rate by dividing the count by the length in mins.

Criterion-Validity Measure

During the initial RCT, parent participants completed the Child Behavior Checklist (CBCL; Achenbach & Rescorla, 2001) prior to intervention. The CBCL is a 64-item questionnaire regarding the child's behavioral and emotional behavior. Parents rate each item as very true (2), somewhat true (1), or not true (0). The CBCL provides a score for two subscales: internalizing and externalizing behaviors.

Generalization Study

To answer our research questions, we applied a fully crossed five-facet design, Participant, Occasion, Length, Rater, and Method. We used an analysis of variance model (ANOVA) to examine the sources of variability in the observed scores across the five facets. The facets were random effects in the model to be able to generalize findings to the entire population of young children with challenging behavior and to the entire population of home observational measurement contexts. According to G theory, the mean squares from the facets in the design can be used to separate person variance from variance due to the measurement facets (Yoder, Woynaroski, & Camarata, 2015). The estimates of variance from the ANOVA were used to calculate the absolute level of stability (i.e., absolute G coefficient) with EduG version 6.1 (Swiss Society for Research in Education Working Group, 2012). The G coefficient is an intraclass correlation that indicates the level of measurement stability. We selected the absolute *g* coefficient instead of the relative as this approach is more conservative and allows for consideration of the main effects of each measurement facet (Yoder et al., 2018).

Decision Study

The variance estimates from the G study were used to determine the number of occasions and raters we need to average to get a threshold level of stability (e.g., .70; Shavelson & Webb, 1991, Yoder et al, 2015). We averaged scores across variations of Occasion and Rater while leaving Length and Method consistent to identify the optimal conditions to minimize error. We made this decision because in practice you would not average scores across 5, 10, and 15-min samples nor scores across different sampling methods. For each D study, we left Rater consistent at 2 and increased number of occasions until gains were minimal. We selected this method because number of occasions contributed more measurement error than the number of raters. This finding is consistent with previous G studies on child behavior (McWilliam & Ware, 1994; Bottema-Beutel et al., 2019). Next, we increased raters until we found a combination that reached or crossed an absolute G coefficient of 0.70. We completed this process across 6 sets of data: (1) 5-min observation measured with timed event, (2) 10-min observation measured with timed event, (3) 15-min observation measured with timed event, (4) 5-min observation measured with partial interval, (5) 10-min observation measured with partial interval, and (6) 15-min observation measured with partial interval. Results of the D studies with number of raters held at two and number of occasions varied are presented in Figure 2.

RESULTS

The absolute *g* coefficients from the generalization studies ranged from 0.29 to 0.53. The absolute *g* coefficients demonstrate one session of any length or behavior sampling method will not produce a stable estimate of the generalized behavioral tendency of child challenging behavior. See Table 2 and Figure 1 for a summary of the G study findings.

Question 1. The differentiated facet (i.e., Person) accounted for 15.2% of the variance. In other words, less than 16% of the variance in child challenging behavior scores were due to individual differences in behavior – well below field standards. The largest source of measurement error was the Person X Occasion interaction which accounts for 39.8% of the variance in scores.

Question 2. The measurement length that resulted in the most stable scores was 5 mins (see Figure 1). Additionally across behavior sampling methods, 5-min observations produced more stable estimates than either 10- or 15-min observations. The absolute *g* coefficient for variations of occasions and raters across the three levels of length are graphed in Figure 2.

Question 3. To identify optimal measurement characteristics for designs using timed event sampling, we ran 6 D studies in which we kept the Method and Length consistent (i.e., timed event sampling at each 5, 10, 15 mins) while varying the number of Occasions and Raters until a threshold of .7 was achieved. For 5-min segments results indicate you would need to average scores across 15 occasions and 2 raters or 10 occasions and 3 raters to obtain a G coefficient at or above the threshold of .7 stability. However, for 10-min segments you would need to average across 20 occasions and 4 raters to achieve the same level of stability, and for 15-min segments you would need to average across 25 occasions and 3 raters.

Question 4. To identify optimal measurement characteristics for designs using partial interval sampling, we ran 6 D studies in which we kept the Method and Length consistent (i.e., partial interval sampling at each 5, 10, 15 mins) while varying the number of Occasions and Raters until a threshold of .7 was achieved. Results indicate for 5-min sessions you would need to average scores across 12 occasion and 4 raters to obtain a G coefficient at or above the threshold of .7 stability. However, for 10-min segments you would need to average scores across 60 occasions and 6 raters, and for 15-min segments you would need to average across 60 occasions and 5 raters.

Question 5. Due to the lack of stability/reliability in our observed scores, a criterion-validity study would not be appropriate in this situation. A measure can only be valid if it is first stable.

DISCUSSION

The goal of the current study was to provide recommendations on number of occasions and raters needed to obtain stable estimates of challenging behavior across combinations of various session lengths and sampling methods. Our results highlight that child challenging behavior is highly variable across occasion. Considering our current results, we recommend researchers examining child challenging behavior in the home environment use timed event sampling with 10 occasions and 3 raters to reduce participant burden. We find this study-design to be the least resource-intense method that yields sufficient stability (Yoder et al., 2019). However, the resources to design a study with those characteristics are likely not feasible. Thus, we recommend researchers increase the stability of estimates by average scores across multiple measurement contexts (e.g., school and home; Yoder et al., 2019). Additionally, we recommend researchers use caution when interpreting results using 10 s partial interval sampling to measure child challenging behavior in homes.

Both sampling methods produced surprisingly unstable estimates of challenging behavior. This highlights the volatility in child challenging behavior in the home environment. This information is helpful for both researchers and practitioners. Children need to be observed multiple times regardless of which sampling method you select. Until the field can identify more stable methods to measure child challenging behavior, researchers should use other dependent variables along with challenging behavior to make decisions regarding intervention effectiveness. Additionally, we confirmed our hypothesis that the optimal study-design characteristics (e.g., number of raters and occasions) vary due to other variables in the design. In our data, the optimal design varied by length and sampling method while Bottema-Beutel and colleagues (2019) found the optimal design varied by participant group and dependent variable. Our findings also support the previous data suggesting averaging across more sessions is better than averaging across longer sessions (McWilliam & Ware, 1994). We were unable to answer our question regarding the minimal length of sessions required to obtain optimal stability; however, we do have some evidence that suggests longer sessions do not produce more stable estimates. Using each behavior sampling method, the 5-min segments produced the least resource-intense study-design with adequate stability. This information is beneficial for researchers in planning where to allocate resources. We recommend researchers consider shorter observations with more sessions and raters.

Limitations. A major limitation of our study was the inability to answer all of our research questions. We were unable to establish stability and thus could not examine evidence of validity. Importantly, our results regarding partial interval sampling may only be generalized to measurement systems with 10s intervals. It is possible shorter intervals would produce more stable estimates. Additionally, our study is limited in that this is one study from one research team. We cannot rule out the possibility that these findings are unique to our coding system without systematic replications of the methods we used here.

Future directions. In our current study the facet rater contributed relatively little error variance in the estimated scores. However, in observational measurement much emphasis is placed on the agreement scores between two raters. Standards for agreement are an average of at least 80% agreement across conditions calculated for at least 20% of total sessions (Horner et al., 2005; Kratchowill et al., 2013; Shadish, Hedges, Horner, & Odom, 2015). However, our study highlights the importance of examining other sources of error in our measurement systems. Researchers regularly use one occasion pre- and post-an intervention as evidence of effectiveness. We suggest researchers use caution in interpreting results of studies using on one occasion to estimate child challenging behavior.

A reasonable conclusion is that these results are unique to our study and coding system. Evidence from previous studies support each of our findings, which suggests these findings are not unique to our study and coding system. We encourage future researchers to replicate our study methods to confirm whether or not these findings generalize across research teams and contexts. McWilliam and Ware (1996) and Bruckner and colleagues (2006) found that to obtain acceptable stability estimates, you needed more occasions and raters than is reasonable for a research team. However, Yoder and colleagues (2016) also found that more highly trained raters/systems increased the stability of scores on child communication. Future studies should examine if highly skilled raters, regular agreement checks, and discrepancy discussions would increase the stability of the estimate of child challenging behavior in an unstructured context. In addition, researchers examine if adding more structure to the context (e.g., giving parents more directive) or aggregating scores across both unstructured and structured contexts (Yoder et al., 2019) would increase the stability of the estimate.

.

Table 1. Operational Definitions of Challenging Behavior

Topographies	Definitions	Examples	Non-examples
	Physical aggression toward objects		
Pounding objects on surface	Two or more forceful contacts (e.g., object pushed toward surface from at least 6 inches away from the surface and brought quickly toward it) with an object on a surface (e.g., table, floor, wall), with no more than 2 s between contacts, with the object remaining in the child's hand.	Child bangs empty cup on the table repeatedly after asking for more milk.	Child bangs toy hammer on table to "fix the broken leg."
Throwing objects forcefully	An object forcefully (i.e., requiring a "pulling-back" motion of the arm) projected from a child's hand. Can be toward a person or not.	Child is told "No," to a request and she throws a toy across the room.	Child is cleaning up toys and tosses a toy into the toy bin.
Destroying objects or property	Any inappropriate use of objects or property that has the potential to harm them (e.g., knocking off surfaces, pulling off walls, kicking, punching, shredding, ripping, biting, breaking, etc.).	Child swipes a knick-knack off the shelf forcefully and the object hits the floor and breaks.	Child is reaching for high book on shelf and knocks off a knick-knack that breaks.
	Physical aggression toward others		
Pinching	Using pointer finger and thumb to forcefully squeeze the body or clothing of another.	Child uses pointer and thumb to squeeze brother's arm until he cries out.	Child pinches his brother's shirt and brings it toward him, saying, "I'm trying to see the picture on your shirt!"
Hitting (pushing, punching, slapping)	Forceful physical contact or attempts at contact with hand, body, or object (i.e., requiring a "pulling-back" motion of the body part) to the body of another, clothing, or object another is holding.	Child walks up to sister and forcefully bangs into her with her side body.	Child walks up to sister, trips, and falls on her knocking her over.
Biting	Closing teeth with the body or clothing of another within them; does not need to be forceful.	Child is told "No," by Dad, walks up to him, and latches on to dad's leg with his mouth.	Child and friend are playing dentist, and child closes mouth on friend's hand before she's had time to remove it.
Spitting	Release of gob (sufficient to see and more than that projected from talking) of saliva that is forcefully projected toward another person (e.g., sticking out tongue and blowing raspberries).	Child forcefully releases saliva in direction of brother.	Child lets saliva slowly drip onto the table and starts to play in it.
Kicking / Stomping	Forceful physical contact with foot (i.e., requiring a "pulling-back" motion of the leg) to an object or part of a body of another.	Child walks up to family dog and stomps on dog's tail.	Child walks by family dog and accidentally steps on dog's tail.
Hair-pulling	Forceful pulling (i.e., requiring a clear pulling away motion of the hand or arm from head)—either downwards or outwards—of the hair of another person.	Child's sister won't share desired toy and child pulls sister's hair forcefully.	Child is brushing sister's hair and sister cries out, "Ow! It's tangled!"
Taking toys/items away forcefully	Pulling toys or items forcefully (i.e., rapidly and/or using a "pulling-away" motion) from another person without permission.	Child is told he cannot play with the iPad and he grabs it from his mom.	Child is told he can play with the iPad and grabs for it forcefully.
Scratching	Forcefully (i.e., rapidly and/or using a "digging-in" motion) "raking" fingernails across the body or clothing of another.	Child is told to clean up and forcefully rakes fingers across adult's (e.g., parent's) arm.	Child is "tickling" adult (e.g., parent) using tips of fingers.
	Verbal aggression		

Inappropriate demands	Requests that are aggressive in tone.	Child yells, “Give it to me now.”	Child says, “Give me a turn.”
Yelling or screaming	Vocal output that is notably louder than that of the child’s loudest typical conversational level. Do not code output communicating excitement (e.g., “Woohoo”) or play schemes.	Child is told it is bedtime and loudly emits a piercing wail.	Child is playing “Dinosaur” and stomps around the room loudly roaring. Child whining
Calling others names or bad words	Directing negative terms (e.g., bad, wrong, ugly) towards another person (the person does not have to be present).	Child tells sibling, “You’re ugly.”	Child tells sibling, “You’re being mean.”
Talking negatively about self	Using negative terms (e.g., bad, wrong, ugly) to describe oneself or one’s actions.	Child says “I broke it. I’m a bad boy.”	Child says “I don’t know how to do this.”
Threatening others / Negative sentiments	Any statement of intention to inflict pain, injury, damage, or other hostile action on another person.	Child is told “No” by adult and child says “I hate you.”	Child is playing with Army figures and the “bad guys” tell the “good guys” they are going to kill them.
Tantrumming			
Flopping / Flailing	Thrashing and throwing oneself around, by moving arms and legs in a manner inconsistent with purposeful ambulation. Can be standing, sitting, on the floor, etc. Do not count any instances that occur as part of a play sequence (e.g., dancing).	Child’s toy is taken away (by adult or another child). He falls to the floor and begins to throw arms and legs around wildly.	Child is given favorite snack and begins to jump and swing arms wildly around.
Crying	Loud, disruptive, or excessive crying that does not appear to be appropriate given the context.	Child is told to stop grabbing for a toy and he begins to cry.	Child begins to cry after banging her head.
Non-compliance			
Elopement	Moving away from adult or a task (e.g., crossing the threshold of the door, without evidence the moving away is to comply) when a directive was stated to do otherwise. If the child moves out of the video when given a directive (and the required task remains in the video), code.	Child runs to bedroom when adults says, “It’s time for dinner.”	Adult says, “It’s time for dinner” and child runs to bathroom, saying “Ok, let me wash my hands!”
Verbal refusal to follow directions	A verbal refusal (e.g., “No,” “I don’t want to,” “Not right now,” etc.) to comply with a directive that is emitted within 5s of being given one.	Adult asks child to turn off the TV and child replies, “When my show is over.”	Adult asks child if she wants a snack and child replies, “Not right now.”
Other			
Other challenging behavior	Any clearly challenging behavior not captured by this code.	Unsafe climbing on furniture, slamming sibling’s fingers in cabinet, self-inflicted emesis	Self-stimulatory behavior, object mouthing, exploratory play, whining

Table 2. Source, degrees of freedom, mean squares, and percentage of variance explained in challenging behavior.

Source of variance	Degrees of freedom	Mean squares	Percentage of variance explained
Persons	32	10.35	15.2
Occasions	3	1.46	0.0
Lengths	2	0.02	0.0
Raters	1	77.81	13
Methods	1	1.59	0.3
Persons x Occasions	96	4.23	39.8
Persons x Lengths	64	0.54	2.2
Persons x Raters	32	0.92	1.8
Persons x Methods	32	0.34	0.0
Occasions x Lengths	6	0.30	0.0
Occasions x Raters	3	0.60	0.2
Occasions x Methods	3	0.04	0.0
Lengths x Raters	2	0.09	0.0
Lengths x Methods	2	0.00	0.0
Raters x Methods	1	0.02	0.0
Persons x Occasions x Lengths	192	0.28	7.8
Persons x Occasions x Raters	96	0.41	3.7
Persons x Occasions x Methods	96	0.25	0.7
Persons x Lengths x Raters	64	0.52	0.1
Persons x Lengths x Methods	64	0.01	0.0
Persons x Raters x Methods	32	0.40	2.0
Occasions x Lengths x Raters	6	0.02	0.0
Occasions x Lengths x Methods	6	0.00	0.0
Occasions x Raters x Methods	3	0.18	0.0
Lengths x Raters x Methods	2	0.01	0.0
Persons x Occasions x Lengths x Raters	192	0.04	1.8
Persons x Occasions x Lengths x Methods	192	0.02	0.4
Persons x Occasions x Raters x Methods	96	0.21	8.9
Persons x Lengths x Raters x Methods	64	0.02	0.2
Occasions x Lengths x Raters x Methods	6	0.02	0.0
Persons x Occasions x Lengths x Raters x Methods	192	0.01	1.7

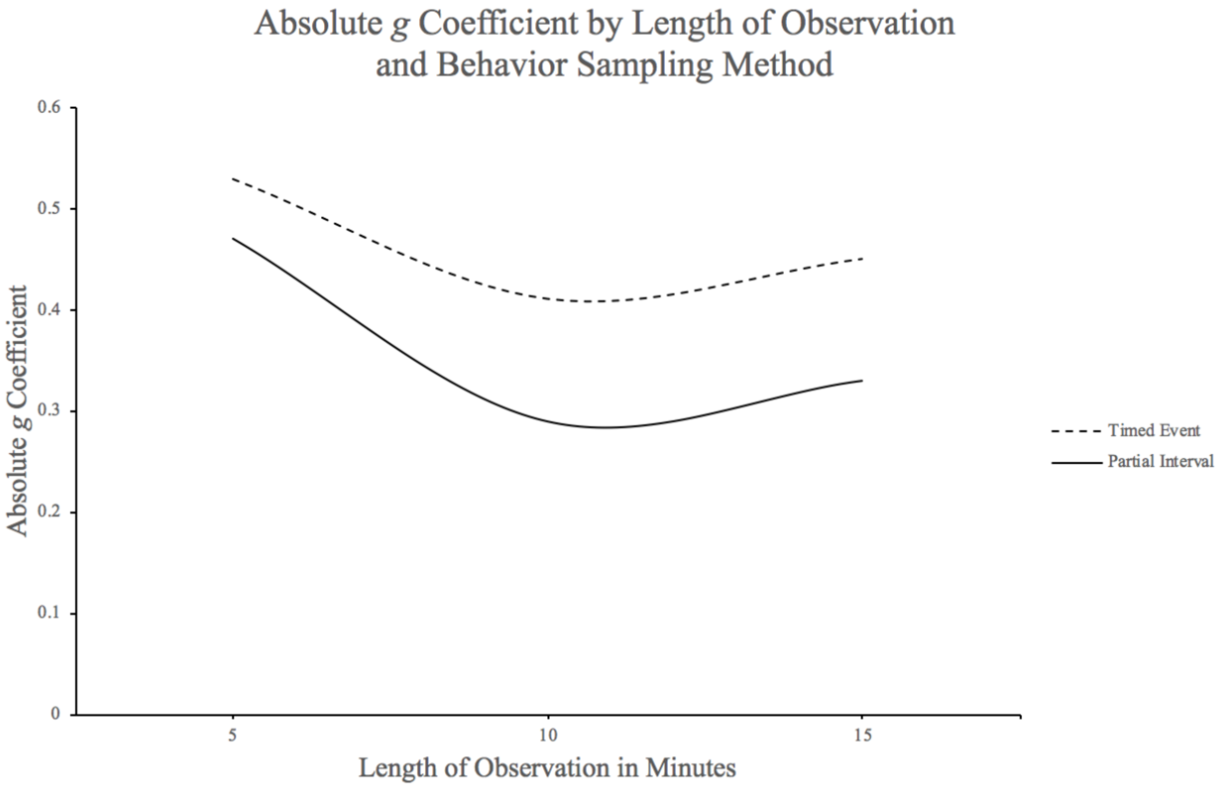

Figure 1. The absolute *g* coefficients from the generalization studies across each combination of observation length and behavior sampling method.

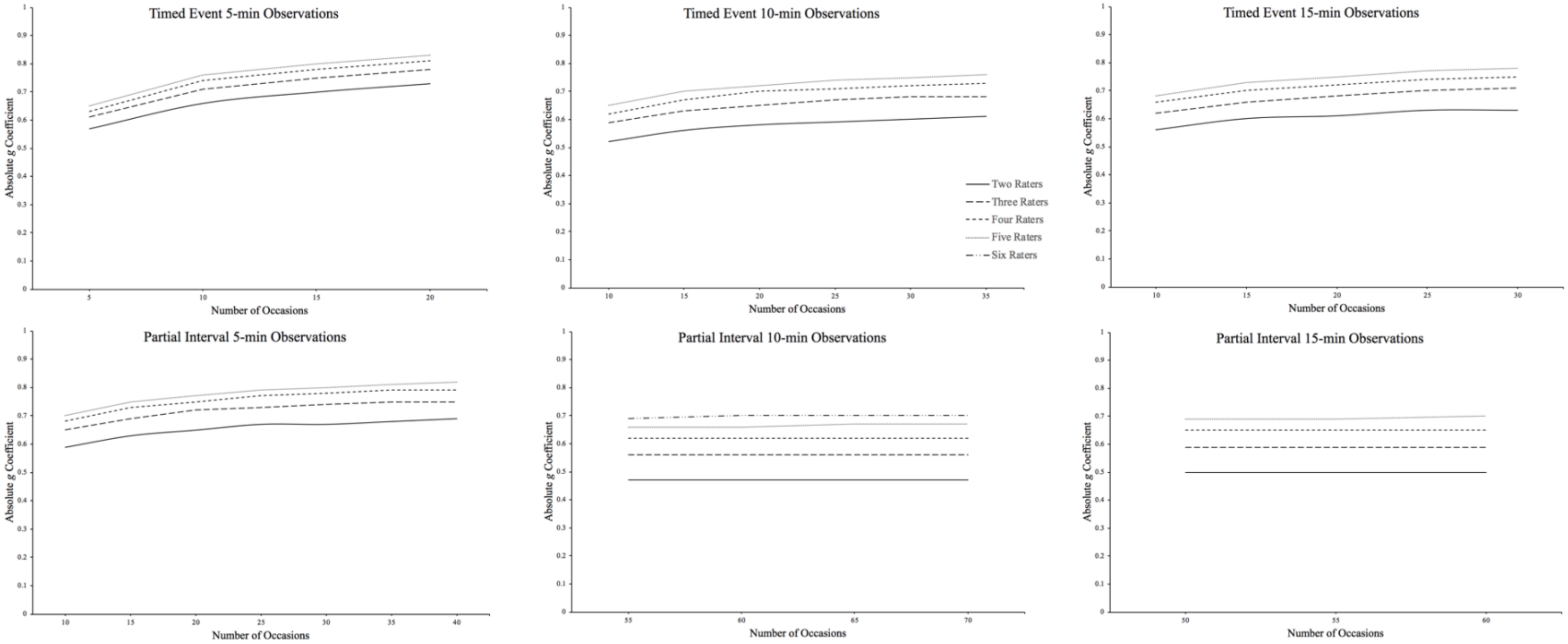

Figure 2. Projected absolute g coefficients from the Decision studies across various levels of four measurement facets: number of occasions, length of observation, number of raters, and behavior sampling method.

REFERENCES

Achenbach, T. M. (1999). The Child Behavior Checklist and related instruments.

Ayres, K., & Ledford, J. R. (2014). Dependent measures and measurement systems. *Single case research methodology: Applications in special education and behavioral sciences, 2*, 124-153.

Barton, E. E., & Fettig, A. (2013). Parent-implemented interventions for young children with disabilities: A review of fidelity features. *Journal of Early Intervention*, *35*(2), 194-219.

Bailey, J. S., & Burch, M. R. (2017). Research methods in applied behavior analysis. Routledge.

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

Bottema-Beutel, K., Kim, S. Y., Crowley, S., Augustine, A., Kecili-Kaysili, B., Feldman, J., & Woynaroski, T. (2019). The stability of joint engagement states in infant siblings of children with and without ASD: Implications for measurement practices. *Autism Research*, *12*(3), 495-504.

Bruckner, C. T., Yoder, P. J., & McWilliam, R. A. (2006). Generalizability and decision studies: An example using conversational language samples. *Journal of Early Intervention, 28*(2), 139-153.

Cardinet, J., Johnson, S., & Pini, G. (2011). Applying generalizability theory using EduG. Taylor & Francis.

Carr, E. G., Innis, J., Blakeley-Smith, A., & Vasdev, S. (2004). Challenging behaviour: Research design and measurement issues. *The International Handbook of Applied Research in Intellectual Disabilities,* 423-441.

Cone, J. D. (1977). The relevance of reliability and validity for behavioral assessment. *Behavior Therapy, 8*(3), 411-426.

Cooper, J. O., Heron, T. E., & Heward, W. L. (2007). Applied behavior analysis.

Cronbach, L. J. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*, 1-33. Wiley.

Cunningham, J. E., Zimmerman, K. N., Ledford, J. R., & Kaiser, A. P. (2019). Comparison of measurement systems for collecting teacher language data in early childhood settings. *Early Childhood Research Quarterly, 49*, 164-174.

Duda, M. A., Clarke, S., Fox, L., & Dunlap, G. (2008). Implementation of positive behavior support with a sibling set in a home environment. *Journal of Early Intervention, 30*(3), 213-236.

Dunlap, G., Ester, T., Langhans, S., & Fox, L. (2006). Functional communication training with toddlers in home environments. *Journal of Early Intervention, 28*, 81-96.

Fiske, K., & Delmolino, L. (2012). Use of discontinuous methods of data collection in behavioral intervention: Guidelines for practitioners. *Behavior Analysis in Practice, 5*(2), 77-81.

Gage, N. A., Prykanowski, D., & Hirn, R. (2014). Increasing reliability of direct observation measurement approaches in emotional and/or behavioral disorders research using generalizability theory. *Behavioral Disorders, 39*(4), 228-244.

Gage, N. A., Han, H., MacSuga-Gage, A. S., Prykanowski, D., & Harvey, A. (2018). A Generalizability Study of a Direct Observation Screening Tool of Teachers’ Classroom Management Skills', Emerging Research and Issues in Behavioral Disabilities (Advances in Learning and Behavioral Disabilities, Volume 30).

Harris, F. C., & Lahey, B. B. (1978). A method for combining occurrence and nonoccurrence interobserver agreement scores. *Journal of Applied Behavior Analysis, 11*(4), 523-527.

Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2013). Single-case intervention research design standards. *Remedial and Special Education, 34*(1), 26-38.

Lane, J. D., & Ledford, J. R. (2014). Using interval-based systems to measure behavior in early childhood special education and early intervention. *Topics in Early Childhood Special Education, 34*(2), 83-93.

Ledford, J. R., Ayres, K. M., Lane, J. D., & Lam, M. F. (2015). Identifying issues and concerns with the use of interval-based systems in single case research using a pilot simulation study. *The Journal of Special Education*, *49*(2), 104-117.

Ledford, J. R., Lane, J. D., & Gast, D. L. (2018). Dependent variables, measurement, and reliability. In *Single case research methodology* (pp. 97-131). Routledge.

Lei, P. W., Smith, M., & Suen, H. K. (2007). The use of generalizability theory to estimate data reliability in single-subject observational research. *Psychology in the Schools, 44*(5), 433-439.

Lloyd, B. P., Weaver, E. S., & Staubitz, J. L. (2016). A review of functional analysis methods conducted in public school classroom settings. *Journal of Behavioral Education, 25*(3), 324-356.

Maggin, D. M., Pustejovsky, J. E., & Johnson, A. H. (2017). A meta-analysis of school-based group contingency interventions for students with challenging behavior: An update. *Remedial and Special Education, 38*(6), 353-370.

Mann, J., Ten Have, T., Plunkett, J. W., & Meisels, S. J. (1991). Time sampling: A methodological critique. *Child Development, 62*(2), 227-241.

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

Pustejovsky, J. E., & Swan, D. M. (2015). Four methods for analyzing partial interval recording data, with application to single-case research. *Multivariate Behavioral Research, 50*(3), 365-380.

Rapp, J. T., Colby-Dirksen, A. M., Michalski, D. N., Carroll, R. A., & Lindenberg, A. M. (2008). Detecting changes in simulated events using partial-interval recording and momentary time sampling. *Behavioral Interventions: Theory & Practice in Residential & Community-Based Clinical Programs, 23*(4), 237-269.

Rapp, J. T., Carroll, R. A., Stangeland, L., Swanson, G., & Higgins, W. J. (2011). A comparison of reliability measures for continuous and discontinuous recording methods: Inflated agreement scores with partial interval recording and momentary time sampling for duration events. *Behavior Modification, 35*(4), 389-402.

Sandbank, M., & Yoder, P. (2014). Measuring representative communication in young children with developmental delay. *Topics in Early Childhood Special Education*, *34*(3), 133-141.

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

Wolery, M., Barton, E. E., & Hine, J. F. (2005). Evolution of applied behavior analysis and treatment of autism. *Exceptionality, 13,* 11-24.

Yoder, P. J., Ledford, J. R., Harbison, A. L., & Tapp, J. T. (2018). Partial-interval estimation of count: Uncorrected and Poisson-corrected error levels. *Journal of Early Intervention*, *40*(1), 39-51.

Yoder, P. J., Lloyd, B. P., & Symons, F. J. (2019) *Observational measurement of behavior* (2nd ed.). Paul H. Brookes Publishing Company.

Yoder, P. Woynaroski, T., & Camarata, S. (2016). Measuring speech comprehensibility in students with Down syndrome. *Journal of Speech, Language, Hearing Research, 59,* 460–467. doi: 10.1044/2015_JSLHR-S-15-0149.