

4-25-2014

Reliability of Data Derived from Time Sampling Methods with Multiple Observation Targets

Austin H. Johnson

University of Connecticut - Storrs, austin.johnson@uconn.edu

Follow this and additional works at: <https://opencommons.uconn.edu/dissertations>

Recommended Citation

Johnson, Austin H., "Reliability of Data Derived from Time Sampling Methods with Multiple Observation Targets" (2014). *Doctoral Dissertations*. 358.

<https://opencommons.uconn.edu/dissertations/358>

Reliability of Data Derived from Time Sampling Methods with Multiple Observation Targets

Austin Hunter Johnson, PhD

University of Connecticut, 2014

Data derived from systematic direct observation procedures are utilized frequently by researchers and practitioners, although the generalizability of these data when derived from distinct measurement decisions has rarely and only selectively been subjected to empirical research. This study utilized generalizability theory to examine the extent to which (a) time-sampling methodology, (b) number of simultaneous behavior targets, and (c) individual raters affect the amount of variance in ratings of academic engagement. Raters with similar levels of advanced training in observation techniques viewed and rated video clips of student behavior within a fully-crossed three-facet design. Results indicated that a majority of variance in ratings was attributable to the object of measurement, whereas very high generalizability and dependability coefficients were observed when ratings were averaged over two raters, regardless of the specific measurement procedure utilized.

Reliability of Data Derived from Time Sampling Methods with Multiple Observation Targets

Austin Hunter Johnson

B.A., University of Arizona, 2007

M.A., University of Connecticut, 2010

A Dissertation

Submitted in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

at the

University of Connecticut

2014

Copyright by
Austin Hunter Johnson

2014

APPROVAL PAGE

Doctor of Philosophy Dissertation

Reliability of Data Derived from Time Sampling Methods with Multiple Observation Targets

Presented by

Austin Hunter Johnson, B. A., M. A.

Major Advisor



Sandra M. Chafouleas

Associate Advisor



Amy M. Briesch

Associate Advisor



Lisa M. H. Sanetti

University of Connecticut
2014

Acknowledgements

I am extraordinarily grateful for the ongoing expertise and kindness of my major advisor, Dr. Sandra Chafouleas. I simply could not have asked for a more thoughtful, encouraging, and above all supportive mentor. I am very grateful to Dr. Daniel Maggin for his encouragement throughout my graduate career, and for his role in shaping my identity as a scholar. I am also indebted to Dr. Amy Briesch for her thoughtful feedback and inspiring research, as well as Drs. Lisa Sanetti and Hariharan Swaminathan for their comments and advice throughout the development of this dissertation. Dr. Melissa Collier-Meek provided crucial support with data entry, and I am profoundly grateful for her advice and friendship.

To my wife Ally, thank you for being who you are. To my Grandma and Grandpa Johnson, for your devotion to our family and your work. To my adopted cohort, for making the past few years so much fun. And finally, to my participants, without whom I would not have a dissertation.

Table of Contents

Abstract page	
Title page	i
Copyright page.....	ii
Approval page.....	iii
Acknowledgments.....	1
Table of Contents.....	2
List of Tables	4
List of Figures.....	5
List of Appendices	6
Introduction.....	7
Psychometrics and behavior assessment.....	8
Systematic direct observation	9
Time-sampling procedure	11
Interval length	15
Observation period.....	16
Operationalizing behavior.....	17
Multiple observation targets	17
Generalizability theory.....	22
GT, time-sampling, and behavior	24
Academic engagement.....	28
Purpose of study.....	31
Method	34

Participants.....	34
Materials	35
Procedures.....	41
Results.....	44
Data preparation.....	44
Analysis.....	45
Discussion	61
Descriptive results.....	62
Research Question 1	62
Research Question 2	64
Research Question 3	67
Research Question 4	69
Limitations	72
Future research.....	75
Conclusion	76
References.....	79

List of Tables

Table 1. Participant characteristics	87
Table 2. Operational definitions of behaviors, with examples and non-examples utilized during training	88
Table 3. Descriptive statistics by clip, methodology, and behavior targets.....	89
Table 4. Variance component estimates for all facets	90
Table 5. Variance component estimates for separate o x r models, by method and behavior	91
Table 6. G and D coefficients for original measurement procedures	91
Table 7. D study of G and D coefficient change as function of number of simultaneous raters ...	92

List of Figures

Figure 1. Distribution of prevalence score density by clip, behavior targets, and methodology...	93
Figure 2. Assignment of clip order, behavior targets, and methodology by participant.....	94
Figure 3. Mean prevalence ratings for academic engagement by clip and methodology	95
Figure 4. Variance component estimates of o x r x m x b as percentage of total variance.....	96
Figure 5. Variance component estimates for o x r models as percentage of total variance, by method and behavior targets	97
Figure 6. D study results for generalizability coefficients, with rater facet manipulated	98
Figure 7. D study results for dependability coefficients, with rater facet manipulated	99

List of Appendices

Appendix A. Sample training binder sheet	100
Appendix B. Sample training data collection form	101
Appendix C. Sample experimental binder sheet.....	102
Appendix D. Sample experimental data collection form, one page	103
Appendix E. Sample experimental data collection form, two pages	104
Appendix F. Participant experience verification form.....	106
Appendix G. Participant consent form.....	107
Appendix H. Training presentation slides	110
Appendix I. Analysis code in R	117

The role that behavioral data take in the work of educators varies, but necessarily undergirds many of the decisions made by both researchers and practitioners. As practices like Functional Behavior Assessments (FBAs) are mandated by federal law for use in schools in certain circumstances (Individuals with Disabilities Education Act Amendments, 1997), and as a continually-growing body of evidence exists to suggest that behavior in schools can affect the impact of academic instruction upon student outcomes (e.g., Sutherland, Lewis-Palmer, Stichter, & Morgan, 2008), it is clear that skills in the selection, design, and use of behavioral assessment tools represent a key competency for educational professionals. This may be especially true for school psychologists, who are tasked with the support of teachers and students in schools using data-based decision-making frameworks (National Association of School Psychologists, 2006).

Behavioral data are used in schools when undertaking FBAs and progress-monitoring the effects of interventions (Chafouleas, Riley-Tillman, & Sugai, 2007), and researchers require behavioral data collection techniques to determine, for instance, if a functional relationship exists between independent and dependent variables in single-case design (Kazdin, 2010), and in the pursuit of the identification of evidence-based interventions for use in schools (Gresham, 2004). Across all types of assessment data, measures may be characterized along a continuum of direct to indirect data collection (Cone, 1977). In the behavioral sphere, these range from indirect tools like behavior checklists and rating scales (e.g., Behavior Assessment System for Children, Second Edition; Reynolds & Kamphaus, 2004) on one end, direct tools like those involving the observation and recording of behavior (Suen & Ary, 1989) on the other, and hybrid measures like Direct Behavior Rating (Chafouleas, 2011) occupying a midpoint between direct observation and a more global estimate of behavior.

A class of direct behavior observation systems titled *systematic direct observation* (SDO) has traditionally been considered the gold standard of behavioral measurement methodologies; this is unsurprising given that instruments which utilize SDO are considered to require little inference from data to target construct (Suen & Ary, 1989). The behavior is observed, often in vivo, and target dimensions of that behavior are recorded by the observer. The defining characteristics of SDO may be described as (a) *systematic* in that the procedures for data collection are rule-bound and replicable, and (b) *direct*, insofar as the behavior is measured as it occurs with little inference required between observation and instance. The fundamental role of SDO is reinforced by data indicating that it continues to be one of the most often used classes of behavior assessment tools by school psychologists (Wilson & Reschly, 1996; Shapiro & Heick, 2004). The popularity of SDO may be unsurprising given that it consists of a wide range of possible methodologies, ranging from the qualitative (e.g., systematically-coded naturalistic observation) to the quantitative.

Psychometrics and Behavior Assessment

With any measurement tool, it is crucial that data derived from an instrument demonstrate adequate psychometric properties, such that the user can have confidence that the information he is collecting is both (a) valid, or allows for meaningful and defensible inference- and decision-making (Cronbach, 1971; Messick, 1995), and (b) reliable, or is able to produce data that are consistent across dimensions of interest (Hintze, 2005).

Reliability and validity are established, albeit continuously evolving, concepts in traditional test and measurement theory. Within the context of behavior, however, the fit of these concepts remains, to some extent, a deeply contentious issue (Silva, 1993). Some authors contend that the constructs of reliability and validity do not apply to behavior assessment, going

so far as to assert in the case of reliability that “‘error’ has no place in behavioral theory” (Hayes, Nelson, & Jarrett, 1986, p. 472). Others insist that both behavioral and traditional assessment share concerns with reliability and validity, and that both concepts should apply to behavioral assessment (Cone, 1977). Silva (1993) assembled a thorough evaluation of both viewpoints, and offered the following recommendations. Validity, as conceptualized in traditional psychometrics, is applicable to behavior assessment without substantial modification; for instance, both behavioral and traditional assessment must be concerned with the degree to which their resulting data reflect only the properties of the target construct. Reliability, on the other hand, is a much more complicated entity. Consensus exists around the fundamental differences between the assumptions involved when considering the stability of a score on a test versus a rating of behavior; the former can, in many cases, be considered as stable over time given certain conditions, whereas the latter should almost always be expected to change with time and repeated measurement. The temporal stability assumption permeates traditional interpretations and indices of reliability, and may therefore need to be reconsidered in the context of behavior assessment. Furthermore, despite the difficulties encountered when directly transporting traditional conceptualizations of reliability to behavior assessment, determinations of “true score variance” and “error variance” are still exceedingly relevant, and therefore some semblance of “reliability” should penetrate the psychometrics of behavior assessment.

Systematic Direct Observation

Although a number of pre-constructed instruments exist which utilize SDO methodologies (e.g., BOSS: Shapiro, 2004; SECOS: Saudargas & Lentz, 1986; see also Volpe, DiPerna, Hintze, & Shapiro, 2007), SDO is unique in that researchers and practitioners have the flexibility to individualize design of the SDO instrument by selecting from a vast array of

observation procedures and data collection parameters in order to build an instrument that will hopefully yield data that represent a target construct. When assessing behavior, utilizing the continuous observation of a target behavior in order to collect data would be ideal, as one would be sampling all behavior from an entire observation period of interest (Suen & Ary, 1989). In such a case, the resolution of the data would only be constrained by one's ability to indicate either the frequency or the start and end points of each behavior with as fine a notation as possible, depending on whether one was interested in the "state" (duration/latency) or "event" (frequency) dimensions of a given behavior (Altmann, 1974). Unfortunately, a number of obstacles exist to undertaking successful continuous observation, one of the most salient of which being that continuous observation becomes "impossible when a number of behaviors are observed simultaneously, the behaviors occur very frequently, and/or more than one subject is being observed" (Suen & Ary, 1989, p. 62).

Given the difficulties inherent in the continuous observation of behavior, time sampling has been offered as a more practical methodology when systematically and directly estimating parameters for a "state" behavior's prevalence or "event" behavior's frequency. Instead of observing at every instant during a behavior stream, one may instead chunk an observation period into intervals of time and use decision rules to determine what should be considered an occurrence or non-occurrence of the target behavior. The resulting data allow for inference-making regarding the properties of the continuous behavior stream, and such a methodology is generally referred to as "time sampling". The procedures for sampling from the continuous behavior stream include partial-interval, whole-interval, and momentary time sampling. In partial-interval sampling (PI), behavior is recorded as present for an interval if it occurs at any point during a given interval. In whole-interval sampling (WI), a behavior is recorded as present

only if it occurs for the entire duration of the interval. Finally, in momentary time sampling (MTS), behavior is recorded as present for an interval only if it occurs at a discrete point in time, typically the start or end point of the time interval.

The dimensions one must consider when designing a time-sampling instrument are numerous, and include: sampling procedure, interval length, observation period length, number of behavior targets and/or observees to observe and record during the period, and operational definitions of those behavior target(s). Determining the optimal choice for any one of these options in a given instrument is a complex task, especially given that the effects of each of these dimensions may not be independent; as discussed below, certain combinations of dimensions may be more appropriate within certain circumstances. Perhaps most crucially, the degree to which dimension choices influence the amount of error attributable to desirable versus undesirable sources is in many cases unclear. As discussed below, researchers have suggested, and in some cases theoretically or empirically tested, the degree to which these decisions impact the reliability of resulting data. Additionally, as discussed next, the complexity of choosing dimensions for use in a time-sampling instrument is made even more difficult by the mixed rigor of recommendations for dimension selection within the behavioral observation literature.

Time-Sampling Procedure

Recommendations exist regarding which time-sampling procedure (i.e., PI, WI, MTS) is most preferable for any given behavior stream, with the rigor of these recommendations ranging from simulation-based research briefs (e.g., Green & Alverson, 1978) to conceptual explorations of a given methodology and how that methodology interacts with the characteristics of a target behavior (e.g., Suen & Ary, 1989; Rogosa & Ghandour, 1991).

History. In 1928, Goodenough published the first article describing the conceptual and experimental underpinnings of what would later be known as one-zero or PI sampling, along with an applied description of this methodology's utility (Suen & Ary, 1984). Citing work by Olson and Parten, Goodenough (1928) provided a description of the evolution of PI sampling from whole-class observations to individual-student observations, as well as a reduction in interval length from 5 to 1 min as observers recorded the prevalence of behaviors such as physical activity, laughter, and conversation as exhibited by nursery-age children. Correlation coefficients were calculated between observers and across time in order to demonstrate consistency across dimensions of interest, with the researcher concluding that this novel method for recording behavior could be considered a potentially-viable method of data collection as it “lends itself to all ordinary forms of statistical treatment, and may be used by persons with only a moderate degree of training, is not excessive in its time-requirements and may be adapted to the study of many different forms of behavior” (Goodenough, 1928, p. 230).

This initial explanation of PI sampling was quickly followed by Olson, who in 1929 published a brief monograph entitled “The Measurement of Nervous Habits in Normal Children.” After an extensive literature review examining various definitions of tic behavior and pilot observations with typical children enrolled in grades K-8, Olson utilized PI sampling with 5 or 10 min intervals in order to estimate the prevalence of tic behavior in students enrolled at the university-based Institute of Child Welfare. The reliability and validity of the resulting data were subsequently examined, chiefly using data collected from a second observer, through the calculation of traditional reliability statistics and correlation matrices. Olson ended his treatment of this new methodology with the prescient acknowledgment that the interval length and number of observations required to utilized PI sampling (referred to as simply “time sampling” in the

foreword) should vary based on the type and frequency of a given target behavior, as well as the reliability of the data themselves.

After the establishment of the first of the three major time-sampling methodologies, an adaptation of more general time-sampling procedures was made in order to systematically observe behavior in rats (Bindra & Blond, 1958) using what would later be referred to as MTS. In this new form of time-sampling, behavior was categorically coded immediately onto the data sheet, rather than written descriptively for later interpretation, at the moment each interval began. Finally, in the 1970s, references to WI sampling procedures began to appear in the literature in a study of positive reinforcement and its effects on student on-task behavior (Peterson, Cox, & Bijou, 1971), as well as an examination of the effects of different types of instructional settings on the study behavior of undergraduate students (Born & Davis, 1974). In an effort to both unify discussions of these three time-sampling methodologies as an overarching system of time-sampling, as well as provide consistent terminology when discussing time-sampling, Powell, Martindale, and Kulp (1975) suggested the use of “whole interval time-sampling,” “partial interval time-sampling,” and “momentary time-sampling” as the three overarching terms to describe these procedures (p. 463).

Selection. Crucial to the appropriate selection of any one procedure over another is a consideration of the characteristics of the behavior stream itself. Although early research recognized the importance of behavior characteristics (e.g., latency between instances of behavior) and other measurement circumstances (e.g., rater experience) to the reliability and validity of the resulting data (Arrington, 1943), empirical comparisons of time-sampling methodologies were not initiated until a few decades later. Many of these studies utilized comparisons of continuous observation and time-sampling or simulation studies to determine the

relative level of accuracy displayed by the sampling method (e.g., Powell, Martindale, & Kulp, 1975; Harrop & Daniels, 1986). Prior to the introduction of theoretical formulae to guide the selection of time-sampling procedures, researchers generally concluded that: (a) PI sampling will provide an overestimate of a behavior's true prevalence and an underestimate of its true frequency; (b) WI time-sampling will provide an underestimate of prevalence and overestimate of frequency; and (c) MTS will provide the most accurate estimate of prevalence (Suen & Ary, 1989; Salvia & Hughes, 1990). However, in summaries of empirical time-sampling evaluations, researchers have disagreed on the appropriate use of MTS in frequency estimation.

In their comprehensive book on quantitative behavioral observation, Suen and Ary (1989) codified the findings of multiple previous papers (e.g., Suen & Ary, 1984; 1986) into an overarching framework for determining whether utilizing a WI, PI, or MTS methodology may provide an unbiased estimate of a behavior's true frequency and a bias-corrected estimate for prevalence or duration (depending on the desired unit of analysis). First, in order for an accurate frequency count to be derived from the results of a given time-sampling methodology, three methodological and behavioral characteristics must be determined: (a) the interval length utilized during time-sampling, (b) the shortest amount of time the behavior occurred during one instance of the behavior (or "bout" length), and (c) the shortest amount of time that elapsed between instances of behavior (or "interresponse time duration"; Suen & Ary, 1989, p. 67). If a given time-sampling procedure's interval length is less than (a) the bout length times a methodology-specific coefficient (.5 or 1.0) and (b) the interresponse time duration times a second methodology-specific coefficient (.5 or 1.0), then an accurate frequency count can be derived from the resulting data.

After deriving an accurate estimate of frequency, a bias-corrected estimate of duration or prevalence may be derived by calculating the sum of (a) the number of intervals scored as an occurrence of the target behavior and (b) the frequency count times a methodology-specific coefficient (0, 1, or -1). This value can then be multiplied by the interval length to provide a bias-corrected estimate of duration, or divided by the total number of observation intervals to provide a bias-corrected estimate of prevalence.

As is underscored by these estimation conditions, interval length plays a critical role in the reduction of systematic error when utilizing time-sampling procedures. However, recommendations for suitable interval lengths for individual behaviors are mixed across the literature.

Interval Length

Outside of the equation-derived recommendations of Suen and Ary (1989), the justification for various interval length recommendations range from applied studies of the concordance between data yielded from time sampling procedures and continuous observation (Saudargas & Zanolli, 1990) to more general recommendations for pre-constructed instruments. For instance, Saudargas and Lentz (1986) recommended a 30 s interval length for use with their State-Event Classroom Observation System (SECOS) which uses frequency counts and MTS, but made their recommendation without cited empirical justification and went on to state that “shorter intervals could certainly be used” (p. 43).

Multiple researchers have expressed a preference towards shorter intervals in order to provide more accurate estimates of behavior (e.g., Powell, Martindale, & Kulp, 1975; Sanson-Fisher, Poole, & Dunn, 1980). Simulation studies have suggested that shorter interval lengths will provide more accurate estimates using MTS procedures (Kearns, Edwards, & Tingstrom,

1990), and live simultaneous-recording procedures have suggested similar results, with particularly high levels of accuracy when observing general student academic behaviors (e.g., reading, listening) for intervals of 10 and 20 s in length (Brulle & Repp, 1984). In their oft-cited book on applied behavior analysis, Cooper, Heron, and Heward (2007) review the literature regarding bias in estimates made from MTS procedures as compared to estimates derived from continuous observation, and conclude that MTS procedures utilizing intervals of less than 2 min in length have been demonstrated to be highly concordant with continuous observation data as compared to those using intervals longer than 2 min, and that such an interval length may generally provide estimates of behavior prevalence with little bias. Clearly, differing opinions exist on just how fine an interval must be in order to derive data representative of a continuous behavior stream.

Observation Period

When developing time-sampling procedures, an interval length must be applied to a given observation period (e.g., sampling in 15 s intervals for 10 min), and with that chosen observation period come a number of additional considerations. As Rogosa and Ghandour (1991) observed, finite observation length is a critical component of data's representativeness, given that the data generated from a sampled observation period will often be generalized to represent a larger unobserved behavior stream; as the authors demonstrate, failure to account for this finite observation length may significantly bias resulting estimates of overall behavior. The choice of *how* to select an observation period may be anchored in the extent to which one desires to generalize a given sample of behavior. Data collected from an observation period that was randomly chosen from a greater length of time (e.g., a 15 min observation chosen randomly during the school day) may generalize to that greater length of time (e.g., the entire school day)

given enough randomly chosen observations (Suen & Ary, 1989), whereas data derived from a deliberately-chosen observation period (e.g., the first 20 min of math class) may not support inferences to behavior outside that period. The use of random selection when identifying observation periods may also have significant implications in later reliability analyses, depending on the way in which time as a dimension is treated (e.g., as a random facet). Random sampling of observation periods from a total potential body of periods is in many ways ideal; however, the choice of that larger body of periods, as well as their length, may be intentional and driven by both the research question and methodological feasibility (Suen & Ary, 1989). With that said, aside from interval length, observation length may serve as one of the most readily-manipulable dimensions that may impact the generalizability of estimates derived from a given data stream.

Operationalizing Behavior

In order for a behavior to be observed, relevant aspects of the target behavior or behaviors must be defined with an operational definition of the target behavior (Salvia, Ysseldyke, & Bolt, 2010), often accompanied by examples and non-examples of the topography of the behavior (Chafouleas, Riley-Tillman, & Sugai, 2007). From an empirical perspective, the common use of a behavioral operationalization may also aid in the facilitation of cross-study comparisons. For instance, in their studies of error in observation instrument data, Hintze and Matthews (2004) as well as Briesch, Chafouleas, and Riley-Tillman (2010) both utilized operational definitions provided by the Behavioral Observation of Students in Schools (BOSS: Shapiro, 2004) system, allowing for some measure of comparison between the two studies.

Multiple Observation Targets

Although recommendations exist regarding how best to systematically observe multiple students across intervals (Thomson, Holmberg, & Baer, 1973), little to no empirical

recommendations exist pertaining to how many behaviors can be observed at one time without affecting the psychometric properties of the resulting data. Nevertheless, the literature is replete with examples of pre-constructed SDO instruments wherein multiple behaviors are observed during a single interval. The State-Event Classroom Observation System (SECOS; Saudargas & Lentz, 1986), for instance, utilizes MTS to concurrently observe the prevalence of eight distinct behaviors, and frequency counts to measure the rate of occurrence for 11 other behaviors. The BOSS (Shapiro, 2004), on the other hand, utilizes MTS to record whether the student was engaged in one of two mutually-exclusive engagement behaviors, and PI to record if any of three off-task behaviors occurred during the interval. Evidence exists to support the suggestion that moderate-to-high levels of inter-rater reliability can be achieved using the BOSS and the SECOS (Volpe, DiPerna, Hintze, & Shapiro, 2007), which is one component of the overarching reliability of data from a direct observation instrument (Hintze, 2005). However, it is unclear to what extent and in what manner the simultaneous measurement of multiple behaviors contributes to the introduction of error in the resulting data, and how this addition of complexity varies by time-sampling method.

In discussing dimensions of observation protocols that may influence the relative accuracy of resulting data, some researchers have subsumed the concept of multiple observation targets under the general term “coding complexity” (Harris & Lahey, 1982). This concept articulates multiple observation targets according to the (a) overall number of categories within the system, (b) number of categories scored simultaneously, (c) overall number of subjects observed, and (d) number of subjects observed simultaneously; thus, within this conceptualization, complexity may be attributed to both the total number of behaviors and subjects observed, as well as how many behaviors and subjects are observed at a time. In order to

guard against inaccurate or unreliable data resulting from any added coding complexity, emphasis has been placed on the importance of providing training to raters across all levels of complexity that they may encounter when rating behavior (Harris & Lahey, 1982).

The range of studies examining the impact of multiple observation targets on direct observation data is limited, and the existing literature provides mixed results and recommendations regarding the effects of added coding complexity. Mash and McElwee (1972) utilized a contrived rating situation in order to determine the effect of (a) number of behavior code categories, (b) the predictability of the behavior, and (c) prior experience with predictable or unpredictable behavior on the concordance of a rater's audiotape codes with those of a master coder. Every 3 s, a verbal statement was presented, and the rater coded the statement as one of four or eight categories of statement, depending on the participant's assigned condition. Results suggested that the use of a four-code system resulted in greater observed accuracy than an eight-code system. Although these findings provide preliminary evidence for an accuracy effect as a function of the number of possible targets, this study utilized audiotapes rather than direct observation, and did not utilize a time-sampling technique but rather a single summative rating in response to each stimulus. As such, it is difficult to compare the results from this study with those that may be expected from use of a time-sampling instrument.

In a direct examination of coding complexity and time-sampling procedures, Frame (1979) examined the effects of direct observation protocols with varying complexity on resulting interobserver agreement using PI time sampling. Six observers were randomly assigned into three pairs, who then simultaneously observed and rated videotaped classroom behavior of approximately ten students who were identified as "emotionally impaired." In order to train participants on the procedures utilized within this study, participants were first quizzed on the

definitions of each of the 15 behaviors they would be asked to rate over the course of the study. Pending successful completion of the quiz, participants then completed interobserver agreement trials to criterion with their randomly-assigned partner. For these trials, participants independently rated a single student on a single behavior, using 10 s intervals over a 10 min observation period (although every sixth interval was considered a “rest interval” and not scored by the observer). If raters did not demonstrate 70% agreement using the traditional percent-agreement index (number of agreements divided by the sum of number of agreements and disagreements) for a given behavior, they participated in an additional set of trials for that behavior using novel 5-min-long videos. During these shortened periods, observers were given the opportunity to pause, discuss, and review behavior depicted on the videos until they demonstrated 70% agreement. After demonstrating mastery of each target behavior, experimental procedures began wherein rater pairs observed and coded novel 15 min clips of student behavior in six-clip sessions. Clips were coded with the use of protocols depicting one, four, six, nine, eleven, or fourteen simultaneous target behaviors. Behaviors were purposefully assigned across conditions of simultaneous target behaviors, and analyses were conducted using both mean IOA values across behaviors and separate ANOVA analyses for each balanced set of behavior observations.

In contrast to the findings of Mash and McElwee (1972), results suggested that only minimal differences were present across protocols with differing levels of behavior-target complexity. Significance tests took the form of quasi-*F* tests, wherein the sum of the mean sum of squares (MS) for the simultaneous-behaviors condition and the MS for the residual term was divided by the sum of the MS for the interaction terms. The authors concluded that the results of

these tests indicated no significant difference in agreement scores across simultaneous-behaviors conditions.

Although these results bear directly upon time-sampling psychometrics, a number of potential limitations are present within this study. First, the training procedures utilized were wholly targeted towards increasing inter-rater reliability; in other words, raters were trained to agree with one another as to whether behavior occurred (and not, interestingly, if it was a nonoccurrence), rather than with an external criterion. As a result, raters were provided with 10 to 12 hr of experience (less introductory and quiz time) in achieving high levels of interobserver agreement, prior to engaging a study that examined their levels of interobserver agreement (Frame, 1979). Second, the non-random partitioning of ANOVAs and subsequent significance tests is a potentially serious concern, as results may have differed if alternate sets of behaviors were grouped together for analyses. Third, only PI sampling was utilized for this study, and it is unclear if these results would generalize to other time-sampling methodologies. Finally, given the reliance on inter-observer agreement data, it was unclear whether raters were uniformly accurate as complexity increased or if their agreement was a result of decreased accuracy that was maintained across both raters (Harris & Lahey, 1982).

As has been described at length, a number of dimensions must be taken into consideration when determining how to assemble and evaluate a time-sampling instrument. However, the rigor and content of recommendations regarding ideal dimension selection varies. A strong case can be made to support the notion that issues of reliability and validity impact data derived from behavioral assessment instruments (Silva, 1993), and that the choices made when designing such an instrument may impact the amount of error present within the resulting data (Suen & Ary, 1989). How, then, may users of SDO instrumentation best determine the relative

impact that these decisions have on the reliability and validity of data derived from time-sampling instruments?

Generalizability Theory

Generalizability theory (GT), first fully described by Cronbach, Gleser, Nanda, and Rajaratnam (1972), provides a method for identifying and quantifying multiple sources of variance in measurement data. Prior to the introduction of GT, traditional conceptualizations of reliability had utilized a classical test theory (CTT) model, wherein an observed score is defined as the sum of a true score and an error term, with the error term treated as a unidimensional, monolithic entity (Crocker & Algina, 2008). As a result, reliability coefficients in CTT capture different sources of error depending on the comparison being made (e.g., test-retest vs. split-half). For decades, psychometricians had recognized that score and error variance are derived from multiple sources (e.g., Thorndike, 1947) and that a determination of which sources of variance are of interest should be founded directly within what questions are trying to be answered with the data from a given measurement instrument.

The foundation of GT lies in the concept of universe scores, and the universes in which those scores operate. When an investigator uses observed scores as data in the decision-making process, “the score on which the decision is to be based is only one of many scores that might serve the same purpose” (Cronbach, Gleser, Nanda, & Rajaratnam, 1972, p. 15). The investigator probably does not want to use that score as a representation of that exact response at that exact moment to that exact rater using that exact procedure, but may rather want to use that score as something representative over a given set of conditions (e.g., a score that would be generated by any rater who was rating any number of behaviors at a time). As such, the universe score in GT is defined as the mean score that would be generated over all observations consisting of

conditions deemed admissible by the researcher, whereas a true score in CTT is generally considered to be the mean generated from an infinite number of measurements of a given target.

By identifying the conditions over which scores will be generalized, the facets of measurement in GT are defined within which alterations could occur and still generate a score that the investigator would consider to be meaningful (Brennan, 2001). For example, a researcher may wish to consider a score generated by any rater who was rating any number of behaviors at a time as representative of that student's behavior. Both (a) raters and (b) the number of simultaneous behavior targets would in this case constitute the two facets of interest, with specific types of raters and numbers of behaviors simultaneously observed referred to as conditions. After making multiple observations along different levels of each facet or condition of measurement, a researcher could examine the consistency of scores within each condition and in that way estimate the relative level of variance attributable to each facet using an analysis of variance procedure; this process of estimation is referred to as a G study. The researcher may have a number of different assemblages of facets and conditions in mind under which a measurement might be taken and considered generalizable (e.g., raters, settings). The specific collection of facets within the universe of admissible observations is called the universe of generalization, and the process of estimating the degree to which one can generalize from the obtained scores is referred to as a D study. In this way, results of a D study can extrapolate the generalizability of a given score using any assemblage of facets specified in the model.

Within GT, two major indices of reliability are derived from partitioned variance components: these include one index for relative decision-making, called a "generalizability coefficient" or $E\rho^2$ (Cronbach et al., 1972), and another for absolute decision-making, known as a "dependability coefficient" or Φ (Brennan & Kane, 1977). The generalizability coefficient

represents the ratio of (a) the variance attributable to the object of measurement to (b) the sum of the variance attributable to the object of measurement and all interactions with the object of measurement. When decisions are made based on an individual's relative rank when compared to other students, and not to an external standard, decision-makers may not wish to consider sources of error that are not based on the object of measurement itself. Thus, a coefficient that only considers variance that is influenced by the object of measurement (e.g., the person) may more adequately represent the desired reliability characteristics of those data.

The dependability coefficient represents a similar ratio to that of the generalizability coefficient; however, whereas relative decision-making may be most adequately represented by sources of variance affected by the object of measurement, absolute decision-making may be appropriately influenced by all sources of variance, including those not directly attributed to the object of measurement. As a result, the dependability coefficient represents the ratio of (a) the variance attributable to the object of measurement to (b) the sum of the variance attributable to the object of measurement and *all other sources of variance*.

Using the generalizability and dependability coefficients, researchers can not only generate indices of reliability for a given set of observations, but also project the reliability of data for distinct assemblages of facets. Thus, although a G study may estimate the reliability of a score averaged over five randomly-selected raters, future research could potentially utilize fewer raters; a D study and its subsequent generalizability and dependability coefficients may provide information regarding how many raters would be required to achieve adequate levels of reliability for relative and absolute decision-making, respectively.

GT, Time-sampling, and Behavior

Just as the design of a time-sampling instrument requires the identification of multiple design dimensions (e.g., observation length, interval length, specific methodology utilized), GT emphasizes the analysis of individual facets (i.e., sources of variance) within a measurement context. When a researcher or practitioner makes decisions regarding how to assemble his time-sampling instrument, he is essentially selecting from individual conditions (e.g., MTS) within larger facets (e.g., time-sampling procedure). Ideally, he is making those decisions that will minimize the amount of error present within his data, such that the data reflect information representative of the object of measurement (e.g., students), rather than, for instance, the raters and how they used the instrument. Therefore, when designing a time-sampling instrument within a GT framework, one might ask: what facets might be expected to contribute to the overall variance of a given score, and which of those facets should be targeted as possible dimensions of generalization?

Cone (1977) identified six facets that may be considered most relevant to the analysis of behavioral observation in a GT framework: (a) scorer, (b) item, (c) time, (d) setting, (e) method, and (f) dimension. Researchers and practitioners may wish to generalize the data derived using their instrument's specific collection of dimensions to those derived using other conditions within these identified facets; rather than limiting the interpretation of their data to "information collected by Rater X on Student Y's disruptive behavior using MTS during a 30 min observation period using 15 s intervals," a researcher or practitioner may instead wish to generalize their data to an overall estimate of the prevalence of disruptive behavior by Student Y. However, as discussed earlier, it is unclear to what extent such a generalization can be fairly made.

Applying GT to time-sampling. To this end, researchers have explored the utility of GT in estimating sources of variance in data derived from time-sampling instruments, although the

number of studies utilizing this technique is limited, and the facets examined within and across studies are few. In the first identified utilization of GT with time-sampling instrumentation, Marcus, Johnson, and Roke (1980) utilized unspecified time sampling procedures with 10 s intervals to measure the cooperative behavior of preschoolers during free play. These data were analyzed within a one-facet design, with preschoolers as the object of measurement and occasion as a random facet. For the first of two studies, researchers found 30.2% of the variance in scores to be attributable to subjects, 0.3% to occasions, and 69.5% as unexplained residual variance.

McWilliam and Ware (1994) utilized a two-facet fully-crossed design, with persons as the object of measurement crossed with observation sessions and raters. Significantly, separate G studies were conducted for three different observation lengths during 15 min video clips of student behavior, with results presented for data from the first 5 min, the first 10 min, and the entire 15 min observation length. Data were collected by trained raters as they independently viewed the video clips (which they could pause and rewind as desired at the end of each interval), using MTS with 10 s intervals to simultaneously measure the prevalence of two dimensions of engagement behaviors: type (e.g., with adults, with peers) and level (e.g., attentional, symbolic) for one student. Results from the G study for the entire 15 min observation period indicated that interaction between subjects and sessions contributed the largest percentage of total variance for seven of nine observed behaviors ($M = 55.3\%$, range = 17.6%-78.1%).

In a more recent application of GT to data derived using time-sampling, Hintze and Matthews (2004) similarly utilized a fully-crossed two-facet design, with persons as the object of measurement and time and setting as facets. Data were collected pertaining to the prevalence of on-task/off-task behavior, considered mutually exclusive in the instrument, using MTS procedures with 15 s intervals across a 15 min observation period as one student was rated at a

time. G study results indicated that, for their study, the object of measurement contributed the largest percentage of variance (62%), followed by the residual term (24%), the person by setting interaction (13%), and the time facet (1%).

A three-facet partially-nested design was utilized by Briesch et al. (2010) using persons as the object of measurement and day, occasion nested within day, and rater as facets. Separate G studies were conducted for teacher raters and researcher raters in order to determine if differential variance component contributions were observed by rater type. Similar to the procedures of Hintze and Matthews (2004), raters in this study utilized MTS with a 15 s interval during observation periods that ranged from 10-15 min as they rated academically engaged behavior. Four students were observed during each interval, and data were simultaneously recorded for all four students at the beginning of each interval. Results indicated that the object of measurement demonstrated the largest single variance component (47 to 48%), with a rater effect and student by rater interaction observed for teacher raters, but not researcher raters.

Most recently, the effect of observation length upon the subsequent dependability of data derived from MTS procedures with a 15 s interval was investigated (Ferguson, Briesch, Volpe, & Daniels, 2012). Two graduate students coded the behavior of 20 students for 30 min each over the course of two days, with each 30 min observation divided into six 5 min “blocks”. A fully-crossed two-facet model was utilized in order to derive variance component estimates and calculate dependability and generalizability coefficients, with person as the object of measurement ($n = 20$), and block ($n = 6$) and day ($n = 2$) as random facets of interest. Contrary to the results of earlier, similar studies, results for the two-facet model suggested that the residual term comprised the majority of variance (50%), followed by person (29%) and the person x occasion interaction (15%). Notably, no variance was attributed to the occasion or block facets.

D-study results essentially described over how many 5 min observation periods a rating of academic engagement would need to be averaged in order to derive an adequate dependability coefficient ($\Phi = 0.70$), with results suggesting that a few longer observations (two 30 min observation) or more observations of shorter length (four 10 min observations) would be required in order to observe a dependability coefficient of 0.70.

Across these previously-conducted studies of time-sampling data using GT, certain themes in procedures and analyses emerge. Most of these studies limited their conclusions to data derived from MTS procedures; neither PI nor WI methodologies were examined. Facets investigated across these studies consisted of sessions, raters, time/day, and setting, and person (or target student) was consistently the object of measurement. Thus, although these studies shed light on three of Cone's (1977) identified facets of interest, another facet, method, has yet to be systematically analyzed. As discussed earlier, decisions related to methodology have been demonstrated to have serious implications for the validity and reliability of data obtained via behavioral observation, and the interactions among these methodological decisions has yet to be explored. This investigation is critical given the widespread use of methodologies, behaviors, and simultaneous behavior targets in pre-constructed time-sampling instruments. Clearly, more research is needed in order to determine the extent to which these methodological decisions contribute to the reliability and validity of data derived from instruments that utilize time-sampling to measure behavior.

Academic Engagement

Students engage in a variety of behaviors that facilitate access to instruction and their subsequent ability to demonstrate academic achievement (DiPerna & Elliott, 2002). These behaviors, often referred to as academic enablers, complement students' academic skills and are

hypothesized to include constructs such as interpersonal skills, motivation, study skills, and engagement. Academically engaged behavior, which is typically characterized as a class of behaviors consisting of both active (e.g., verbally responding, writing) and passive (e.g., silent reading, listening) activities (Shapiro, 2004), has been identified throughout the literature as possessing a significant relationship with academic achievement (Greenwood, Horton, & Utley, 2002).

Multiple researchers have modeled the relationship between academically engaged behavior and academic achievement, with results generally supportive of the effects of engagement on subsequent achievement. In one study, 104 teachers rated 394 students in Grades K-6 using the Academic Competency Evaluation Scales (ACES; DiPerna & Elliott, 2000), a rating scale that has previously demonstrated data with psychometric adequacy (DiPerna, Volpe, & Elliott, 2002). Teachers randomly selected no more than five students from their class rosters, and rated 81 items on a 5-point scale relating to students' relative levels of academic proficiency and the frequency with which students engaged in behaviors identified as academic enablers. Structural equation modeling was utilized to test the relationship between hypothesized academic enablers and reading achievement, with results suggesting that engagement behaviors demonstrate a "large" impact on reading achievement for students in Kindergarten to Grade 2, and a "moderate" impact for students Grades 3-6 (DiPerna & Elliott, 2000).

In another investigation of the relationship between achievement and engagement, 53 students were observed longitudinally across Grades 1-3 using a pre-constructed protocol of ecological and student-behavioral variables (Greenwood & Terry, 1994). These variables were combined post hoc into exposure, task quality, and engagement composite ratings, and were subsequently submitted for structural equation modeling. Results suggested that a model in

which academically engaged behavior mediated the effect of instruction (comprised of both exposure to instruction and task quality) provided superior fit to models wherein either (a) achievement was impacted by instruction and engagement simultaneously, or (b) engagement, task quality, and exposure to instruction were all modeled as simultaneous direct effects.

Given its hypothesized role as a key variable in academic achievement, it is unsurprising that multiple prior generalizability analyses of behavioral observation procedures have utilized academic engagement as their dependent variable. McWilliam and Ware (1994) measured engagement behavior in their study of 20-65 month-old children, although engagement behavior in this study was broadly defined to match the developmental level of study participants. In their generalizability analysis of SDO procedures, Hintze and Matthews (2004) utilized a modified definition of academic engagement, termed “on-task” behavior, which was derived from the operationalization of active and passive engagement utilized by the Behavioral Observation of Students in Schools (BOSS) form (Shapiro, 2004). A similar definition was utilized by Briesch et al. (2010), wherein academically engaged behavior served as the dependent variable and was operationalized as active or passive classroom participation; this modified operational definition was also derived from the operational definition of academic engagement found within the BOSS (Shapiro, 2004). This modified definition was also utilized in a study of MTS procedures (Ferguson et al., 2012).

Given both (a) strong support for the primary role of academically engaged behavior in promoting academic achievement, and (b) the historic use of this behavior as a dependent variable in prior GT studies, the use of academic engagement as a dependent variable appears defensible for an investigation of time-sampling instrumentation for two reasons. First, the results of such a study may more effectively generalize to other examinations of behavior

observation using GT if a common operational definition for the dependent variable is utilized (insofar as this is appropriate, given the recommendations of Lei, Smith, and Suen, 2007).

Second, the previously-identified relationship between this behavior and academic achievement suggests that psychometric research that utilizes academic engagement as a dependent variable may prove particularly informative to researchers and practitioners who seek to improve student outcomes.

Purpose of Study

Designing a time-sampling instrument requires the specification of a number of parameters, or in the language of GT, facets. Although the relative validity of certain assemblages of facets has been postulated, it is unclear whether these assemblages change the amount of variance attributable to desirable versus undesirable sources, and therefore the psychometric quality of the instrument's resulting data. Given the lack of understanding in the literature of how these instruments differentially function, the purpose of this study is to apply the principles of GT to time-sampling instruments that vary based on (a) method (MTS, WI, PI), and (b) number of behaviors observed simultaneously (1, 3, 5, 7) across raters and occasions. The research questions posed by such a study are as follows:

1. Using raters with similar levels of prior training, what proportion of overall variance in academic engagement ratings is attributable to time-sampling methodology?

It has been demonstrated that the application of distinct time-sampling methodologies to the same behavior stream will result in differing prevalence estimates in the presence of mixed intervals (Suen & Ary, 1989). Furthermore, prior investigations of variance in time-sampling-derived ratings using GT have generally demonstrated large proportions of variance attributable to the object of measurement (e.g., Hintze & Matthews, 2004;

Briesch, Chafouleas, & Riley-Tillman, 2010). Given the possibility for large differences in estimates based upon the application of distinct methodologies (Suen & Ary, 1989), it is hypothesized that a majority of rating variance will be attributable to time-sampling methodology, followed by variance attributable to the occasion facet.

2. Is any variance in academic engagement estimates attributable to an interaction between the number of behaviors rated and the type of time-sampling procedure used? In other words, are systematic differences observed in estimates of academic engagement depending on the combination of (a) number of simultaneous behaviors observed and (b) time-sampling methodology utilized? Is any variance attributable to other facets or interactions?

It is hypothesized that greater than 6.67% of total rating variance will be attributable to a methodology by behavior interaction. If variance was equally divided among all 15 estimated variance components generated using a full three-facet model, each component would contribute 6.67% of the overall variance; thus, this hypothesis suggests that more than an equally-distributed amount of variance will be attributable to the methodology by behavior interaction. Time-sampling methodologies vary in the degree to which raters must attend to behavior during the interval: for an instant when using MTS, and for the whole interval when using PI and WI. Given that these methodologies require differing levels of attendance from raters, they may also systematically differ from one another depending on the number of behaviors observed simultaneously.

3. Within each time-sampling methodology, does rating additional behaviors simultaneously affect the amount of variance in academic engagement ratings attributable

to facets other than the object of measurement? If so, which methodologies are least susceptible to increasing error, and which are most susceptible?

It is hypothesized that, across methodologies, rating additional behaviors simultaneously will result in incremental decreases in the proportion of variance attributable to the object of measurement. It is further hypothesized that the percentage of variance attributable to the object of measurement will be highest for PI recording, followed by MTS, and lowest for WI recording. Raters may be able to most consistently rate academic engagement using PI recording regardless of the number of behaviors observed, given that any instance of the behavior during the interval will result in marking the interval as an occurrence. With MTS recording, raters are asked to simultaneously rate all target behaviors at once, whereas in WI recording, raters must attend to all target behaviors for the duration of the interval. Given the differing demands placed upon raters across methodologies, ratings conducted with increasing amounts of simultaneous behavior targets may become increasingly and differentially less consistent.

4. How many raters would need to simultaneously rate a student in order to achieve a generalizability and dependability coefficient of at least 0.90? How does this vary across time-sampling procedure and simultaneous behavior target combinations?

It is hypothesized that more than one rater will need to simultaneously rate a student in order to demonstrate a generalizability and dependability coefficient above 0.90.

Generalizability coefficient values for one-facet rater models will be dependent upon the relative size of variance attributable to the object of measurement and the residual term, as well as the number of raters over which an estimate of academic engagement is averaged. In order for a generalizability coefficient to demonstrate a value of 0.90 with

one rater, the variance component for the object of measurement would need to be nine times the size of the variance component for the residual term. Although procedures in this study vary greatly from previous generalizability analyses of SDO-derived ratings, observed ratios between the object of measurement and the residual term for estimates of academic engagement in previous studies have ranged from 2.58:1 (Hintze & Matthews, 2004), 1.09:1 (Briesch, Chafouleas, & Riley-Tillman, 2010), and 0.57:1 (Ferguson et al., 2012). Dependability coefficients require an even greater amount of variance to be attributable to the object of measurement in order for high values to result. Given that most previous studies have not demonstrated ratios of observed to residual variance to the extent necessary for a reliability-like coefficient of 0.90 without averaging over additional facets, it is expected that similar results will be demonstrated within the current study. Similar to the hypothesis in the third research question, more raters are expected to be necessary for (ranked from more to less): WI recording, MTS, and PI sampling, with more raters required for higher numbers of simultaneous behavior targets.

Method

Participants

Participants in this study consisted of 10 graduate students enrolled in a school psychology program at a large public university located in the Northeastern United States. All participants had completed prior coursework in behavior assessment, including training in systematic direct observation protocols. Seven of ten participants (70%) were enrolled in a doctoral-level program, with the remainder enrolled in a Masters/6th-Year program. All participants had previously used MTS in research- or practice-based observations, with an average of 58.2 hours of previous observations conducted using SDO-based procedures.

Additional detail regarding participant characteristics is provided in Table 1. Participants were recruited for participation via email, with participants selected according to schedule availability. Participants were provided with digital copies of consent forms prior to arriving for formal research procedures, and provided written consent to participate in this study in person as witnessed by the principal investigator (see Participant Consent Form, Appendix G).

Materials

Video clips. The video clips utilized in this study each consisted of 10 min long simulations of typical classroom behavior. Students in the videos were either (a) children with previous acting experience who were assigned the roles of “target students” or (b) children from the community who played the roles of “typical peers.” Consent and assent for participation in these videos was attained from students and at least one of their parents or guardians, and parents/guardians and students were compensated for their time. These consent and assent documents included provisions for the videos derived from their participation to be utilized in other research studies. Written permission to utilize these clips in the present study was acquired from the principal investigators of the study from which these videos originated.

Two sets of clips were utilized in this study, the first depicting elementary-aged students, and the second depicting middle-school-aged students. Procedures for developing both sets of clips were identical. For each age group, 1 hr of coaching preceded filming, that consisted of providing modeling of examples and non-examples of target behaviors for all students. This was followed by role-playing, practice, and finally filming. During filming, student behavior within the videos was systematically varied by clip for both (a) academically engaged behavior, and (b) disruptive behavior. Target students were cued off-screen by the research team to initiate or terminate specific behaviors at given points throughout each clip. For this study, the middle-

school clip set was utilized during training procedures; the elementary clip set was utilized for all experimental procedures.

Lesson content for each clip set varied by age group. Within elementary clips, a female teacher-actor conducted a lesson on spiders and folk-tales for a group of 10 students. Within the middle-school clip set, a male teacher-actor conducted a lesson on Martin Luther King Jr.'s "I Have a Dream" speech for a group of 12 students. Across both sets of clips, four target students were utilized who represented each possible combination of binarily-defined gender (i.e., male/female) and race (i.e., white/non-white).

All observation procedures within this study utilized 15 s intervals. In order to ensure consistent application of these intervals across all study procedures, auditory tones were superimposed over digital files for all 12 clips utilized in this study. Tones were initiated every 15 s, with the first tone played 15 s after the moment the clip began (i.e., time = 0 min 15 s). Tones and tone-loop files were created by the investigator using Audacity editing and recording software (Audacity Team, 2012), and superimposed upon video clips by the investigator using Final Cut Pro video editing software.

Training materials. In order to train participants to criterion on each time-sampling methodology and behavior, training videos and master codes for those videos were established. Twelve unique training videos, each 5 min in length, were created utilizing the middle-school-aged set of videos previously described; to wit, each of the six 10 min videos were edited in half to create twelve 5 min videos. Nine of the twelve videos were randomly selected for training purposes: three for training on methodology, and six for training on behaviors. For methodology training, each video was coded using a given time-sampling methodology while observing academically engaged behavior. For behavior training, each of the six additional behaviors to be

observed during experimental procedures was assigned to a clip, with time-sampling methods randomly assigned twice-over without replacement, such that each of the three possible methodologies was represented twice among the set of six behavior training clips.

After training stimuli were created, master codes for each stimulus were developed. First, the investigator coded each of the 5-min-long training videos once in succession, utilizing the target methodology and behavior for that clip. After an interval of approximately 24 hr, the training videos were coded a second time. Scores between protocols were then compared, and inter-observer agreement was calculated using the following equation: number of agreements divided by the sum of number of agreements and number of disagreements. If two protocols had less than 100% agreement, the video was re-coded after an interval of 24 hr had elapsed since the prior coding, until 100% agreement between two temporally-adjacent protocols was established. The protocols with 100% agreement were utilized as the “master” for that given training clip.

Target student. The behavior of four target students was systematically manipulated during the creation of both elementary-age and middle-school-age videos, with all students scripted to perform the same duration of manipulated behaviors during each clip. Given that the camera utilized during filming was oriented in a fixed position for the entirety of each set of clips, the duration of each target student’s on-screen visibility varied depending on the student and the behavior they exhibited (e.g., walking around the classroom and out of the camera’s field of view). Therefore, the elementary-age student who displayed the majority of on-camera time was selected for use as the target student during experimental procedures. This individual served as the target student across all experimental clips. For training clips, students were initially randomly assigned for use as the target student in each clip. If upon master coding of training clips, the target student was observed to be off-camera for any portion of the clip, another target

student was randomly selected until one was identified who was on-camera for the entirety of the training clip.

Target behaviors. Ratings of academic engagement served as the primary dependent variable for this study, and as such, this behavior was rated during every participant observation. Six additional behaviors were identified for use in additional conditions of the “behavior target” facet, that consisted of four sets of behavior targets with 1, 3, 5, or 7 behaviors to be rated simultaneously. For convenience, these sets are referred to as Set 1, Set 3, Set 5, and Set 7, respectively. Those behaviors above and beyond academic engagement (AE) were drawn from the SECOS (Saudargas & Lentz, 1986), that features eight “state,” or prevalence-based, behaviors. Six of the eight behaviors identified on the SECOS (i.e., Out of Seat, Looking Around, Motor Behavior, Playing with Object, Social Interaction with Child, and Social Interaction with Teacher) were utilized for the current study. The remaining two SECOS behaviors were judged to be either too similar to the dependent variable (i.e., School Work), or too vaguely operationally-defined (i.e., Other Activity) to be used in the current study. The operational definitions for the six utilized SECOS behaviors provided in Saudargas and Lentz (1986) were used for all training and study ratings (see Table 2).

Set 1 consisted exclusively of AE, utilizing the previously-discussed modified definition supplied by Briesch et al. (2010). The remaining sets were comprised as follows: Set 3, AE and two behaviors; Set 5, AE and four behaviors; and Set 7, AE and six behaviors. In order to ameliorate a potential confound among (a) the number of behavior targets, which was the intended manipulation, (b) the specific behaviors utilized in each condition, and (c) the presentation order of those behaviors in each condition, quasi-randomization procedures took place. Procedures were not fully random due to the condition that AE be present in all behavior

sets. Behaviors were randomly assigned without replacement into sets of 3, 5, and 7 behaviors for each unique observation, with AE included across all potential randomized sets. The randomly-assigned order was preserved for presentation in subsequent experimental procedures. Set 1 did not require randomization, given that the only possible behavior for this condition was AE. Thus, at the end of randomization, each unique observation (e.g., Rater 1 observing 5 behaviors with MTS for Clip A) involving 3, 5, or 7 simultaneous behavior targets utilized a randomly-selected and randomly-ordered behavior set, with AE included across all sets.

Data collection. All data collection completed by participants during both training and experimental sessions was conducted using paper-based forms. These forms were constructed using researcher-defined templates within Microsoft Word, and included fillable areas for (a) observation-specific identifiers (e.g., participant ID, clip title), (b) definitions of behaviors and methodology utilized for the observation of interest, and (c) titles of behaviors within the data-collection area of the form itself. Templates were merged with a Microsoft Excel spreadsheet containing the relevant information for each participant's observation schedule in order to generate prepared data collection forms. Forms were printed and provided un-stapled to participants in the order in which they were to be completed (e.g., experimental forms for Observation 1 to Observation 72). Although data collection forms for those observations utilizing one, three, or five behavior targets were accommodated within a single one-sided sheet of paper (see Sample Experimental Data Collection Form, One Page; Appendix D), forms for those observations utilizing seven behavior targets required the use of a single double-sided sheet of paper due to the size of the data-collection area of the form (see Sample Experimental Data Collection Form, Two Pages; Appendix E).

In order to facilitate accurate participant completion of each data collection form, participants were provided with an accompanying binder displaying sheets of paper corresponding to each participant observation. Two sections were included in each binder: training and experimental. Each binder sheet displayed (a) a still image of the stimulus environment, with the target student's face circled, (b) a large-print definition of that observation's given methodology, and (c) a large-print definition of the behavior targets utilized for that observation. Thus, the binder sheets repeated the content displayed on each data collection sheet in a larger-print format, with the addition of an image depicting the target student. Although binder sheets for the training portion of the study provided necessary and unique information regarding the identity and appearance of the target student for a given clip (see Sample Training Binder Sheet, Appendix A), binder sheets for the experimental portion of the study displayed the same image of the target student on each sheet due to the use of a single, static target student for all experimental procedures (see Sample Experimental Binder Sheet, Appendix C).

Presentation order. In order to minimize potential order effects attributed to the repeated use of given methodologies and behavior targets as applied to individual stimulus clips, the presentation order of these variables was manipulated prior to the initiation of study procedures. Randomization of presentation order occurred in two phases: (a) block-randomized by clip, and (b) fully-randomized by methodology and behavior target. In the first phase, clip presentation order was independently assigned for each participant. The six experimental clips were randomly assigned without replacement into twelve blocks; thus, each block consisted of all six clips, randomly ordered. These blocks were arranged in sequence (from Block 1 to Block 12), in order to ensure that the distribution of clips was spaced across the experimental

procedures while maintaining a degree of randomness. If blocks were assigned such that the same clip was observed twice in succession (e.g., Clip A was assigned as both the last clip in Block 4 and the first clip in Block 5), one of the two phases was randomly selected for re-randomization until no clips were observed twice in succession. In the second phase, each unique methodology and behavior target combination was randomly assigned without replacement to the established clip order, by clip and participant (see Figure 2 for random assignment of presentation order by participant).

Procedures

Rater training. Immediately prior to initiating experimental study procedures, all raters participated in rater training that consisted of: (a) a review of SDO and the decision rules used for PI, WI, and MTS procedures and (b) training to criterion on the use of all three time-sampling methodologies and each of the target behaviors utilized in this study. Training occurred in a large, quiet university conference room under the supervision and direction of the investigator. Each participant was provided with the following materials: (a) training and experimental data collection forms, (b) binder depicting procedures for all training experimental observations, (c) color copies of training slides, (d) a small personal laptop upon which to view experimental clips, (e) headphones, and (f) pencils.

Review of procedures. In order to provide all participants with a standardized understanding of the procedures utilized in this study, as well as the behaviors to be addressed, a presentation lasting approximately 45 min was delivered to all participants (see Training Presentation Slides, Appendix H). During this training presentation, participants (a) were generally oriented to materials, (b) reviewed and signed consent documentation, (c) reviewed the basic theoretical underpinnings of and procedures in SDO, (d) reviewed the specific information

provided on all data collection forms, and (e) reviewed the behaviors to be observed, with definitions, examples, and non-examples. Raters were also instructed to utilize a 0/1 rating scheme for all training and experimental data collection procedures, in order to introduce a uniform rating procedure across all participants. Participants therefore were instructed to code every interval and behavior on all data collection forms as either 0 for non-occurrence, or 1 for occurrence. Following this review of procedures, training to criterion occurred.

Training to criterion. In order to ensure that each rater was able to initially demonstrate accurate and consistent ratings of the behavior targets and methodologies to be utilized in this study, rater training to criterion was undertaken. Mastery of a given training procedure was defined as 90% or greater agreement with the master code. If mastery was not demonstrated for a given training clip, the participant repeated coding for that clip until mastery was demonstrated, and subsequently advanced to the next training clip.

The sequence of procedures for training occurred as follows. First, with master codes established (as previously described), participants completed ratings of academically engaged behavior using each of the three time-sampling methods. If 90% agreement was observed between the participant and the master code for a given method, the participant advanced to the next clip and time-sampling method, until 90% agreement was demonstrated using all three time-sampling methods. After demonstrating mastery of each methodology, participants rated each additional behavior using its corresponding randomly-selected methodology, following the same procedures as those for methodology training. After mastery was demonstrated for all training clips, specific procedures for experimental ratings were explained, and experimental procedures began.

Experimental ratings. Prior to initiating experimental ratings, instructions for break schedules and instructions for overall rating completion were provided. First, in order to combat rater fatigue, participants were instructed to take breaks (a) after each clip, (b) after every six clips, and (c) after every eighteen clips. After each clip, participants were instructed to take a breath and prepare protocols and binder pages for the subsequent clip. After every six clips (or after 1 hr of coding), participants were instructed to take a 10 min break wherein they might engage in some physical exercise (e.g., “stretch your legs”). After every eighteen clips (or after 3 hr of coding), participants were instructed to take a lunch break. Participants were also instructed to try their best on all observations, but to not rewind or restart clips unless the incorrect clip was selected for playback or they inadvertently began a clip before they were prepared to rate. Raters were instructed not to discuss behaviors or coding questions after the initiation of experimental procedures, nor to discuss coding with other participants. Participants were asked to remember to code to the best of their ability, using the information available to them in their training materials and data collection sheets.

Upon delivery of all reminders and logistical instructions from the investigator, participants began their experimental ratings. In order to provide initial support to participants as necessary, each participant completed at least half of all experimental ratings in the presence of the investigator. After completing at least half of all experimental ratings under investigator supervision, participants independently completed the remaining ratings and returned their completed protocols to the investigator. Participants were compensated for their involvement and thanked for their time.

Completed data sheets were collected and ratings of academic engagement were manually entered by the investigator into a Microsoft Excel spreadsheet. This was subsequently exported into the comma-separated values (CSV) file format for data analysis.

Results

Data Preparation

Data entry verification. In order to verify the integrity of data entry, 29.6% of all data sheets ($n = 213$) were randomly selected for independent double-entry by a second researcher. Inter-observer agreement was calculated between the two entries using the agreement index (number of agreements divided by number of agreements and disagreements). Agreement was calculated with a researcher-defined function (see Appendix I) at the level of both (a) individual 15 s interval, as well as (b) calculated prevalence statistic, resulting in 8520 possible agreements and disagreements (40 intervals $\times$ 213 data sheets) for individual interval calculations and 213 possible agreements and disagreements for the prevalence statistic. The observed IOA value between the two researchers was 99.7% by interval, and 96.2% by prevalence statistic, suggesting extremely high levels of consistency across data entry procedures. All disagreements between researchers were individually checked for accuracy against the original data sheet, with the master data file revised according to the value observed in the original sheet.

Dependent variable creation. All analyses were conducted using the calculated prevalence statistic as the dependent variable, defined as the proportion of total intervals wherein an occurrence of academically-engaged behavior was observed. This proportion was calculated for each observation ($n = 720$) within the statistical software package, rather than by hand.

Missingness. Four missing ratings were observed within the completed data set, with a total of 28,800 ratings (40 intervals $\times$ 720 observations) possible. As such, missing values

represented 0.01% of the overall interval-by-interval data set. Missing ratings were distributed across protocols, such that four total observations (0.56%) were each missing one rating for one interval. Given that the total number of missing values was distributed evenly across protocols, with each of the four protocols missing one interval rating of a possible forty, the relative influence of each missing data point was equivalent to a change in prevalence estimate of ± 0.025 . Given the very small amount of missingness and the relative lack of influence each missing data point possessed upon overall estimates of prevalence, these values were accepted as missing rather than imputing their values or deleting their corresponding cases. Prevalence statistics for these four observations were calculated using a denominator of 39 rather than 40 intervals, and were treated as typical cases for all further analyses.

Analysis

All analyses were conducted within the R statistical language and environment (R Core Development Team, 2013) using the RStudio interactive development environment (RStudio, 2013), as well as the EduG software program. Analyses in R were supplemented and facilitated by the *plyr* (Wickham, 2011), *lme4* (Bates, Maechler, Bolker, & Walker, 2014), *stringr* (Wickham, 2012), *ggplot2* (Wickham, 2009), and *exactRankTests* (Hothorn & Hornik, 2013) packages, with some code adapted from the QME package (Vue, Zieffler, Brown, Chang, & Stanke, 2013).

Descriptive statistics. Means and standard deviations were calculated by clip for each combination of methodology and simultaneous behavior targets (e.g., MTS with 3 targets), as well as across targets by method (e.g., MTS across 1, 3, 5, and 7 targets) and across all measurement combinations. Results of descriptive analyses are depicted in Table 3.

Overall. The overall mean estimate of prevalence for academic engagement varied by clip from 0.23 (Clip F) to 0.86 (Clips A and C), with standard deviation values ranging from 0.13 (Clips A and D) to 0.20 (Clip E). In other words, estimates of the percentage of total intervals the target student was involved in academically-engaged behavior ranged from 23% to 86% across the six clips observed for this study, when averaged across all protocols that observed that clip. Mean prevalence for academic engagement ratings for Clips A, C, and D ranged between 0.83 and 0.86, whereas mean values for Clips B and E were 0.61 and 0.68, respectively. The mean estimate for Clip F demonstrated the lowest overall ratings for academic engagement, with a prevalence of 0.23. Standard deviation values were lowest for Clips A and D ($SD = 0.13$), as well as Clip C ($SD = 0.14$). Clip F demonstrated a relatively moderate amount of variance ($SD = 0.18$), and Clips E and B demonstrated relatively high standard deviation values of 0.20 and 0.23, respectively. Variance followed a somewhat decreasing trend when compared against academic engagement ratings, with clips depicting lower levels of engagement demonstrating relatively high standard deviations, and clips with higher levels of engagement demonstrating lower standard deviation values.

By method. When means and standard deviations were calculated by methodology (i.e., calculated across all protocols that utilized MTS, PI, or WI procedures) for each clip, mean estimates of prevalence could be rank ordered by method from least to most as: WI, MTS, and PI (see Figure 3). Mean estimates for all clips were the lowest when derived using WI time-sampling procedures, followed by MTS procedures, with PI time-sampling resulting in the highest mean estimate of prevalence for each clip. Those clips that demonstrated relatively high overall mean ratings for academic engagement also demonstrated fairly uniform ranges in means across methodologies: Clip A (range: 0.73-0.99), Clip C (range: 0.71-0.98), and Clip D (range:

0.68-0.97). The ranges in means for Clips B (0.35-0.86) and E (0.47-0.90) were somewhat larger across methodologies; means for Clip F, with the lowest overall mean prevalence rating, ranged in values across methodology from 0.05 to 0.42. Standard deviations across all methods and clips ranged from 0.01 (Clip A, PI) to 0.13 (Clip C, WI). These values also varied fairly widely within method and across clips: MTS (range: 0.05-0.09), PI (range: 0.01-0.12), and WI (range: 0.04-0.13).

Within each clip, tests for significant differences in means between methodologies were conducted using the nonparametric exact Wilcoxon signed-rank test, which tests for differences between two related samples without the strict parametric assumptions of normality assumed by analyses such as Student's *t*-test. Due to the presence of ties and zeros across samples, the *wilcox.exact* function was utilized, permitting robust calculation of exact *p*-values under these conditions. Given that three tests were conducted within each clip (i.e., PI/WI, PI/MTS, MTS/WI), inflation of the familywise error rate was addressed using the Bonferroni correction, with a revised critical *p*-value of .016 from an original alpha value of .05. For each clip, significant differences were found between all possible methodology comparisons ($p < .001$).

By method and behavior targets. Within each methodology, means and standard deviations were calculated for prevalence ratings using each possible condition of simultaneous behavior targets (i.e., 1, 3, 5, 7), as depicted in Table 3. Ranges for mean prevalence ratings across all methodologies and number of behavior targets were fairly limited, with the largest absolute difference between means observed using WI for Clips B (range: 0.31-0.38) and C (range: 0.68-0.75), both of which demonstrated a difference between minimum and maximum mean ratings of 0.07. The lowest range within methodology and clip was observed for Clip A using PI time-sampling procedures; the mean rating using one target behavior was 1.00, whereas

the means for ratings using three, five, and seven target behaviors were 0.99, for an absolute difference of 0.01 between minimum and maximum mean ratings.

Variance component estimation. When estimating variance components in a mixed model, various approaches may be appropriate depending upon one's interpretative framework and intended decision-making. The first step in undertaking a full G study for a three-facet model would involve the estimation of variance components for the object of measurement, all facets, and all facet interactions, for a total of 15 estimated components. The object of measurement for this study was defined as occasion ("o", $n = 6$), or the six video clips that were observed by each rater. Additional facets of interest were defined as rater ("r", $n = 10$), method ("m", $n = 3$), and simultaneous behavior targets ("b", $n = 4$). Given that all facets were fully crossed, this full model may be depicted as $o \times r \times m \times b$.

Notably, when interpreting variance components for a fixed facet, scores are averaged over the conditions of that fixed facet, which may or may not make conceptual sense within a given model (Shavelson & Webb, 1991). Thus, two model types were submitted for analysis. First, a fully random design was utilized in order to determine to what extent each facet of measurement contributed to overall score variance. This first model may inform general interpretations of sources of variance across the data set; however, these values may be of limited conceptual use given the unlikelihood of a measurement context in which an individual would utilize an average rating derived from three different time-sampling methodologies or four simultaneous behavior targets. Furthermore, the assumptions of random sampling that are associated with the identification of a facet as random would be violated with the deliberate selections made within the "method" and "behavior" facets. Thus, in order to determine the relative amount of variance attributable to each condition of each fixed facet, analyses were also

broken out by each condition combination, with a full one-facet G study of the design “occasion x rater” run for each combination of the fixed facet conditions (e.g., PI with 5 behaviors, WI with 7 behaviors, MTS with 1 behavior).

Fully random model. Utilizing an ANOVA Type III Sums of Squares procedure within the EduG software program, a fully-random $o \times r \times m \times b$ model was calculated for the total data set, with results displayed in Table 4. Across the 15 variance components that were derived for this model, two small negative variance components were observed: the interaction between occasion and behavior targets ($\sigma^2_{ob} = -0.00003$) and the interaction between method and behavior targets ($\sigma^2_{mb} = -0.00004$). Following the recommendations of Briesch, Swaminathan, Welsh, and Chafouleas (2014), in the presence of negative variance components, two interpretative techniques were utilized in order to determine the relative influence of negative variance components. First, the total variance was calculated by summing both positive and negative variance components, and the proportion of variance attributable to each component was calculated by dividing each component by this overall sum. Second, values for the two negative variance components were replaced by zero prior to taking the sum of all components, and proportions of total variance were calculated using this modified total in the denominator. Changes between the two methods in the percentage of total variance were only observed for three components (i.e., o, r, m) when taking the percentage to two decimal places, and each of these changes in percentage were relatively small (absolute differences of 0.04%, 0.01%, and 0.02%, respectively). Given the relatively small differences noted between the two methods, problems with model specification were not considered a major concern, and interpretation of the full model was conducted with negative variance components set at zero, following the recommendations of Cronbach et al. (1972; see also Briesch et al., 2014).

The majority of variance observed among prevalence statistics was attributed to the occasion facet (56.74%), suggesting that variance across individual ratings was most systematically influenced by the video clips observed during those ratings. The second-highest amount of variance among prevalence ratings was attributed to the method facet (30.85%), suggesting that after occasion, method was the most systematic source of variance in prevalence ratings. The occasion and method variance component estimates together comprised 87.59% of the total variance observed among the prevalence ratings. The final two individual facets of the full model, rater and number of simultaneous behavior targets (“behavior targets”), each contributed 1.58% and 0.01% of the observed variance, respectively, suggesting minimal contributions from these two facets when considered independently.

First- and second-order interaction variance component estimates comprised a total of 7.72% of the variance in prevalence ratings. Individual component estimate percentages for interactions that were over 1.00% were as follows, ranked from largest to smallest: occasion x method (2.55%), occasion x rater x method (1.63%), occasion x rater (1.26%), occasion x rater x behavior targets (1.08%), and rater x method (1.01%). The remaining interaction variance component estimates contributed a relatively minimal amount of variance, ranging in percentage from 0.08% (rater x method x behavior targets) to 0.00% for both variance component estimates that were set at zero (i.e., occasion x behavior targets, method x behavior targets).

In generalizability theory, the variance component estimate for the complete interaction term is confounded with the residual unexplained error within the model; in the current study, this term comprises the four-way interaction between occasion, rater, method, and behavior targets, as well as the residual. The percentage of total variance contributed by the variance component estimate for this term was 3.10%, which was the third-largest individual variance

component estimate in the model. However, this term was also lower than the sum of all first- and second- order interaction terms.

The fully random model provides general information regarding sources of variance across the entire data set, using occasion (e.g., video clip) as the object of measurement and prevalence ratings as the dependent variable of interest. However, as previously described, the method and behavior targets components are most appropriately considered fixed, since the conditions utilized therein were deliberately selected when defining the universe of generalization. Thus, a G study utilizing such a model may most appropriately be conducted using separate G studies for each level of the fixed facets.

Separate o x r models. A fully random one-facet occasion x rater (“o x r”) model was conducted for each condition combination of the two fixed facets utilized within this study, for a total of 12 separate o x r G-studies (three methodology conditions multiplied by four behavior target conditions). These models were conducted within the R statistical program using REML estimation; however, concerns have been noted regarding potential bias in variance component estimates derived from REML estimation techniques (Briesch et al., 2014). In order to check for any potential bias in REML estimation compared to ANOVA Type III model analyses, results of REML estimates were compared those derived from ANOVA estimation procedures in SPSS, with no differences noted between variance component estimates. Thus, REML-derived variance component estimates were utilized for each individual o x r G-study. Results from these G studies are presented in Table 5.

Across all 12 one-facet models, the occasion facet (i.e., the object of measurement) contributed the overwhelming majority of the total estimated variance. Percentages of total estimated variance for each occasion variance component estimate ranged from a minimum of

85.44% (WI, seven behaviors) to a maximum of 92.51% (MTS, seven behaviors), with a median value of 90.09% (calculated as the mean of the two center-most rank-ordered percentages). Thus, across all combinations of each fixed-facet condition, the majority of the variance in prevalence scores across all occasions and raters was attributed to the occasion facet. The size of the variance component estimate for the rater facet was relatively low, varying from a minimum of 1.70% (PI, five behaviors) to a maximum of 6.31% (WI, seven behaviors), with a median value of 2.88%. These relatively low values suggested that a minimal amount of variance was attributable to systematic rater influence upon prevalence ratings. Finally, the percentage of total variance attributable to the confounded interaction/residual term (or, e) was also relatively low, ranging in magnitude from 4.78% (MTS, one behavior) to 15.22% (PI, five behaviors), with a median value of 6.98%. The range of percentages of total variance for the interaction/residual facet was somewhat more expansive than that for the rater facet, but still fairly minimal when compared to that derived from the occasion facet. Across all combinations of methodology and behavior target, the rank order of each facet in terms of relative proportion was (from lowest to highest): rater, interaction/residual, and occasion.

Between methodology. When examining relative variance component size within all behavior target conditions, between methodology, some patterns emerge. However, given the relatively small magnitude of the differences in variance component proportions, these results must be interpreted with caution. Across methodologies, MTS consistently demonstrated the highest percentage of variance attributable to the object of measurement, followed by PI and then WI. This rank ordering holds for three of the four behavior conditions observed; the “five behaviors” condition, wherein the variance component estimate for occasion was proportionally highest for MTS (90.18%), followed by WI (88.62%), and then PI (83.06%). Variance

component estimate percentages for the object of measurement for ratings conducted within the MTS condition all exceeded 90%, whereas those in the PI condition ranged from 83.08% (five behaviors) to 91.17% (one behavior). Occasion percentages for ratings conducted using WI sampling did not exceed 90%, ranging from 85.44% (seven behaviors) to 89.28% (three behaviors). Notably, the percentage of variance attributable to rater was consistently higher in the WI condition (range: 3.86%-6.31%) than the MTS (range: 1.75%-3.64%) or PI (range: 1.70%-2.77%) conditions.

Between behavior targets. When examining percentages of total variance by behavior target conditions, some discernable trends are apparent in the relative magnitude of each facet's variance contributions within methodology. The distribution of facet variance across behavior conditions for MTS was highly consistent, with very small differences noted within facets and across behavior conditions. Across behavior conditions for MTS, the occasion facet contributed between 90.18% (5 behaviors) and 92.51% (7 behaviors) of the total variance, the interaction/residual contributed a limited amount of variance (range: 4.78%-6.92), and the rater facet even less (range: 1.75%-3.64%). Similar results were found within the WI condition, with individual facets contributing consistent magnitudes of variance across behavior conditions. Within the PI condition, however, consistency across behavior conditions was hindered by the five behavior condition, which demonstrated a relatively large residual term (15.22%) and correspondingly smaller proportion of variance attributable to the object of measurement (83.08%). Thus, although variance was consistently distributed across facets and behavior conditions within the MTS and WI conditions, the anomalous five-behavior condition within the PI methodology condition precluded similar conclusions for this time-sampling procedure.

No single behavior target condition's object of measurement facet possessed a consistent rank across all three methodologies. Within those ratings that were conducted using MTS procedures, the rank ordering of magnitude for the object of measurement facet across behavior target conditions was: 5, 3, 1, 7. For those conducted using PI procedures, the percentage of total variance attributable to the object of measurement was ranked across conditions as: 5, 7, 3, 1. Finally, percentage of total variance attributable to the object of measurement for ratings conducted using WI procedures were ranked across conditions as: 7, 1, 5, 3. When rank-ordering was divided into those conditions that were in the first half (i.e., lowest and second-lowest percentages of total variance) and second half of the total rank ordering (i.e., second-highest and highest percentages of total variance) within methodology, no single behavior target condition was consistently ranked within the first- or second-half of the rank ordering across all methodologies. In other words, no consistent trend in the relative size of the object of measurement facet was observed across behavior target conditions. These conclusions were similarly true for rankings of the rater facet, for which the percentages of total variance within methodology were ranked as follows: MTS (7, 1, 5, 3); PI (5, 1, 3, 7); and WI (5, 3, 1, 7).

Unlike the object of measurement and rater facets, the percentage of total variance attributed to the interaction/residual component exhibited some consistency across behavior target conditions. The rank ordering of behavior targets within each methodology were as follows: MTS (1, 3, 7, 5), PI (3, 1, 7, 5), and WI (3, 5, 1, 7). When individual rank-order placements are examined, no single behavior target condition remained in an individual placement across methodologies. However, when first- and second-half placements are examined, interaction/residual variance component percentages derived from ratings conducted utilizing three behavior targets were consistently ranked within the first half of relative

percentage magnitude, and those from ratings using seven behavior targets were consistently ranked within the second half of relative percentage magnitude. However, this was not true for components from ratings conducted using one or five behaviors.

Generalizability and dependability coefficients. Both generalizability (e.g., relative decision-making) and dependability (e.g., absolute decision-making) coefficients were generated using the original model specification ($n_r = 10$) for each individual $o \times r$ G-study. Coefficients were created using a researcher-defined function within R (see Appendix I). As presented in Table 6, coefficients utilizing the original measurement procedures exceeded 0.900 across all combinations of methodology and behavior targets. These results are concordant with those of the variance component analyses, given the large proportion of variance attributable to the object of measurement (i.e., “desirable” variance).

G coefficients. As described earlier, generalizability coefficients (ρ^2) denote the extent to which an estimate of the dependent variable observed within the object of measurement may be considered “reliable” for the purposes of relative decision-making (e.g., a student’s estimated prevalence of academically engaged behavior during one 10-min observation when compared to that of her peers). Across all methodology and behavior target combinations, observed generalizability coefficients ranged in value from 0.982 (PI, five behaviors) to 0.995 (MTS, one behavior; MTS, three behaviors). Thus, the proportion of variance attributable to the object of measurement, when compared against that attributable to the sum of the object of measurement and the interaction/residual term divided across 10 raters, was consistently large.

Generalizability coefficients within methodologies did not consistently increase or decrease as a function of behavior targets, and within methodology, extremely small differences were noted across behavior target conditions. When rounding to two decimal places, both MTS

and WI consistently demonstrated generalizability coefficients of 0.99; only the PI condition demonstrated distinct generalizability coefficients across behavior target conditions when rounding to two decimal places. For those procedures employing MTS, generalizability coefficients ranged from 0.993 (three behaviors) to 0.995 (one and three behaviors). For those ratings that utilized PI time-sampling, generalizability coefficients ranged from 0.982 (five behaviors) to 0.992 (one and seven behaviors). Finally, ratings utilizing WI time-sampling procedures demonstrated generalizability coefficients ranging from 0.990 (seven behaviors) to 0.993 (three behaviors). Models utilizing MTS demonstrated equal or higher generalizability coefficients than those utilizing PI or WI; for the one, three, and seven behavior target conditions, MTS-based generalizability coefficients exceeded both PI and WI, and coefficients for the five-behavior condition were equal in the MTS and WI conditions. However, although the coefficient exhibited for the three-behavior-target condition was the highest or equally-highest within each methodology, these differences were extremely small and unnoticeable if rounded to the second decimal place.

D coefficients. Dependability coefficients (Φ) reflect the extent to which ratings may be considered “reliable” for the purposes of absolute decision-making, such as when a rating of academically engaged behavior may be utilized during progress-monitoring to make within-, rather than between-, student decisions. The numerator used during the calculation of dependability coefficients, like that used for generalizability coefficients, is equal to the variance component estimate for the object of measurement. However, the denominator used within dependability coefficients calculation for the original measurement context comprises the sum of (a) the variance component estimate for the object of measurement, (b) the rater variance component estimate divided by 10 (i.e., the number of raters in the original measurement

design), and (c) the interaction/residual term divided by 10. As a result, dependability coefficients for this study reflect variance attributable to systematic rater influence, in addition to simply interaction/residual variance.

Dependability coefficients across all combinations of methodology and behavior-target conditions exceeded 0.900, ranging in value from 0.980 (PI, five behaviors) to 0.992 (MTS, one behavior; MTS, seven behaviors). Dependability coefficients in this model differ from generalizability coefficients in that they include the rater variance component in the denominator; given the small observed percentage of variance attributable to the rater facet, these small differences between generalizability and dependability coefficients are unsurprising. The condition combinations comprising the minimum and maximum dependability coefficient values were nearly equivalent to those for the minimum and maximum generalizability coefficient values; in other words, ratings conducted using PI time-sampling with five behavior targets demonstrated the lowest generalizability and dependability coefficients, and ratings conducted using MTS with one behavior target demonstrated the dependability coefficients that were tied for the maximum with other behavior target conditions. However, the behavior target conditions with which the coefficient was tied varied between coefficient type ($\rho^2_{\text{MTS},1} = \rho^2_{\text{MTS},3}$; $\Phi_{\text{MTS},1} = \Phi_{\text{MTS},5}$).

As with the observed generalizability coefficients, differences between dependability coefficients were extremely small, and many differences are indistinguishable when rounded to a second decimal place. No increasing or decreasing trend in the magnitude of dependability coefficients was observed across behavior targets within each methodology, nor was there a behavior target condition that consistently demonstrated the lowest or highest dependability coefficients across time-sampling methodologies. Within all behavior target conditions, the

magnitude of the dependability coefficient was highest for the MTS methodology condition ($\Phi_{\text{MTS},1} = 0.992$, $\Phi_{\text{MTS},3} = 0.991$, $\Phi_{\text{MTS},5} = 0.989$, $\Phi_{\text{MTS},7} = 0.992$). Dependability coefficients for ratings utilizing PI time-sampling ranged from 0.980 (five behaviors) to 0.990 (one behavior, seven behaviors), and those utilizing WI time-sampling ranged from 0.983 (seven behaviors) to 0.988 (three behaviors). Within each behavior target condition except that utilizing five behavior targets, the value of the dependability coefficient across methodologies was rank-ordered as follows: WI, PI, and MTS. When rounded to two decimal places, WI consistently demonstrated dependability coefficients of 0.98, and PI and MTS demonstrated coefficients of 0.98 and 0.99.

D studies. Although generalizability and dependability coefficient values under the original measurement context consistently exceeded 0.900, it is not likely that a model utilizing an estimate of academic engagement averaged over 10 simultaneous raters would be utilized in a real-world setting. Thus, a set of D studies were conducted for each combination of methodology and behavior-target conditions in order to determine the minimum number of raters required to demonstrate an acceptable level of reliability for both relative- (generalizability coefficients, or ρ^2) and absolute- (dependability coefficients, or Φ) decision making. Given the large magnitude of the generalizability and dependability coefficients demonstrated by the original measurement model, which utilized 10 raters, the D study was undertaken by manipulating the number of raters over which an estimate of academic engagement would be averaged to a value between 1 and 9 (as well as the original 10 raters), utilizing a researcher-defined function within R. The resulting generalizability and dependability coefficients are presented in Table 7.

G coefficients. As displayed in Figure 6, generalizability coefficients for all combinations of methodology and behavior target conditions exceeded 0.900 when scores were averaged over at least two raters. When utilizing a single rater, all combinations except one (PI, five behaviors)

demonstrated generalizability coefficients above 0.900. Values for ρ^2 when one rater was utilized ranged from 0.845 (PI, five behaviors) to 0.951 (MTS, one behavior). The second-lowest coefficient after that of the PI condition utilizing five behavior targets was 0.912 (WI, seven behaviors). Within procedures utilizing MTS to derive estimates of academically engaged behavior, generalizability coefficients for one rater ranged from 0.929 (five behaviors) to 0.951 (one behavior). Procedures utilizing PI demonstrated generalizability coefficients ranging from 0.845 (five behaviors) to 0.935 (three behaviors), and those utilizing WI ranged from 0.912 (seven behaviors) to 0.939 (three behaviors). Thus, within each methodology, no consistent pattern was observed regarding which behavior target condition yielded the highest generalizability coefficient when only one rater was utilized to derived a prevalence estimate.

Within behavior targets, rating procedures utilizing MTS consistently demonstrated the highest generalizability coefficients when only one rater was utilized, compared to PI or WI time-sampling procedures. Coefficients derived from procedures utilizing one behavior target and one rater ranged from 0.919 (WI) to 0.951 (MTS), and those utilizing three behavior targets ranged from 0.935 (PI) to 0.949 (MTS). When five behavior targets were simultaneously rated, generalizability coefficients for one rater ranged from 0.845 (PI) to 0.929 (MTS), and those from seven-behavior-target procedures ranged from 0.912 (WI) to 0.942 (MTS).

As previously described, all combinations of behavior target and methodology demonstrated generalizability coefficients about 0.900 when estimates of prevalence were averaged over two raters. The magnitude of these two-rater-derived coefficients ranged from 0.916 (PI, five behaviors) to 0.975 (MTS, one behavior). All generalizability coefficients for two raters derived from MTS (range: 0.963-0.975) and WI procedures (range: 0.954-0.968) exceeded

0.950, and those derived from PI procedures ranged from 0.916 (five behaviors) to 0.967 (three behaviors).

D coefficients. As displayed in Figure 7, patterns of dependability coefficient magnitude were similar to those displayed by generalizability coefficients with regards to the number of raters required for all combinations of fixed facet conditions to demonstrate high levels of dependability (i.e., $\Phi \geq 0.900$). When only one rater was utilized, dependability coefficient estimates ranged from 0.831 (PI, five behaviors) to 0.925 (MTS, seven behaviors). Dependability coefficients for seven of twelve possible combinations of methodology and behavior target equaled or exceeded 0.900 when one rater was utilized; these consisted of all behavior target conditions for MTS procedures; and one, three, and seven behavior targets for PI procedures. Dependability coefficients that were less than 0.900 when one rater was utilized ranged in value from 0.831 (PI, five behaviors) to 0.893 (WI, three behaviors). As suggested by relative differences in the percentage of variance attributable to the rater facet across methodologies, differences between generalizability and dependability coefficients were minimal within the PI condition, somewhat larger within the MTS condition, and largest within the WI condition.

As with generalizability coefficients, dependability coefficients with one rater were highest when MTS procedures were utilized, across all behavior target conditions. These MTS-derived values ranged from 0.902 (five behaviors) to 0.925 (seven behaviors). Values for Φ with one rater were higher for PI procedures than WI procedures across all behavior target conditions except for the five-behavior condition ($\Phi_{PI,5} = 0.831$, $\Phi_{WI,5} = 0.886$). Dependability coefficient values derived from PI procedures with one rater ranged from 0.831 (five behaviors) to 0.912 (one behavior), and those derived from WI procedures ranged from 0.854 (seven behaviors) to

0.893 (three behaviors). Thus, no single behavior-target condition demonstrated the highest dependability coefficient values when one rater was utilized.

Although results for the one-rater D study were mixed, dependability coefficients for all combinations of methodology and behavior-target conditions exceeded 0.900 when estimates of prevalence were averaged over two raters. Values for Φ with two raters ranged from 0.908 (PI, five behaviors) to 0.961 (MTS, seven behaviors). As with the one-rater D study, dependability coefficient values for two raters were highest when derived from MTS procedures, ranging from 0.948 (five behaviors) to 0.961 (seven behaviors). Those derived from PI procedures with two raters ranged from 0.908 (five behaviors) to 0.954 (one behavior), and from 0.921 (seven behaviors) to 0.943 (three behaviors) for WI procedures.

Discussion

Time-sampling procedures for systematic direct observation enjoy widespread use and an extensive theoretical literature base. Furthermore, instruments utilizing time-sampling procedures may be flexibly designed in order to apply to a number of potential measurement contexts, depending on the behavior or behaviors of interest, the characteristics of the behavior, and other potential dimensions. However, research examining the extent to which these instrument-development decisions impact the resulting generalizability and dependability of resulting data is limited, and no prior studies have utilized generalizability theory to examine the effect of methodology and number of simultaneous behavior targets upon relative proportions of variance and subsequent reliability-like coefficients. Thus, this study sought to examine the extent to which (a) rater, (b) time-sampling methodology, and (c) number of simultaneous behavior targets influenced relative proportions of variance and reliability-like coefficients for ratings of academically engaged behavior.

Descriptive Results

Overall estimates of the proportion of time the target student was engaged in academically engaged behavior varied across the six 10-min observations utilized as the object of measurement for this study. Descriptive analyses of differences in estimates by methodology, number of behavior targets, and clip suggested that within-clip estimates were fairly stable when conducted with a specific methodology; few differences were observed in prevalence estimates as the number of behavior targets was manipulated. However, large differences in prevalence estimates were identified depending on the specific methodology used, with estimates rank-ordered from least to most as WI, MTS, and PI, as would be theoretically expected when applying all three methodologies to a common behavior stream in the presence of mixed intervals (Suen & Ary, 1989). For a behavior stream that produces mixed intervals, wherein a behavior does not occur for the entire duration of the interval, WI sampling will underestimate the overall duration of a specific behavior, MTS may provide a more-moderate estimate, and PI sampling will overestimate behavior duration. This appeared to hold across all clips. However, the limited differences in means that appeared to exist within methodology and across behavior target conditions suggested that a relatively small amount of variance may be attributable to the influence of simultaneous behavior targets.

Research Question 1 (Fully-random model). Using raters with similar levels of prior training, what proportion of overall variance in academic engagement ratings is attributable to time-sampling methodology?

It was hypothesized that a majority of rating variance would be attributable to time-sampling methodology, followed by variance attributable to the occasion facet. The conclusions suggested by the descriptive statistics were borne out with the completion of the fully-random

model, utilizing occasion as the object of measurement and rater, methodology, and behavior targets as facets. Although this fully-random model did not provide information that was conceptually relevant in a generalizability framework, it did provide information regarding general sources of variance in the overall data set. Results derived from the fully-random variance components analysis suggested that the majority of variance in prevalence ratings of academic engagement across the entire data set was attributable to the behavior depicted in the clip itself, which is the most desirable source of variance for these ratings. This result suggested that raters most often completed their observation protocol based on what occurred in the clip, rather than based on effects of other non-desirable sources of variance like rater-specific influences such as fatigue or idiosyncrasies in individual applications of behavior definitions. As would be expected based on consistent application of the decision rules involved in each time-sampling methodology, the second-largest source of variance observed from the fully-random model was the “method” facet. Given both (a) the theoretical differences that should be observed when decision rules are consistently applied to a behavior stream, and (b) the results of descriptive analyses suggesting large and consistent differences between means derived from distinct time-sampling methodologies, the relative size of this variance component is also expected. The size of this variance component suggests that the individual time-sampling method utilized during an observation systematically influenced the resulting prevalence estimate from that observation to a larger degree than any other factor besides the behavior itself. Thus, these results combined with those of the descriptive analyses suggest that raters were consistently applying time-sampling decision rules in a manner that would be expected, and would not be undesirable in an actual measurement context. Although it was hypothesized that the majority of overall variance would be attributable to methodology followed by occasion, rather than vice

versa, these two facets were found to be the two largest contributors of variance in ratings of academic engagement to the overall data set.

Research Question 2 (Interactions). Is any variance in academic engagement estimates attributable to an interaction between the number of behaviors rated and the type of time-sampling procedure used? In other words, are systematic differences observed in estimates of academic engagement depending on the combination of (a) number of simultaneous behaviors observed and (b) time-sampling methodology utilized? Is any variance attributable to other facets or interactions?

It was hypothesized that greater than 6.67% of total rating variance would be attributable to a methodology by behavior interaction. Outside of the clip itself and the method used to rate each clip, variance in ratings of academic engagement was distributed in small amounts across all other facets, interactions, and the residual term within the full model. Given that variance attributable to time-sampling methodology is somewhat self-evident and desirable, this study was partially concerned with determining if any variance in academic engagement ratings would be attributable to interactions between the behavior targets facet and the object of measurement, or the behavior targets facet itself and any other interactions. A small negative variance component was found for the interaction between behavior targets and method ($m \times b$), suggesting that no systematic differences were observed in estimates of academic engagement depending on the combination of number of simultaneous behaviors observed and time-sampling methodology utilized. Therefore, little support was found for the first portion of the hypothesis for the second research question, which postulated that more than 6.67% of total rating variance would be attributable to this interaction. Additionally, almost no systematic variance was attributable to the main behavior-targets facet, suggesting that there were little-to-no systematic

differences between ratings using one, three, five, or seven behaviors at a time when the effect of behavior targets was considered independently. Interaction effects including the behavior targets facet were similarly small, with a small negative variance component observed for the interaction between occasion and behavior targets (o x b), suggesting that ratings made using one, three, five, or seven behavior targets did not systematically differ as a function of the specific clip observed. Variance components for interactions with behavior targets were also near- or equal-to-zero with rater facet and most three-way interactions. Thus, results of this variance components analysis suggest that the number of behavior targets simultaneously observed had little to no systematic influence upon ratings of academic engagement, both when considered independently and when considered in terms of their interactions with other facets of measurement.

Individual raters also appeared to have little systematic influence upon estimates of academic engagement prevalence in this study. When considered individually, the rater facet contributed a minimal amount of variance to ratings, suggesting that ratings across raters were fairly uniform. Raters did not appear to systematically rate differently when presented with multiple behavior targets, nor did they systematically rate substantially different from one another when presented with different time-sampling methodologies. Across all potential sources of rater variance, little to any influence was attributed to individual rater idiosyncrasies upon estimates of academic engagement.

Finally, a small amount of variance was attributed to the four-way interaction between facets (occasion x residual x method x behavior targets), which is confounded with the residual term. Although it is impossible to disentangle this interaction from any random variance not

accounted for by the other facets in the fully-random model, it appears that a minimal amount of variance was unexplained.

The relative component sizes observed within the full model are fairly distinct to those observed in other studies of SDO-derived ratings of academic engagement. Within the fully random model, the percentage of variance attributable to the object of measurement was 56.78%, which is less than that observed by Hintze and Matthews (2004; 62%) but greater than that observed by two other studies (48%: Briesch et al., 2010; 29%: Ferguson et al., 2012). Furthermore, the percentage of unexplained variance in the present model was very low (3.10%), whereas previous studies have demonstrated values ranging from 24% (Hintze & Matthews, 2004) to 50% (Ferguson et al., 2012). It is difficult to make comparisons across these studies due to the distinct procedures utilized, perhaps most especially when the use of occasion as the object of measurement in the current study is compared to that of person, which was utilized as the object of measurement in the three previously-mentioned studies. However, distinctions between the current and previous studies may provide some initial explanation for the differences in residual variance proportions across studies.

First, training of behaviors was conducted with operational definitions, examples, and non-examples, followed by explicit training aligned to a master code on those specific behavior definitions. As identified by Hintze and Matthews (2004), large residual components may be attributed to inconsistent application of the dependent variable's definition; in other words, raters may not consistently agree upon what constitutes "academic engagement." The present study provided multiple opportunities for teaching and practicing the application of this definition, and as a result, the amount of random error present within this study may be somewhat lessened. Furthermore, rater variance was explicitly examined within this study, unlike two of three prior

research studies examining time-sampling data (i.e., Hintze & Matthews, 2004; Ferguson et al., 2012). For example, Hintze and Matthews (2004) utilized five raters during data collection, but did not include variance among rater estimates as a facet in subsequent analyses (due to the distribution of ratings across raters, rather than the crossing of observations with raters). As a result, variance that may have been attributable to differences between raters may have been subsumed under the residual component. Finally, unlike any of the three identified studies of time-sampling-derived data using academic engagement, the present study involved observations of a single student across occasions, rather than multiple students across occasions. This single student was also coached to engage in academically engaged or disruptive behavior. The contrived nature of the behaviors being observed, coupled with observations of a single student across all observations, may have resulted in easily identifiable forms of academic engagement that were able to be consistently applied to a single student. When rating multiple students, application of a single behavior definition may prove more challenging given distinct behavior topographies across students, which may subsequently result in increased variance attributable to a residual component.

Research Question 3 (Condition-combination differences). Within each time-sampling methodology, does rating additional behaviors simultaneously affect the amount of variance in academic engagement ratings attributable to facets other than the object of measurement? If so, which methodologies are least susceptible to increasing error, and which are most susceptible?

It was hypothesized that, across methodologies, rating additional behaviors simultaneously would result in incremental decreases in the proportion of variance attributable to the object of measurement. It was further hypothesized that the percentage of variance

attributable to the object of measurement would be highest for PI recording, followed by MTS, and lowest for WI recording. Although the fully-random model provides information about general sources of variance in estimates of academically engaged behavior, conceptually-sensible D studies cannot be developed from the components as written. A fully-random model interpreted within a generalizability framework would ignore the fixed nature of the method and behavior-targets facets, and a mixed model treating the method and behavior-targets facets as fixed would assume that ratings were averaged across protocols using all conditions of both time-sampling method and simultaneous behavior targets. Given the improbability of a measurement context wherein average ratings were derived using multiple protocol types, separate G studies were conducted in order to estimate variance components for both interpretation and later use in D studies.

Results from separate G studies for each combination of methodology and simultaneous behavior targets suggested, similar to the results of the fully-random model, that few systematic patterns could be discerned as a function of the behavior targets facet. Thus, no support was found for the hypothesis that an increase in simultaneous behavior targets would be related to a decrease in variance attributable to the object of measurement. Despite the significant methodological differences between Frame (1979) and the present study, as well as limitations on generalizability for the current study to the six clips observed, similar conclusions may be drawn regarding the effect of additional behaviors observed simultaneously upon variance in SDO-derived ratings (or, in the case of Frame, agreement). As in Frame (1979), no clear differences were found between simultaneous behavior-target conditions.

Although ratings derived from MTS procedures demonstrated slightly larger variance components for the object of measurement than those derived from other time-sampling

procedures, when compared within behavior-target conditions, these differences were extremely small to minimal. Similarly, the WI condition demonstrated relatively larger variance components for the rater facet when compared to the MTS and PI conditions, although these differences were also on the order of a few percentage points. Given the small size of the differences between components, as well as the anomalous behavior of the PI condition with 5 behavior targets, it is difficult to discern if this pattern suggests that MTS is less-susceptible to non-desirable sources of variance than PI or WI sampling, and WI perhaps more susceptible, or instead if this is simply an anomaly within the data. Similar to the findings of the full model, across all separate o x r models, the percentage of unexplained variance (i.e., that attributed to the interaction/residual term) was consistently very low.

Research Question 4 (Generalizability and Dependability coefficients). How many raters would need to simultaneously rate a student in order to achieve a generalizability and dependability coefficient of at least 0.90? How does this vary across time-sampling procedure and simultaneous behavior target combinations?

It was hypothesized that more than one rater would need to simultaneously rate a student in order to demonstrate a generalizability and dependability coefficient above 0.90. Similar to the hypothesis in the third research question, more raters were expected to be necessary for (ranked from more to less): WI recording, MTS, and PI sampling, with more raters required for higher numbers of simultaneous behavior targets. Coefficients depicting the relative amount of variance attributable to desirable (versus undesirable) sources may help to simplify variance component interpretation, as well as couch conclusions in a framework of relative- or absolute- decision-making. These frameworks describe the relative reliability of decisions made when scores are compared across students (relative) and within students (absolute), respectively. Given the

relatively high amount of variance attributable to the object of measurement when separate G studies were conducted, it is unsurprising that the generalizability and dependability coefficients for each of these models in the original measurement condition are extremely high. Furthermore, these coefficients are constructed for an estimate of academic engagement that is averaged over 10 raters, permitting an even greater reduction in the amount of variance attributable to sources other than the object of measurement. Results of generalizability and dependability coefficients across all combinations of methodology and behavior-targets conditions suggest that, when estimates of academic engagement are averaged over 10 raters, these estimates are overwhelmingly influenced by the occasion itself, rather than the rater, the rater and occasion interaction, or random error.

However, a measurement procedure wherein 10 raters observed the same 10 min of behavior and the average of their academic engagement ratings was utilized as the final estimate is likely impossible in a real-world setting. Thus, D studies were utilized in order to discover the minimum number of raters over whom a rating would need to be averaged in order to derive a highly dependable (≥ 0.900) estimate of academically engaged behavior for the six experimental clips observed within this study. When results were extrapolated such that only one rater was utilized in order to derive an estimate of academically-engaged behavior, both generalizability and dependability coefficients were very high, with a minimum observed value of 0.845 and 0.831 for their respective coefficients. These lowest values were observed for the anomalous condition combination of PI time-sampling and five behavior targets; generalizability coefficients for all other combinations of methodology and behavior targets exceeded 0.900 with only one rater. When ratings of academic engagement were averaged over two raters (e.g., a D study where $n_r = 2$), all combinations (including PI with five targets) demonstrated

generalizability coefficients above 0.900. These results suggest that, for the six clips measured within this study, an estimate of academic engagement could be derived by rating a student's behavior using any methodology or number of simultaneous behavior targets, as long as the final estimate of academic engagement was averaged across two randomly-selected raters.

Given that data derived from time-sampling instruments may be assumed to be most often used to make absolute decisions about students (e.g., when progress-monitoring an individual student's response to intervention by comparing a baseline and intervention phase), dependability coefficients may provide the most instructive information regarding the relative reliability of an estimate of academic engagement derived from time-sampling procedures. Unlike the results for generalizability coefficients, five out of twelve dependability coefficients were below 0.900 when ratings were extrapolated for only one rater. However, similar to the results for generalizability coefficients, all dependability coefficients rose above 0.900 when ratings were averaged across two raters. These results provide support for the hypothesis that ratings would need be averaged over more than one rater in order for measurement conditions to yield generalizability and dependability coefficients about 0.900 for the six clips observed in this study. However, no support was found for the hypothesis that systematic differences would be observed in the number of raters required for measurement contexts with distinct types of methodology or greater numbers of simultaneous behavior targets.

Thus, regardless of the type of decision-making utilized, averaging ratings across two raters resulted in very high levels of reliability for any combination of methodology and number of simultaneous behavior targets for the six clips observed within this study. If ratings were only used for relative decision-making, all but one of the twelve rating procedures tested within this study resulted in very high generalizability coefficients with only one rater. If ratings were used

for absolute decision-making, five of the twelve tested measurement procedures would require ratings to be averaged over two raters in order to demonstrate very high dependability coefficients.

Limitations

Concerns have been raised regarding the suitability of generalizability theory to behavioral observation data. These arguments, as outlined by Lei, Smith, and Suen (2007), derive from concerns regarding (a) autocorrelation, (b) finite observation length, and (c) uncertainties regarding how generalizability theory functions when data are considered for use in single-subject analyses. Although the first concern may not bear significantly upon the current study due to the contrived nature of the stimulus materials, discussion regarding the utilization of generalizability theory with data in single-subject designs is warranted. Furthermore, conclusions regarding the measurement contexts to which these results are generalizable may need to be restricted to 10-min observation periods. The discussion of Lei, Smith, and Suen (2007) regarding the applicability of generalizability theory to observational data is derived from the use of person as the object of measurement in traditional applications of generalizability theory (Brennan, 2001). Single-subject research is typically concerned with data derived from individual observations of persons, rather than observations over persons, highlighting the importance of correct model specification when designing a generalizability analysis.

Thus, in an effort to continue to apply generalizability theory to behavioral observation data, particularly data intended for use in single-subject design frameworks, Lei, Smith, and Suen (2007) proposed a number of recommendations for model specification in applications of generalizability theory, which were utilized within this study. These recommendations include: (a) the object of measurement should be specified as occasion, rather than subject (as

implemented within this study); and (b) results should only be generalized for that specific individual, for only those time points observed. Given the nature of single-subject design, such restrictions are unsurprising. The conclusions of these authors suggest that results may only be generalizable to these specific data; results may be best presented as a potential outcome when applying direct observation to these specific clips, and then perhaps observations of academic engagement derived from an elementary-aged student, rather than the results that would be guaranteed across all students and contexts. When direct observation is used for generalized characteristics (e.g., taking a mean over multiple observations, using structured observation environments), applications of generalizability theory may require less restrictions for appropriate use (Yoder & Symons, 2010).

Although possible threats to the internal validity of this study were addressed through procedures including quasi-random assignment of clip ordering and rating procedures, some threats to external validity may be present depending on the measurement contexts to which results are generalized. As described above, results of this study are most appropriately interpreted exclusively within the context of the behaviors being observed: in this case, six 10-minute video clips of contrived behavior for a single, male, elementary-aged student. Although these clips were manipulated in order to display a diverse assortment of behaviors of varying duration, and thus may generalize to behavior constellations beyond those observed within this study, the analyses conducted and results generated were wholly dependent on occasions that were deliberately, rather than randomly, selected. Raters for this study also constituted a convenience sample of participants; although procedures were implemented in order to ensure a basic level of similarity across all participants in terms of both their prior experience with SDO and their

preparation for study procedures, their individual characteristics and subsequent ratings may be specific to a graduate student or research assistant population.

Training procedures in this study were designed to ensure that participants rated target behaviors in a consistent and accurate manner while avoiding activities that may have biased subsequent study procedures. Thus, although participants received training on utilizing protocols with one behavior, they did not receive practice on rating multiple behaviors simultaneously, which may not resemble the training traditional research personnel would receive prior to engaging in SDO. Research personnel would likely conduct extensive training in order to achieve inter-observer agreement with one another, in contrast to school-based personnel who may be more likely to engage in observations with a definition of the target behavior, as well as limited practice in measuring that target behavior. Given that the training procedures utilized within this study may not adequately resemble training procedures used by researchers or practitioners, it is unclear to what extent these findings may generalize to these populations.

Other characteristics of the present study may limit its external validity. Stimuli for this study were video clips rather than in vivo student ratings, and it is therefore unclear if raters would have utilized protocols differently had they conducted their observations in school-based settings, rather than a research setting. Ratings were also completed in close proximity to one another; although breaks were provided every hour and three hours in order to combat rater fatigue, ratings were nonetheless completed in larger blocks than are likely found in a typical research or practice setting. Furthermore, although academically engaged behavior has been identified as a critical behavior for observation by both researchers and practitioners, the results of this study may not generalize to other behaviors that are more or less amenable to consistent recording by raters. It is also unclear to what extent the additional six “distractor” variables

utilized within the current study were present within each clip; the relative duration of these behaviors during each clip may not have been sufficient to reduce the variability attributable to the object of measurement. These results may therefore not readily generalize to other distractor behaviors of differing prevalence and frequency.

Finally, although clip order was systematically manipulated in order to avoid instances wherein participants viewed the same clip twice in a row, conditions for the methodology and behavior target facets were assigned randomly to clips, such that participants may have utilized the same methodology, number of target behaviors, or combination of conditions in succession. Random assignment was conducted in order to minimize any potential sources of systematic variance related to presentation order of conditions, although it is unclear whether this random assignment was sufficient in order to mitigate the influence of practice effects, rater fatigue, and other potential confounds.

Future Research

Future research may consider the use of planned missingness designs in order to address potential threats to validity resulting from the repeated rating of video clips (e.g., rater fatigue, practice effects). In planned missingness designs like the multiform questionnaire protocol, researchers take advantage of the statistical benefits of data patterns that are missing completely at random (MCAR) by planfully exposing participants to both a common and random subset of stimuli (Little, Jorgensen, Lang, & Moore, 2013). The exposure of participants to a subset of clips and protocol types, rather than every possible combination, may provide a robust method for addressing the fatigue and practice concerns that result from a large, fully-crossed design like that utilized in the present study, while also increasing the amount of resources available for participant recruitment and compensation.

Given the critical role of interval length in the relative level of rating bias derived from distinct time-sampling methods (Suen & Ary, 1989), future research may also focus upon rating variance as a function of interval length and this facet's interactions with other dimensions of interest (e.g., methodology, number of simultaneous behavior targets). While both the current study and previous GT-based studies of time-sampling methods have utilized 15 s intervals, this interval length may not consistently result in unbiased estimates of a continuous behavior stream. Intervals derived from formulae which take into account estimates of bout length and interresponse time, as suggested by Suen and Ary (1989), may provide suggestions for condition selection that are more likely to demonstrate both reliable and accurate estimates of target behavior.

Although this study focused exclusively on academically-engaged behavior as the target outcome variable, other behaviors may be relatively “harder” to measure reliably when measured simultaneously with other behaviors, or may be more suited to estimates of frequency rather than duration. The current study also examined behaviors which were modeled for and coached to be exhibited by the target student. As a result, behaviors in the current study may have been more amenable to consistent application of a target definition. Further research should therefore be conducted targeting other duration-based outcome variables as well as non-continuous behaviors, as well as behaviors that are naturally-occurring rather than artificially encouraged. Similarly, future research may also benefit from explorations of distinct sets of simultaneous behavior targets, with a particular emphasis on the relative frequency and duration of each additional behavior. These directions for future research notwithstanding, replications of the current study are encouraged in order to generalize current findings.

Conclusion

This study utilized generalizability theory analyses in order to determine whether the effects of time-sampling methodology and simultaneous behavior targets influenced the relative amount of variance observed among ratings of academic engagement. Results suggested that, for observations of six 10-minute video clips of an elementary-aged boy conducted by participants with experience comparable to a typical graduate research assistant, the majority of variance in ratings of academic engagement was attributable to the object of measurement, regardless of the specific combination of time-sampling methodology and number of simultaneous behavior targets utilized. No support was found for an effect of increasing error dependent upon an increasing number of simultaneous behavior targets. Reliability-like coefficients suggested that for any combination of methodology and simultaneous behavior targets, ratings averaged across two simultaneous raters would demonstrate generalizability and dependability coefficients above 0.900 for the six clips observed within this study.

Implications derived from this study for both researchers and practitioners are anchored in the demonstration that academic engagement, an important behavior in the promotion of positive student outcomes, may be rated consistently using time-sampling procedures. The results from this study suggest that the reliability of academic engagement ratings may not be significantly impacted by simultaneously measuring other behaviors alongside academic engagement. Furthermore, the reliability of these ratings may not change as function of the specific time-sampling methodology used. Researchers may utilize this information in order to support the development of measures that employ small to large numbers of simultaneous behavior targets when academic engagement is a primary dependent variable of interest. Although these conclusions may be best restricted to raters who possess levels of experience equivalent to that of a graduate research assistant who has completed at least one course in

behavior assessment, these individuals may often serve as raters of student behavior in a typical research study.

Generalization of findings to applied settings may be more difficult, given the likely disparity between the training procedures utilized within this study and the amount of training and familiarization practitioners may engage in immediately prior to conducting in-class observations. However, overall results from this study suggest that these individuals, with training to criterion on rating individual behaviors as well as utilization of time-sampling methodologies, are able to produce estimates of academic engagement that are most directly influenced by the behaviors being observed, rather than systematic rater-generated errors due to person- or protocol-level influence. Given the frequency with which SDO and specifically time-sampling procedures are utilized by educational professionals (e.g., Wilson & Reschly, 1996; Shapiro & Heick, 2004), these implications provide important initial support for the reliability of academic engagement data that may be derived by practitioners as they engage in these ratings. Although these results should be couched within the specific behavior conditions and student observed within this study, these results provide preliminary evidence for the production of reliable ratings of academic engagement, regardless of the type of methodology utilized and number of simultaneous behavior targets observed.

References

- Altmann, J. (1974). Observational study of behavior: Sampling methods. *Behaviour*, 49, 227-267. doi:10.1163/156853974X00534
- Arrington, R. E. (1943). Time sampling in studies of social behavior: A critical review of techniques and results with research suggestions. *Psychological Bulletin*, 40, 81-124. doi: 10.1037/h0053539
- Audacity Team (2012). Audacity (Version 2.0.2) [Computer software]. Retrieved from <http://audacity.sourceforge.net/>
- Bates, D., Maechler, M., Bolker, B. & Walker, S (2014). lme4: Linear mixed-effects models using Eigen and S4. (Version 1.1-5) [Computer software]. Retrieved from <http://CRAN.R-project.org/package=lme4>
- Bindra, D., & Blond, J. (1958). A time-sample method for measuring general activity and its components. *Canadian Journal of Psychology*, 12, 74-76. doi: 10.1037/h0083729
- Born, D. G., & Davis, M. L. (1974). Amount and distribution of study in a personalized instruction course and in a lecture course. *Journal of Applied Behavior Analysis*, 7, 365-375. doi: 10.1901/jaba.1974.7-365
- Brennan, R. L. (2001). *Generalizability Theory*. New York: Springer-Verlag.
- Brennan, R. L., & Kane, M. T. (1977). An index of dependability for mastery tests. *Journal of Educational Measurement*, 14, 277-289. doi: 10.1111/j.1745-3984.1977.tb00045.x
- Briesch, A. M., Chafouleas, S. M., & Riley-Tillman, T. (2010). Generalizability and dependability of behavior assessment methods to estimate academic engagement: A comparison of systematic direct observation and Direct Behavior Rating. *School Psychology Review*, 39, 408-421.

- Briesch, A. M., Swaminathan, H., Welsh, M., Chafouleas, S. M. (2014). Generalizability theory: A practical guide to study design, implementation, and interpretation. *Journal of School Psychology*. Advance online publication. doi: 10.1016/j.jsp.2013.11.008
- Brulle, A. R., & Repp, A. C. (1984). An investigation of the accuracy of momentary time sampling procedures with time series data. *British Journal of Psychology*, 75, 481-485. doi: 10.1111/j.2044-8295.1984.tb01917.x
- Chafouleas, S. M. (2011). Direct Behavior Rating: A review of the issues and research in its development. *Education & Treatment of Children*, 34, 575-591. doi: 10.1353/etc.2011.0034
- Chafouleas, S. M., Riley-Tillman, T. C., & Sugai, G. (2007). *School-based behavioral assessment: Informing intervention and instruction*. New York: Guilford.
- Cone, J. D. (1977). The relevance of reliability and validity for behavioral assessment. *Behavior Therapy*, 8, 411-426. doi: 10.1016/S0005-7894(77)80077-4
- Cooper, J. O., Heron, T. E., & Heward, W. L. (2007). *Applied behavior analysis* (2nd ed.). New Jersey: Pearson.
- Crocker, L. & Algina, J. (2008). *Introduction to classical and modern test theory*. Ohio: Cengage.
- Cronbach, L. J. (1971). Test validation. In R. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443-507). Washington, DC: American Council on Education.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: John Wiley.

- DiPerna, J. C. & Elliott, S. N. (2000). Academic Competency Evaluation Scales. San Antonio, TX: Psychological Corporation.
- DiPerna, J. C., & Elliott, S. N. (2002). Promoting academic enablers to improve student achievement: An introduction to the mini-series. *School Psychology Review*, 31, 293-297.
- DiPerna, J., Volpe, R. J., & Elliott, S. N. (2002). A model of academic enablers and elementary reading/language arts achievement. *School Psychology Review*, 31, 298-312.
- EduG (Version 6.1e) [Computer software]. Neuchatel, Switzerland: IRDP. Retrieved from <http://www.irdp.ch/edumetrie/englishprogram.htm>
- Ferguson, T. D., Briesch, A. M., Volpe, R. J., & Daniels, B. (2012). The influence of observation length on the dependability of data. *School Psychology Quarterly*, 27, 187-197. doi: 10.1037/spq0000005
- Final Cut Pro (Version 7.0.3) [Computer software]. Cupertino, CA: Apple.
- Frame, R. E. (1979). Interobserver agreement as a function of the number of behaviors recorded simultaneously. *The Psychological Record*, 29, 287-296.
- Goodenough, F. L. (1928). Measuring behavior traits by means of repeated short samples. *Journal of Juvenile Research*, 12, 230-235.
- Green, S. B. & Alverson, L. G. (1978). A comparison of indirect measures for long-duration behaviors. *Journal of Applied Behavior Analysis*, 11, 530. doi: 10.1901/jaba.1978.11-530
- Greenwood, C. R., Horton, B. T., & Utley, C. A. (2002). Academic engagement: Current perspectives in research and practice. *School Psychology Review*, 31, 328-349.
- Greenwood, C. R., & Terry, B. (1994). Confirming a performance-based instructional model. *School Psychology Review*, 23, 652-669.

- Gresham, F. M. (2004). Current status and future directions of school-based behavioral interventions. *School Psychology Review*, 33, 326–343.
- Harris, F. C., & Lahey, B. B. (1982). Recording system bias in direct observational methodology: A review and critical analysis of factors causing inaccurate coding behavior. *Clinical Psychology Review*, 2, 539-556. doi: 10.1016/0272-7358(82)90029-0
- Harrop, A., & Daniels, M. (1986). Methods of time sampling: A reappraisal of momentary time sampling and partial interval recording. *Journal of Applied Behavior Analysis*, 19, 73-77. doi: 10.1901/jaba.1986.19-73
- Hayes, S. C., Nelson, R. O., & Jarrett, R. B. (1986). Evaluating the quality of behavioral assessment. In R. Nelson & S. Hayes (Eds.), *Conceptual foundations of behavioral assessment* (pp. 463-503). New York: Guilford.
- Hintze, J. M. (2005). Psychometrics of direct observation. *School Psychology Review*, 34, 507-519.
- Hintze, J. M. & Matthews, W. J. (2004). The generalizability of systematic direct observations across time and setting: A preliminary investigation of the psychometrics of behavioral observation. *School Psychology Review*, 33, 258-270.
- Hothorn, T. & Hornik, K. (2013). exactRankTests: Exact distributions for rank and permutation tests (Version 0.8-27) [Computer software]. Retrieved from <http://CRAN.R-project.org/package=exactRankTests>
- Individuals with Disabilities Education Act Amendments of 1997, Pub. L. no. 105-17, 111 Stat 37 (1997).
- Kazdin, A. E. (2010). *Single-case research designs: Methods for clinical and applied settings* (2nd ed.). Oxford: New York.

- Kearns, K., Edwards, R., & Tingstrom, D. H. (1990). Accuracy of long momentary time-sampling intervals: Implications for classroom data collection. *Journal of Psychoeducational Assessment*, 8, 74-85. doi: 10.1177/073428299000800109
- Lei, P., 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, 433-439. doi:10.1002/pits.20235
- Little, T. D., Jorgensen, T. D., Lang, K. M., & Moore, E. W. G. (2013). On the joys of missing data. *Journal of Pediatric Psychology*. Advance online publication. doi: doi:10.1093/jpepsy/jst048
- Marcus, R. F., Johnson, C., & Roke, E. J. (1980). *Application of the "generalizability coefficient" to the study of the stability of cooperative behavior of young children*. Paper presented at the 88th Annual Convention of the American Psychological Association, Montreal, Canada.
- Mash, E. J., & McElwee, J. D. (1974). Situational effects on observer accuracy: Behavioral predictability, prior experience, and complexity of coding categories. *Child Development*, 45, 367-377. doi:10.2307/1127957
- 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. doi: 10.1177/105381519401800104
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50, 741-749. doi: 10.1037/0003-066X.50.9.741

- National Association of School Psychologists (2006). *School Psychology: A Blueprint for Training and Practice III*. Retrieved from <http://www.nasponline.org/resources/blueprint/FinalBlueprintInteriors.pdf>
- Olson, W. C. (1929). *The Measurement of Nervous Habits in Normal Children*. Minneapolis, MN: University of Minnesota Press.
- Peterson, R. F., Cox, M. A., & Bijou, S. W. (1971). Training children to work productively in classroom groups. *Exceptional Children*, 37, 491-500.
- Powell, J., Martindale, A., & Kulp, S. (1975). An evaluation of time-sample measures of behavior. *Journal of Applied Behavior Analysis*, 8, 463-469. doi: 10.1901/jaba.1975.8-463
- R Development Core Team (2012). R: A language and environment for statistical computing (Version 2.14.2) [Computer software]. Vienna, Austria. Retrieved from <http://www.r-project.org/>
- Reynolds, C. R., & Kamphaus, R. W. (2004). *Behavior assessment system for children* (2nd ed.). Minneapolis, MN: Pearson.
- Rogosa, D., & Ghandour, G. (1991). Statistical models for behavioral observations. *Journal of Educational Statistics*, 16, 157-252. doi: 10.3102/10769986016003157
- RStudio (2012). RStudio: Integrated development environment for R (Version 0.96.122) [Computer software]. Boston, MA. Retrieved from <http://www.rstudio.org/>
- Salvia, J. & Hughes, C. (1990). *Curriculum-based assessment: Testing what is taught*. New York: Macmillan.
- Salvia, J., Ysseldyke, J., Bolt, S. (2010). *Assessment in special and inclusive education* (11th ed.). Belmont, CA: Wadsworth.

- Sanson-Fisher, R. W., Poole, A. D., & Dunn, J. (1980). An empirical method for determining an appropriate interval length for recording behavior. *Journal of Applied Behavior Analysis*, 13, 493-500. doi: 10.1901/jaba.1980.13-493
- Saudargas, R. A., & Lentz, F. R. (1986). Estimating percent of time and rate via direct observation: A suggested observational procedure and format. *School Psychology Review*, 15, 36-48.
- Saudargas, R. A., & Zanolli, K. (1990). Momentary time sampling as an estimate of percentage time: A Field validation. *Journal Of Applied Behavior Analysis*, 23, 533-37. doi: 10.1901/jaba.1990.23-533
- Shapiro, E. S. (2004). *Academic skills problems workbook* (Revised ed.). New York: Guilford.
- Shapiro, E. S., & Heick, P. F. (2004). School psychologist assessment practices in the evaluation of students referred for social/behavioral/emotional problems. *Psychology in the Schools*, 41, 551-561. doi: 10.1002/pits.10176
- Silva, F. (1993). *Psychometric foundations and behavioral assessment*. Newbury Park, CA: Sage.
- Suen, H. K., & Ary, D. (1984). Variables influencing one-zero and instantaneous time sampling outcomes. *Primates*, 25, 89-94. doi: 10.1007/BF02382298
- Suen, H. K., & Ary, D. (1989). *Analyzing quantitative behavioral observation data*. Hillsdale, NJ: Erlbaum.
- Suen, H. K., Lee, P. S., & Owen, S. V. (1990). Effects of autocorrelation on single-subject single-facet crossed-design generalizability assessment. *Behavioral Assessment*, 12, 305-315.
- Sutherland, K. S., Lewis-Palmer, T., Stichter, J., & Morgan, P. L. (2008). Examining the

- influence of teacher behavior and classroom context on the behavioral and academic outcomes for students with emotional or behavioral disorders. *The Journal of Special Education*, 41, 223-233. doi:10.1177/0022466907310372
- Thomson, C., Holmberg, M., & Baer, D. M. (1974). A brief report on a comparison of time-sampling procedures. *Journal of Applied Behavior Analysis*, 7, 623-626.
doi:10.1901/jaba.1974.7-623
- Thorndike, R. L. (1947). *Research Problems and Techniques (Report No. 3)*. Washington, DC: US Printing Office.
- Volpe, R. J., DiPerna, J. C., Hintze, J. M., & Shapiro, E. S. (2005). Observing students in classroom settings: A review of seven coding schemes. *School Psychology Review*, 34, 454-474.
- Vue, K., Zieffler, A., Brown, E., Chang, Y & Stanke, L. (2013) QME: Reliability functions (Version 0.1) [Computer software]. Retrieved from <https://github.com/zief0002/QME>
- Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis*. New York: Springer.
- Wickham, H. (2011). The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, 40, 1-29.
- Wickham, H. (2012). stringr: Make it easier to work with strings (Version 0.6.2) [Computer software]. Retrieved from <http://CRAN.R-project.org/package=stringr>
- Wilson, M. S. & Reschly, D. J. (1996). Assessment in school psychology training and practice. *School Psychology Review*, 25, 9-23.
- Yoder, P. J. & Symons, F. J. (2010). *Observational measurement of behavior*. New York: Springer.

Table 1

Participant Characteristics

ID Number	Program, Year	Hours of Prior SDO Training	Undergrad Courses (n)	Graduate Courses (n)	Hours of SDO, Last Year	Hours of SDO, Total	Previous SDO Methods Used	Most Behaviors Observed (n)
01	PhD, 3	7	0	2	40	55	MTS, PI, FC, DUR	8
02	PhD, 3	4	0	1	10	12	MTS	3
03	PhD, 3	6	0	6	30	45	MTS, PI, FC	8
04	Masters, 1	8	0	1	75	75	MTS, FC	6
05	PhD, 1	5	0	1	30	30	MTS, PI, WI, FC	5
06	Masters, 1	3	0	1	35	35	MTS	3
07	Masters, 1	10	0	1	15	15	MTS, PI, WI, FC	8
08	PhD, 1	4	2	2	20	40	MTS, PI, WI, FC	5
09	PhD, 1	5	0	3	75	75	MTS	3
10	PhD, 4	20	0	2	80	200	MTS, PI, WI, FC	5
<i>M</i>	n/a	7.2	0.2	2	41	58.2	n/a	5.4
<i>SD</i>	n/a	5.0	0.6	1.6	26.2	54.3	n/a	2.1

Note. Program, Year = program level enrolled, followed by highest year completed in program. Undergrad Courses, Graduate Courses = number of courses completed which included SDO training or practice. MTS = momentary time-sampling. PI = partial-interval recording. WI = whole-interval recording. FC = frequency count. DUR = continuous duration recording. Most Behaviors Observed = maximum number of behaviors simultaneously recorded per interval using SDO procedures.

Table 2

Operational definitions of behaviors, with examples and non-examples utilized during training

Behavior	Definition	Examples	Non-examples
Academic Engagement	The student is actively or passively participating in the classroom activity.	writing, raising hand, talking about a lesson, listening to the teacher, reading silently, or looking at instructional materials.	staring into space, talking about off-topic activities.
Out of Seat	The student is out of his/her seat.	out of seat	in seat
Looking Around	The student is looking around and not engaged in any other activity.	staring into space, looking at posters on walls, looking at ceiling.	looking at teacher, writing on worksheet, talking to student about off-task topic, playing with pencil.
Motor Behavior	The student is engaged in repetitive, stereotyped body movements.	rocking in chair, tapping desk with finger, bobbing head, rubbing fingers together, poking other student repeatedly with finger.	tapping table with pencil, folding paper, throwing eraser, hitting self with magazine.
Playing with Object	The student is repetitively playing with an object.	tapping pencil, folding paper, playing with action figure, texting on phone	writing on worksheet with pencil, using calculator during math assignment.
Social Interaction w/ Child	The student is interacting with one or more other students.	talking to other student about weekend plans, listening to student as they speak to them, poking other student, throwing paper at other student.	looking at student without engaging in conversation, staring into space, attending to teacher.
Social Interaction w/ Teacher	The student is interacting with the classroom teacher.	talking to teacher about assignment, listening to teacher as they are reprimanded, accepting pencil from teacher	looking at teacher when teacher is addressing whole class, looking away from teacher as they are spoken to

Table 3

Descriptive statistics by clip, methodology, and behavior targets

	A	B	C	D	E	F
	M (SD)	M (SD)	M (SD)	M (SD)	M (SD)	M (SD)
Overall	0.86 (0.13)	0.61 (0.23)	0.86 (0.14)	0.83 (0.13)	0.68 (0.20)	0.23 (0.18)
Method						
MTS	0.85 (0.07)	0.63 (0.09)	0.90 (0.06)	0.85 (0.05)	0.67 (0.08)	0.23 (0.08)
PI	0.99 (0.01)	0.86 (0.08)	0.98 (0.02)	0.97 (0.02)	0.90 (0.10)	0.42 (0.12)
WI	0.73 (0.12)	0.35 (0.10)	0.71 (0.13)	0.68 (0.08)	0.47 (0.08)	0.05 (0.04)
MTS						
1	0.86 (0.09)	0.65 (0.08)	0.91 (0.05)	0.84 (0.05)	0.66 (0.05)	0.26 (0.08)
3	0.85 (0.06)	0.62 (0.11)	0.89 (0.06)	0.86 (0.04)	0.70 (0.11)	0.20 (0.07)
5	0.84 (0.07)	0.63 (0.08)	0.89 (0.08)	0.84 (0.08)	0.68 (0.08)	0.23 (0.09)
7	0.85 (0.05)	0.62 (0.09)	0.91 (0.07)	0.84 (0.04)	0.65 (0.07)	0.24 (0.10)
PI						
1	1.00 (0.01)	0.85 (0.06)	0.98 (0.02)	0.96 (0.02)	0.90 (0.07)	0.43 (0.13)
3	0.99 (0.01)	0.87 (0.08)	0.99 (0.02)	0.98 (0.02)	0.90 (0.09)	0.42 (0.12)
5	0.99 (0.02)	0.89 (0.09)	0.98 (0.03)	0.99 (0.01)	0.89 (0.17)	0.42 (0.14)
7	0.99 (0.01)	0.84 (0.10)	0.99 (0.02)	0.96 (0.03)	0.92 (0.07)	0.41 (0.13)
WI						
1	0.75 (0.09)	0.38 (0.11)	0.70 (0.17)	0.67 (0.07)	0.50 (0.08)	0.04 (0.03)
3	0.73 (0.12)	0.35 (0.11)	0.70 (0.10)	0.68 (0.09)	0.48 (0.08)	0.03 (0.03)
5	0.74 (0.16)	0.31 (0.08)	0.75 (0.12)	0.68 (0.08)	0.44 (0.04)	0.06 (0.06)
7	0.71 (0.12)	0.35 (0.11)	0.68 (0.14)	0.68 (0.09)	0.46 (0.11)	0.05 (0.03)

Table 4

Variance component estimates for all facets

Variance Component	N	With negative		Without negative	
		σ^2	%	σ^2	%
occasion (o)	6	0.05838	56.78	0.05838	56.74
rater (r)	10	0.00163	1.59	0.00163	1.58
method (m)	3	0.03174	30.87	0.03174	30.85
behavior targets (b)	4	0.00001	0.01	0.00001	0.01
o x r		0.00130	1.26	0.00130	1.26
o x m		0.00262	2.55	0.00262	2.55
o x b		-0.00003	-0.03	0.00000	0.00
r x m		0.00104	1.01	0.00104	1.01
r x b		0.00005	0.05	0.00005	0.05
m x b		-0.00004	-0.04	0.00000	0.00
o x r x m		0.00168	1.63	0.00168	1.63
o x r x b		0.00111	1.08	0.00111	1.08
o x m x b		0.00006	0.06	0.00006	0.06
r x m x b		0.00008	0.08	0.00008	0.08
o x r x m x b (residual)		0.00319	3.10	0.00319	3.10
Total		0.10282	100.00	0.10289	100.00

Note. As noted within the manuscript, interpretation of the full model was conducted utilizing "Without negative" variance components.

Table 5

Variance component estimates for separate o x r models, by method and behavior target

Facet	MTS		PI		WI	
	σ^2	%	σ^2	%	σ^2	%
One behavior						
o	0.058	92.37	0.045	91.17	0.070	86.88
r	0.002	2.85	0.001	1.81	0.004	5.45
or, e	0.003	4.78	0.003	7.03	0.006	7.68
Three behaviors						
o	0.068	91.42	0.049	91.06	0.072	89.28
r	0.003	3.64	0.001	2.64	0.004	4.88
or, e	0.004	4.94	0.003	6.30	0.005	5.84
Five behaviors						
o	0.060	90.18	0.047	83.08	0.077	88.62
r	0.002	2.90	0.001	1.70	0.003	3.86
or, e	0.005	6.92	0.009	15.22	0.006	7.52
Seven behaviors						
o	0.060	92.51	0.050	89.99	0.066	85.44
r	0.001	1.75	0.002	2.77	0.005	6.31
or, e	0.004	5.73	0.004	7.24	0.006	8.25

Table 6

G and D coefficients for original measurement procedures

Targets	MTS		PI		WI	
	G	D	G	D	G	D
1	0.995	0.992	0.992	0.990	0.991	0.985
3	0.995	0.991	0.993	0.990	0.993	0.988
5	0.992	0.989	0.982	0.980	0.992	0.987
7	0.994	0.992	0.992	0.989	0.990	0.983

Table 7

D study of G and D coefficient change as a function of number of simultaneous raters

Raters	MTS		PI		WI	
	G	D	G	D	G	D
One behavior						
1	0.951	0.924	0.928	0.912	0.919	0.869
2	0.975	0.960	0.963	0.954	0.958	0.930
3	0.983	0.973	0.975	0.969	0.971	0.952
4	0.987	0.980	0.981	0.976	0.978	0.964
5	0.990	0.984	0.985	0.981	0.983	0.971
6	0.991	0.986	0.987	0.984	0.985	0.975
7	0.993	0.988	0.989	0.986	0.988	0.979
8	0.994	0.990	0.990	0.988	0.989	0.981
9	0.994	0.991	0.992	0.989	0.990	0.983
10	0.995	0.992	0.992	0.990	0.991	0.985
Three behaviors						
1	0.949	0.914	0.935	0.911	0.939	0.893
2	0.974	0.955	0.967	0.953	0.968	0.943
3	0.982	0.970	0.977	0.968	0.979	0.962
4	0.987	0.977	0.983	0.976	0.984	0.971
5	0.989	0.982	0.986	0.981	0.987	0.977
6	0.991	0.985	0.989	0.984	0.989	0.980
7	0.992	0.987	0.990	0.986	0.991	0.983
8	0.993	0.988	0.991	0.988	0.992	0.985
9	0.994	0.990	0.992	0.989	0.993	0.987
10	0.995	0.991	0.993	0.990	0.993	0.988
Five behaviors						
1	0.929	0.902	0.845	0.831	0.922	0.886
2	0.963	0.948	0.916	0.908	0.959	0.940
3	0.975	0.965	0.942	0.936	0.973	0.959
4	0.981	0.973	0.956	0.952	0.979	0.969
5	0.985	0.979	0.965	0.961	0.983	0.975
6	0.987	0.982	0.970	0.967	0.986	0.979
7	0.989	0.985	0.975	0.972	0.988	0.982
8	0.990	0.987	0.978	0.975	0.990	0.984
9	0.992	0.988	0.980	0.978	0.991	0.986
10	0.992	0.989	0.982	0.980	0.992	0.987
Seven behaviors						
1	0.942	0.925	0.926	0.900	0.912	0.854
2	0.970	0.961	0.961	0.947	0.954	0.921
3	0.980	0.974	0.974	0.964	0.969	0.946
4	0.985	0.980	0.980	0.973	0.976	0.959
5	0.988	0.984	0.984	0.978	0.981	0.967
6	0.990	0.987	0.987	0.982	0.984	0.972
7	0.991	0.989	0.989	0.984	0.986	0.976
8	0.992	0.990	0.990	0.986	0.988	0.979
9	0.993	0.991	0.991	0.988	0.989	0.981
10	0.994	0.992	0.992	0.989	0.990	0.983

Figure 1

Distribution of prevalence score density by clip, behavior targets, and methodology

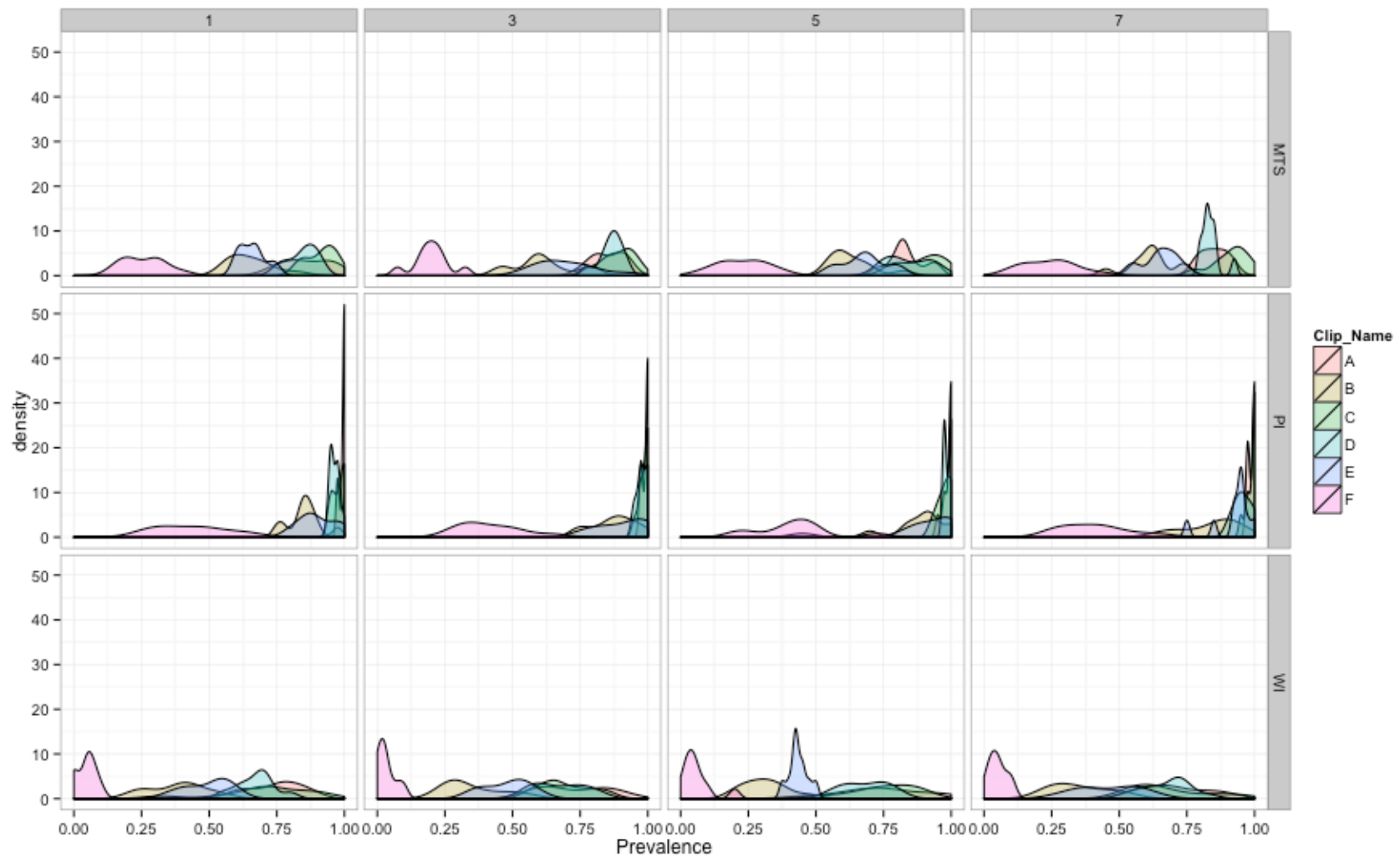


Figure 2

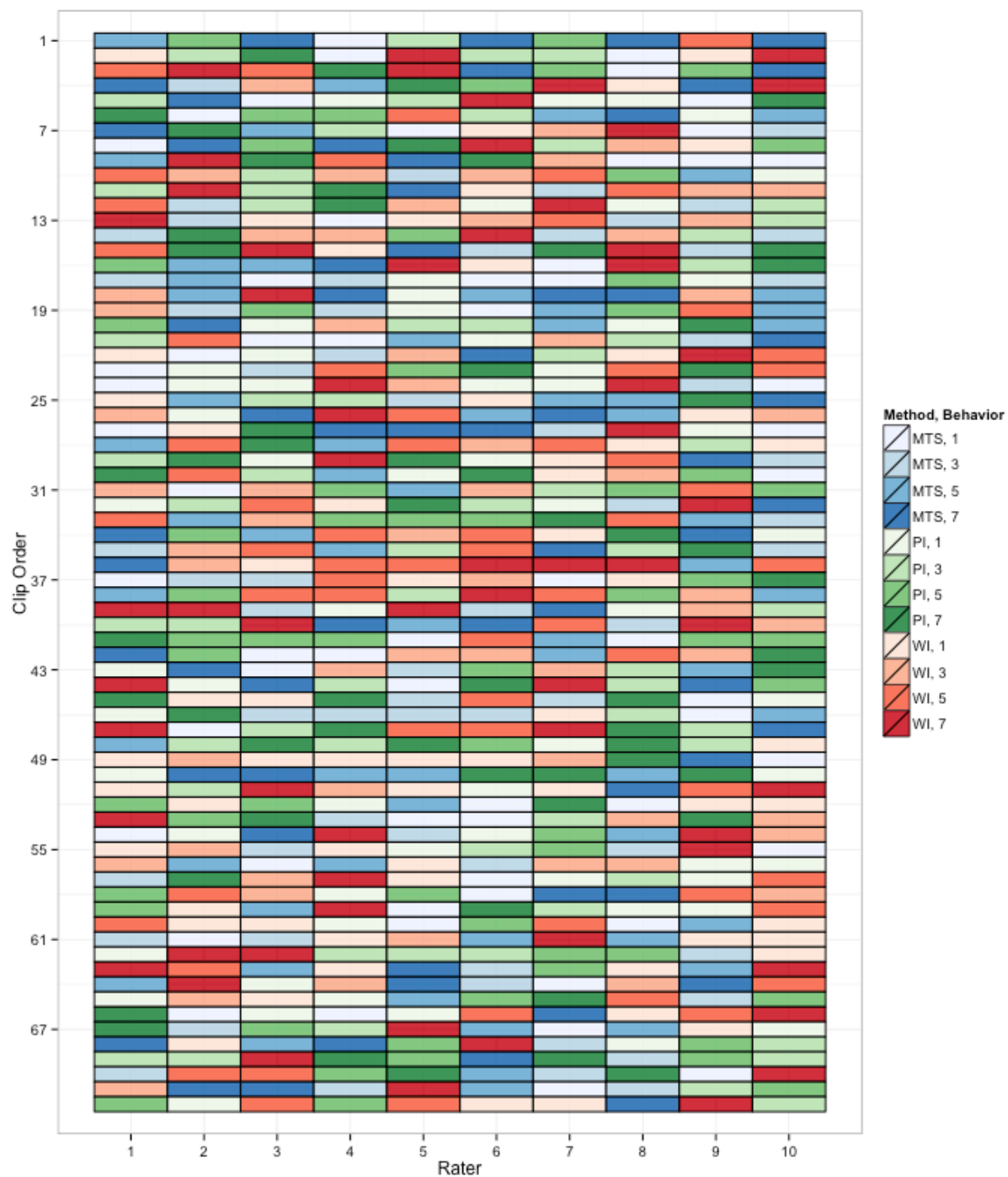
Assignment of clip order, behavior targets, and methodology by participant

Figure 3

Mean prevalence ratings for academic engagement by clip and methodology

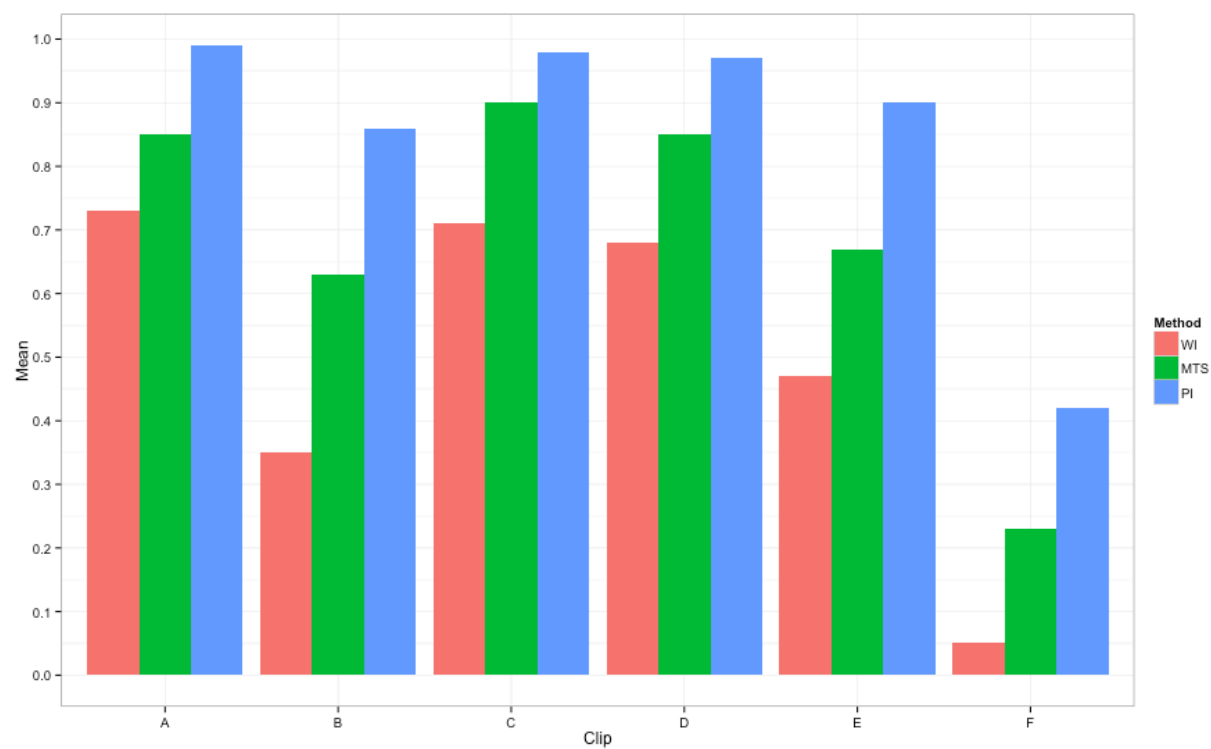


Figure 4

Variance component estimates for $o \times r \times m \times b$ as percentage of total variance

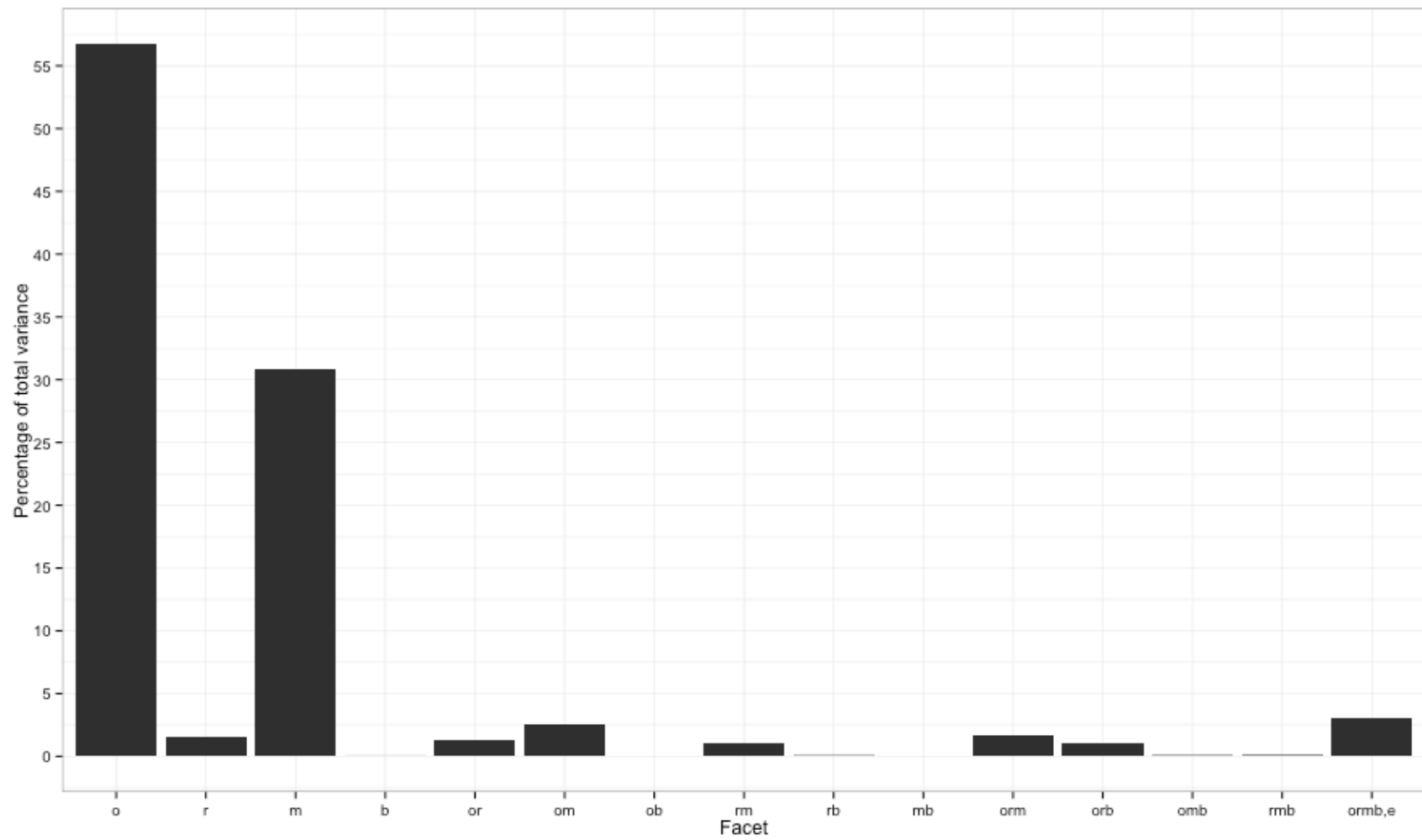


Figure 5

Variance component estimates for o x r models as percentage of total variance, by method and behavior targets

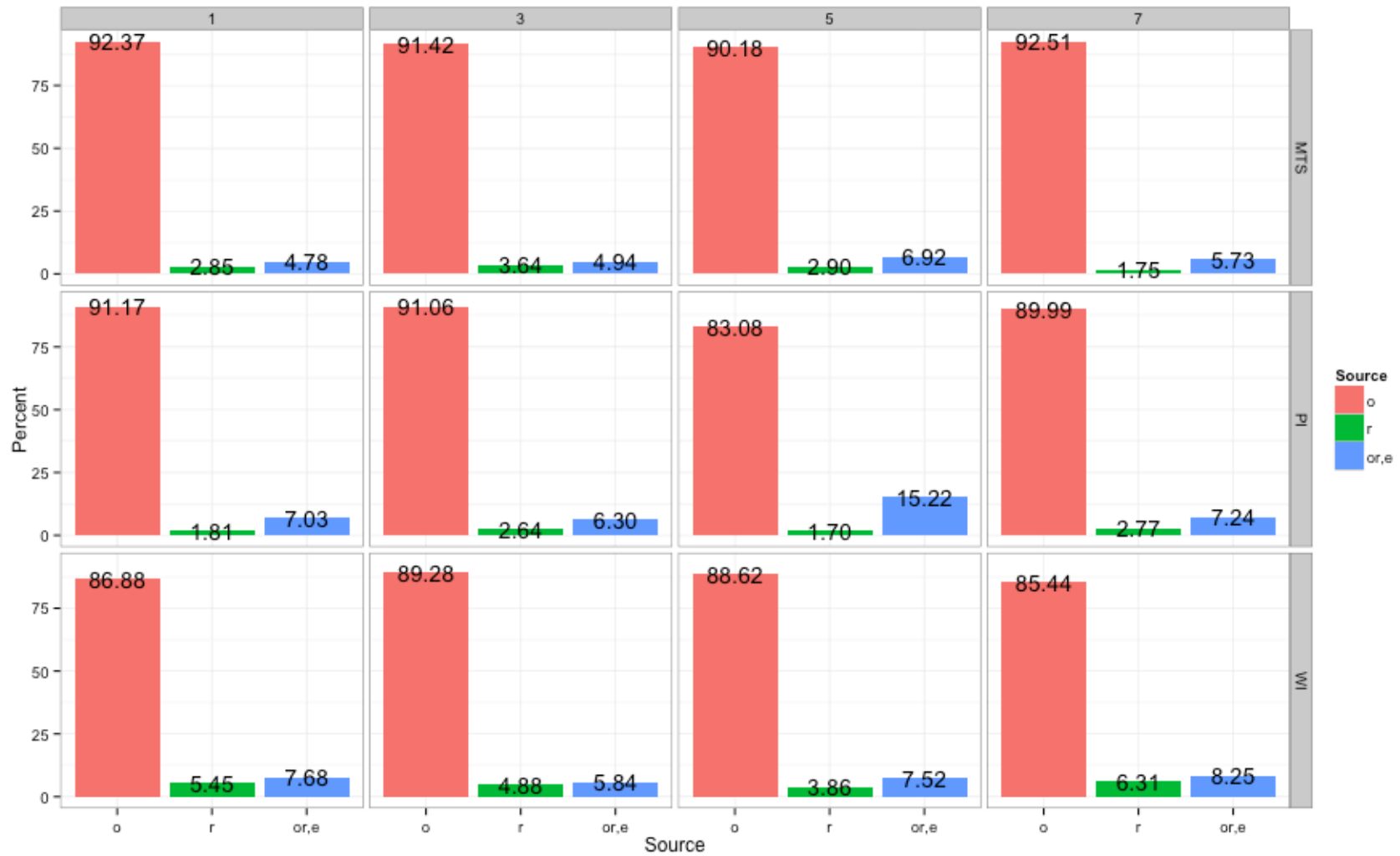


Figure 6

D study results for generalizability coefficients, with rater facet manipulated

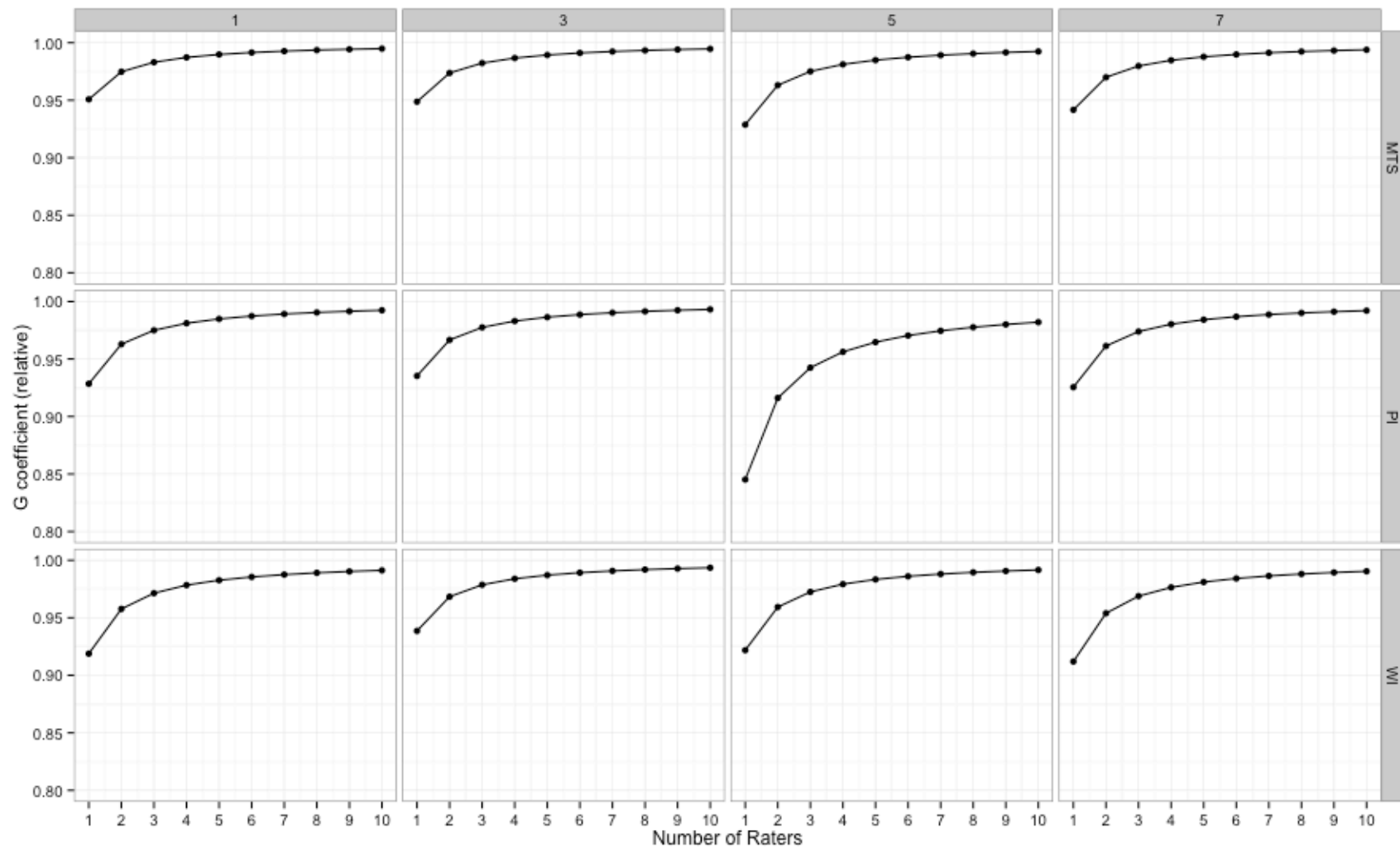
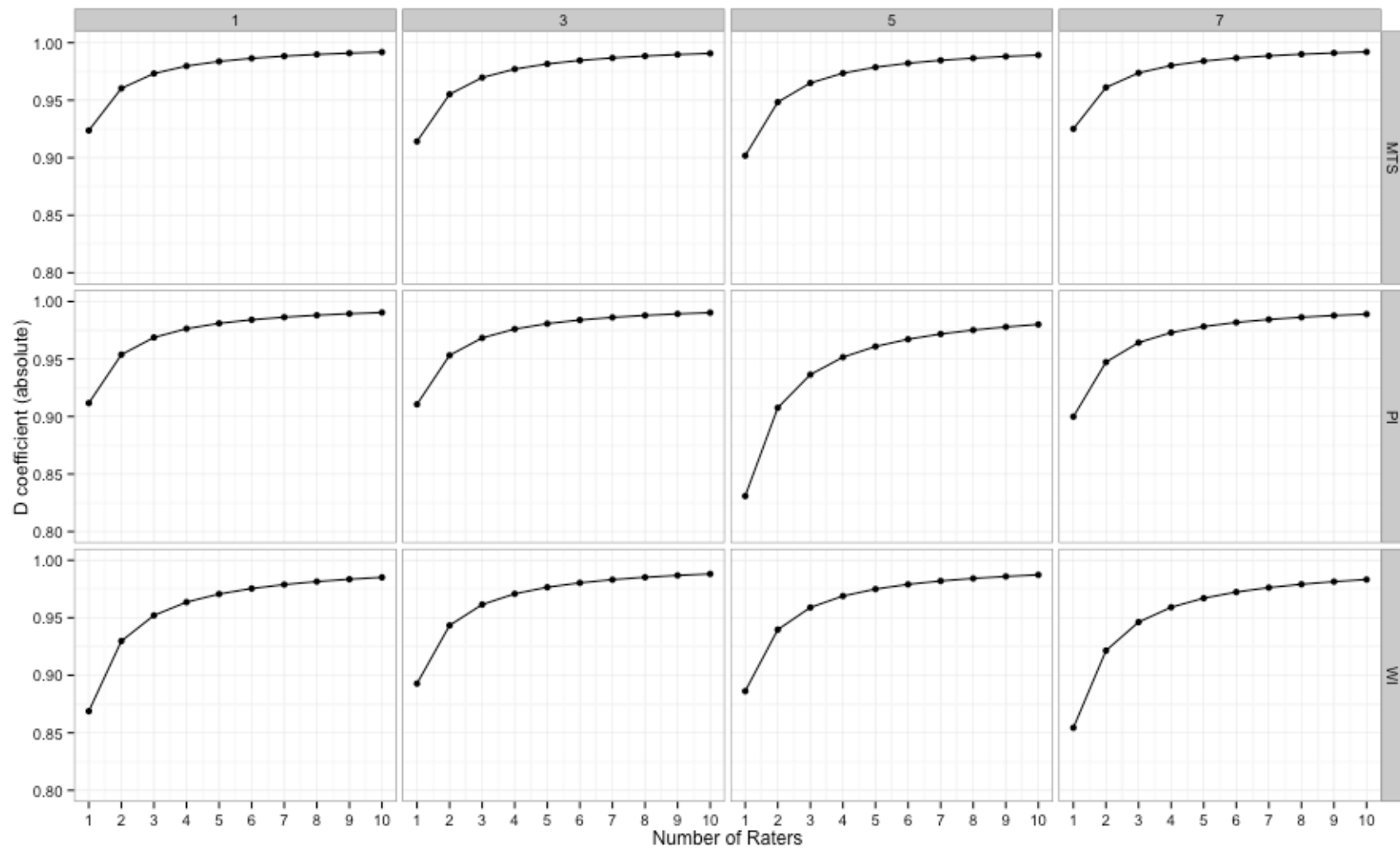


Figure 7

D study results for dependability coefficients, with rater facet manipulated



Appendix A

Sample Training Binder Sheet

Rater 01 : Rating #02

**Clip F**

Rate using **Whole interval**.

If the behavior occurs for the entire duration of the interval, score it as an occurrence.

Rate the following behavior:

1. Academic Engagement: *The student is actively or passively participating in the classroom activity.*

Note. Faces of students are obscured for dissemination purposes only. The original document did not obscure students' faces.

Appendix B

Sample Training Data Collection Form

ID01 : CLIP F : #02

Systematic Direct Observation Form - TRAINING

<u>Method</u>
Whole interval. If the behavior occurs for the entire duration of the interval, score it as an occurrence.
<u>Behaviors</u>
Academic Engagement. The student is actively or passively participating in the classroom activity.

Behavior	:15	:30	:45	1:00	1:15	1:30	1:45	2:00	2:15	2:30
Academic Engagement										

Behavior	2:45	3:00	3:15	3:30	3:45	4:00	4:15	4:30	4:45	5:00
Academic Engagement										

IOA Calculations

$\frac{\text{\# of agreements}}{20} = \text{IOA}$

$\frac{\quad}{20} =$

If IOA is less than 0.90, please use a new SDO form and re-code the clip.

If IOA is greater than 0.90, please advance to the next training clip.

Appendix C

Sample Experimental Binder Sheet

ID01 : CLIP B : PI_3 : #11

**Clip B**

Rate using **Partial interval**.

If the behavior occurs at any time during the interval, score it as an occurrence.

Rate the following behaviors:

- 1. Out of Seat:** *The student is out of his/her seat.*
- 2. Academic Engagement:** *The student is actively or passively participating in the classroom activity.*
- 3. Social Interaction w/ Teacher:** *The student is interacting with the classroom teacher.*

Note. Faces of students and teacher are obscured for dissemination purposes only. The original document did not obscure the teacher's or students' faces.

ID01 : CLIP B : PI_3 : #11

[illegible]

ID01 : CLIP E : MTS_7 : #07

<u>Method</u>   Momentary Time Sampling. If the behavior is occurring at the moment the interval ends, score it as an occurrence.
<u>Behaviors</u>   Playing with Object. The student is repetitively playing with an object.   Out of Seat. The student is out of his/her seat.   Academic Engagement. The student is actively or passively participating in the classroom activity.   Social Interaction w/ Teacher. The student is interacting with the classroom teacher.   Motor Behavior. The student is engaged in repetitive, stereotyped body movements.   Looking Around. The student is looking around and not engaged in any other activity.   Social Interaction w/ Child. The student is interacting with one or more other students.

[illegible][illegible]

Momentary Time Sampling. If the behavior is occurring at the moment the interval ends, score it as an occurrence.

Playing with Object. The student is repetitively playing with an object.

Academic Engagement. The student is actively or passively participating in the classroom activity.

Social Interaction w/ Teacher. The student is interacting with the classroom teacher.

Motor Behavior. The student is engaged in repetitive, stereotyped body movements.

Looking Around. The student is looking around and not engaged in any other activity.

Social Interaction w/ Child. The student is interacting with one or more other students.

[illegible][illegible]

Appendix F

Participant Experience Verification Form

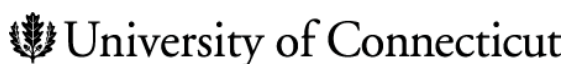
Please complete the following items regarding your experience with the use of systematic direct observation (SDO) protocols and measurement. If you are unsure of how to respond to a specific item, please provide your best estimate.

1. ID# _____
2. What is the highest year of your graduate program that you have completed? _____
3. Are you a Master's or PhD student?
☐ Master's
☐ PhD
4. Hours of previous SDO training: _____
5. How many undergraduate courses (3 credits or more) have you completed that involved the training in or use of SDO? _____
6. How many graduate courses (3 credits or more) have you completed that involved the training in or use of SDO? _____
7. In what year did you first use SDO? _____
8. Hours of SDO observations conducted in the last year: _____
9. Hours of SDO observations conducted total: _____
10. Which SDO time-sampling procedures have you used in actual non-training-based observations? (check all that apply)
☐ Momentary time-sampling
☐ Partial interval recording
☐ Whole interval recording
☐ Frequency count
☐ Latency
☐ Duration (continuous recording)
☐ Other: _____
11. What is the greatest number of behaviors you've observed per interval using SDO? _____

Appendix G

Participant Consent Form

Consent Form for Participation in a Research Study



Principal Investigator: Sandra Chafouleas, PhD

Student Researcher: Austin Johnson, MA

Study Title: Reliability of data derived from time sampling methods with multiple observation targets

Introduction

You are invited to participate in this study of how raters use systematic direct observation (SDO). I am a graduate student at the University of Connecticut, and I am conducting this study as part of my dissertation.

Why is this study being done?

I am interested in learning more about the variance in scores derived from systematic direct observation, and whether this varies by methodology and number of behaviors rated at a time.

What are the study procedures? What will I be asked to do?

Your participation in this study will require you to travel to your university to participate in (a) a training that will take approximately two hours, and (b) paper-and-pencil ratings of video clips depicting student behavior, which will take approximately twelve hours. All materials required for the study (laptop, headphones, forms, pencils) will be provided by the researcher.

During the training, you will be asked to participate in a review of SDO procedures and practice rating using nine 5-minute video clips of student behavior. All clips will be viewed on a laptop using headphones.

After training and practice, you will be asked to watch 10-minute videos of student behavior and rate target students within those videos using various types of SDO protocols. You will make your ratings using a paper SDO form and a pencil, and will watch the videos on a laptop with headphones. You may progress through the clips at your own pace, and will be asked to take a short break after every hour of rating, as well as an hour-long lunch break after three hours of ratings have been completed.

All procedures should take a total of two full workdays to complete.

What are the risks or inconveniences of the study?

You may be inconvenienced by the amount of time it takes to complete the study, which should take two full days to complete.

What are the benefits of the study?

The benefits of your participation may impact society by helping to increase knowledge about how systematic direct observation may be most appropriately designed and used.

Will I receive payment for participation? Are there costs to participate?

After you have completed all twelve hours of ratings, you will be given \$200 in cash to compensate you for your participation. All rating materials will be provided to you, and there are no direct costs associated with your participation outside of those acquired by commuting to and from the university. Lunch will be provided by the research team for both days of the study, with dietary restrictions respected for all participants.

How will my personal information be protected?

You will be identified by a two-digit code on all of your completed SDO forms. The principal investigator and student investigator will have access to a key that relates your two-digit code to your name, but this will not be shared with anyone else and will be secured in an encrypted folder on the student investigator's computer. The results of this study will only reported as a group; you will never be personally identified within any publications or presentations that result from this study.

You should also know that the UConn Institutional Review Board (IRB) and the Office of Research Compliance may inspect study records as part of its auditing program, but these reviews will only focus on the researchers and not on your responses or involvement. The IRB is a group of people who review research studies to protect the rights and welfare of research participants.

Can I stop being in the study and what are my rights?

You do not have to be in this study if you do not want to. If you agree to be in the study, but later change your mind, you may drop out at any time. There are no penalties or consequences of any kind if you decide that you do not want to participate. However, you must complete all twelve hours of rating in order to receive \$200 in compensation.

Whom do I contact if I have questions about the study?

Take as long as you like before you make a decision. We will be happy to answer any question you have about this study. If you have further questions about this project or if you have a research-related problem, you may contact the principal investigator, Dr. Sandra Chafouleas (860-486-6868, sandra.chafouleas@uconn.edu) or the student researcher, Austin Johnson (520-203-6798, austinj@gmail.com). If you have any questions concerning your rights as a research subject, you may contact the University of Connecticut Institutional Review Board (IRB) at 860-486-8802.

Documentation of Consent:

I have read this form and decided that I will participate in the project described above. Its general purposes, the particulars of involvement and possible hazards and inconveniences have been explained to my satisfaction. I understand that I can withdraw at any time. My signature also indicates that I have received a copy of this consent form.

Participant Signature:

Print Name:

Date:

Signature of Person
Obtaining Consent

Print Name:

Date:

Training Presentation Slides

Austin's dissertation!

Training in systematic direct observation (SDO)

VL3M7PDI created by [Austin Johnson](#)
Copyright © 2012 by the University of Connecticut
All rights reserved. Permission granted for personal and educational use as long as the names of the creators and the full copyright notice are included in all copies.

- Interested in studying variance in ratings as raters use different SDO methodologies (i.e., momentary time sampling, partial-interval sampling, whole-interval sampling) to rate different numbers of behaviors (i.e., 1, 3, 5, 7)

- ▶ 10 participants with prior experience and training in SDO
- ▶ 6 ten-minute videos of elementary-level classroom
- ▶ After training, participants will code each video multiple times using different methods

- ▶ Right now,
  - ▶ Orient to materials
  - ▶ Sign consent forms

- ▶ Then,
  - ▶ SDO overview
  - ▶ Method overview
  - ▶ Behavior overview
- ▶ After that,
  - ▶ Code training clips
- ▶ And finally,
  - ▶ Code experimental clips

- ▶ Right now,
  - ▶ Orient to materials
  - ▶ Sign consent forms

- ▶ **Then,**
  - ▶ SDO overview
  - ▶ Method overview
  - ▶ Behavior overview
- ▶ **After that,**
  - ▶ Code training clips
- ▶ **And finally,**
  - ▶ Code experimental clips

► Systematic direct observation is a method for collecting data pertaining to a person's behavior. It is systematic insofar as it is rule-bound and replicable. It is direct because it involves direct observation of student behavior, and therefore requires little inference from data to target construct.

Inference required from data collected to conclusions drawn



- ▶ All behaviors possess the qualities of both “events” and “states”.

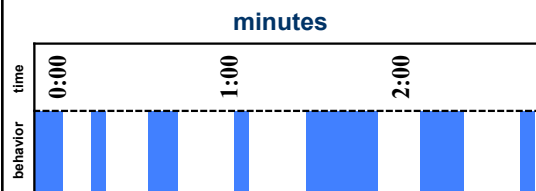
- ▶ Event: did it occur?
  - ▷ Count
- ▶ State: how long did it occur?
  - ▷ Duration, prevalence
- ▶ Depending on the behavior and RQ of interest, we may want to measure either the “trait” or “state” aspects of a given behavior.
- ▶ For this study, we’ll be focusing on the “state” aspects of our target behaviors, because we’re more interested in *how long* they occur than *how often* they occur.



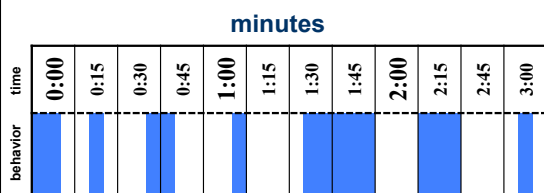
How is SDO used?

- ▶ In order to undertake SDO, we:
  - ▶ Select an observation period (e.g., a 10 minute portion of the school day)
  - ▶ Chunk that observation period into intervals of time
    - ▶ So, that 10 minute period becomes 40 15-second observation intervals (10 minutes x 60 seconds / 15 second intervals)
  - ▶ During each interval, observe the student
  - ▶ Record occurrences or nonoccurrences of behavior according to our given time-sampling procedure

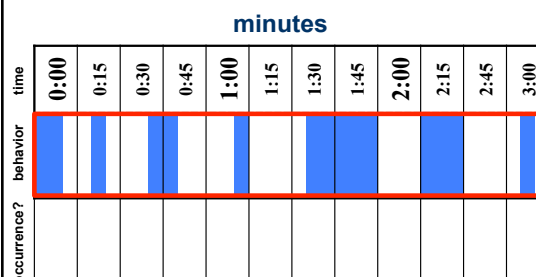
Hypothetical observation period exists



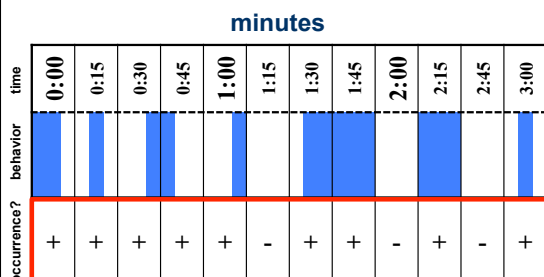
We can chunk it into intervals



And then use time-sampling rules



To determine if behavior occurred



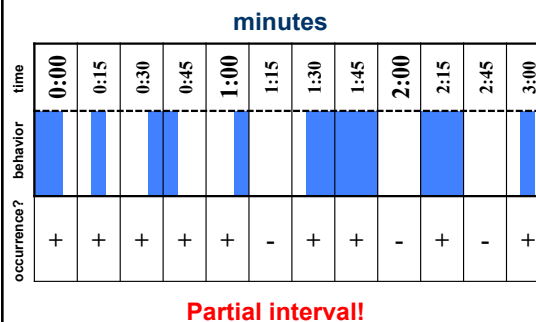
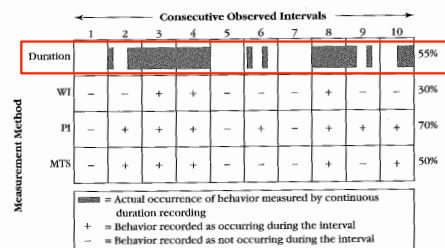
What's time sampling?

- ▶ What we think of as SDO often comes down to the time-sampling procedure we use. These are:
  - ▶ Partial interval
  - ▶ Whole interval
  - ▶ Momentary time sampling
- ▶ These are basically just decision rules for how we determine if what occurred during our 15-second interval counts as an occasion of the target behavior.

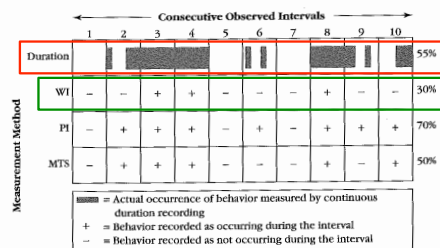
What's scored as an occurrence?

- ▶ **Partial interval**
  - ▶ If the behavior occurs at any point during the interval
- ▶ **Whole interval**
  - ▶ If the behavior occurs for the entire duration of the interval
- ▶ **Momentary time sampling**
  - ▶ If the behavior is occurring at the moment the interval ends

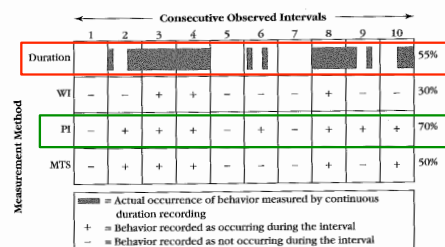
Which time-sampling procedure?



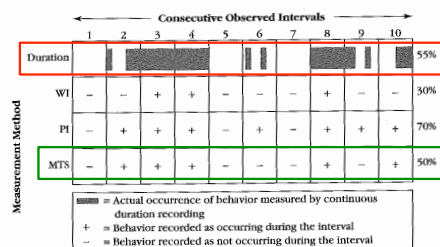
Continuous observation (the actual data stream)



Whole interval



Partial interval



Momentary time sampling

 What is the protocol we'll be using?

- 10 minute observations
- 15 second intervals
- Rating one student during the observation period
- 1, 3, 5, or 7 behaviors at a time
  - Always rating **academic engagement**, sometime rating others
- Using momentary time sampling, partial-interval, or whole-interval to determine if interval is scored as occurrence or nonoccurrence

This is a sample systematic direct observation form!

Systematic Direct Observation Form ID#1-CLIP-D-MTS-5-01

Behavior

Looking Around: The student is looking around and not engaged in any other activity.
Academic Engagement: The student is actively or passively participating in the classroom activity.
Motor Behavior: The student is engaged in repetitive, stereotyped body movements.
Social Interaction w/ Teacher: The student is interacting with the classroom teacher.
Playing with Object: The student is repetitively playing with an object.

Behavior	1:15	1:30	1:45	2:00	2:15	2:30	2:45	3:00	3:15	3:30	3:45	4:00
Looking Around												
Academic Engagement												
Motor Behavior												
Social Interaction w/ Teacher												
Playing with Object												

Up here are details about the observation.

ID#01: This is participant #01.

CLIP D: This observation form is completed while watching Clip D.

MTS_5: This observation form will be completed using momentary time sampling (MTS), and uses five behaviors rated during each interval.

#01: This is Participant #01's first observation.

Systematic Direct Observation Form ID#1-CLIP-D-MTS-5-01

Behavior

Looking Around: The student is looking around and not engaged in any other activity.
Academic Engagement: The student is actively or passively participating in the classroom activity.
Motor Behavior: The student is engaged in repetitive, stereotyped body movements.
Social Interaction w/ Teacher: The student is interacting with the classroom teacher.
Playing with Object: The student is repetitively playing with an object.

Behavior	1:15	1:30	1:45	2:00	2:15	2:30	2:45	3:00	3:15	3:30	3:45	4:00
Looking Around												
Academic Engagement												
Motor Behavior												
Social Interaction w/ Teacher												
Playing with Object												

Here's an easy-to-read explanation of:

(a) The method you're using
(b) The behaviors you'll be rating

for this observation.

Systematic Direct Observation Form ID#1-CLIP-D-MTS-5-01

Behavior

Looking Around: The student is looking around and not engaged in any other activity.
Academic Engagement: The student is actively or passively participating in the classroom activity.
Motor Behavior: The student is engaged in repetitive, stereotyped body movements.
Social Interaction w/ Teacher: The student is interacting with the classroom teacher.
Playing with Object: The student is repetitively playing with an object.

Behavior	1:15	1:30	1:45	2:00	2:15	2:30	2:45	3:00	3:15	3:30	3:45	4:00
Looking Around												
Academic Engagement												
Motor Behavior												
Social Interaction w/ Teacher												
Playing with Object												

And here's where you'll actually complete the SDO ratings.

1 = occurrence
0 = non-occurrence

Systematic Direct Observation Form ID#1-CLIP-D-MTS-5-01

Behavior

Looking Around: The student is looking around and not engaged in any other activity.
Academic Engagement: The student is actively or passively participating in the classroom activity.
Motor Behavior: The student is engaged in repetitive, stereotyped body movements.
Social Interaction w/ Teacher: The student is interacting with the classroom teacher.
Playing with Object: The student is repetitively playing with an object.

Behavior	1:15	1:30	1:45	2:00	2:15	2:30	2:45	3:00	3:15	3:30	3:45	4:00
Looking Around												
Academic Engagement												
Motor Behavior												
Social Interaction w/ Teacher												
Playing with Object												

These interval times indicate the time at the end of the interval.

MTS: Mark if the behavior is occurring at the moment the interval ends.

PI: Mark if the behavior occurred anytime between the start and end of the interval.

WI: Mark if the behavior occurred for the entire duration of the interval (i.e., from start to end).

Looking Around: The student is looking around and not engaged in any other activity.
Academic Engagement: The student is actively or passively participating in the classroom activity.
Motor Behavior: The student is engaged in repetitive, stereotyped body movements.
Social Interaction w/ Teacher: The student is interacting with the classroom teacher.
Playing with Object: The student is repetitively playing with an object.

Behavior	1:15	1:30	1:45	2:00	2:15
Looking Around					
Academic Engagement					
Motor Behavior					
Social Interaction w/ Teacher					
Playing with Object					

Behavior	2:45	3:00	3:15	3:30	3:45
Looking Around					
Academic Engagement					
Motor Behavior					
Social Interaction w/ Teacher					
Playing with Object					

Behavior	5:15	5:30	5:45	6:00	6:15
Looking Around					
Academic Engagement					
Motor Behavior					
Social Interaction w/ Teacher					
Playing with Object					

Behavior	7:45	8:00	8:15	8:30	8:45
Looking Around					

What's the plan?

- ▶ Right now,
  - ▶ Orient to materials
  - ▶ Sign consent forms
- ▶ Then,
  - ▶ SDO overview
  - ▶ Method overview
  - ▶ Behavior overview
- ▶ After that,
  - ▶ Code training clips
- ▶ And finally,
  - ▶ Code experimental clips

Which behaviors will I rate?

Behavior	Definition
Academic Engagement	The student is actively or passively participating in the classroom activity.
Out of Seat	The student is out of his/her seat.
Looking Around	The student is looking around and not engaged in any other activity.
Motor Behavior	The student is engaged in repetitive, stereotyped body movements.
Playing with Object	The student is repetitively playing with an object.
Social Interaction w/ Child	The student is interacting with one or more other students.
Social Interaction w/ Teacher	The student is interacting with the classroom teacher.

Which behaviors will I rate?

- ▶ **Academic Engagement:**
The student is actively or passively participating in the classroom activity.

Examples: writing, raising hand, talking about a lesson, listening to the teacher, reading silently, or looking at instructional materials.

Nonexamples: staring into space, talking about off-topic activities.



Which behaviors will I rate?

- ▶ **Out of Seat:**
The student is out of his/her seat.

Examples: out of seat.

Nonexamples: in seat.



Which behaviors will I rate?

- ▶ **Looking Around:**
The student is looking around and not engaged in any other activity.

Examples: staring into space, looking at posters on walls, looking at ceiling.

Nonexamples: looking at teacher, writing on worksheet, talking to student about off-task topic, playing with pencil.



Which behaviors will I rate?

- ▶ **Motor Behavior:**
The student is engaged in repetitive, stereotyped body movements

Examples: rocking in chair, tapping desk with finger, bobbing head, rubbing fingers together, poking other student repeatedly with finger.

Nonexamples: tapping table with pencil, folding paper, throwing eraser, hitting self with magazine.



Which behaviors will I rate?

▶ Playing with Object:

The student is repetitively playing with an object.

Examples: tapping pencil, folding paper, playing with action figure, texting on phone.

Nonexamples: writing on worksheet with pencil, using calculator during math assignment.



Which behaviors will I rate?

▶ Social Interaction with Child:

The student is interacting with one or more other students.

Examples: talking to other student about weekend plans, listening to student as they speak to them, poking other student, throwing paper at other student.

Nonexamples: looking at student without engaging in conversation, staring into space, attending to teacher.



Which behaviors will I rate?

▶ Social Interaction with Teacher:

The student is interacting with the classroom teacher.

Examples: talking to teacher about assignment, listening to teacher as they are reprimanded, accepting pencil from teacher.

Nonexamples: looking at teacher when teacher is addressing whole class, looking away from teacher as they are spoken to.



What's the plan?

▶ **Right now,**

- ▶ Orient to materials
- ▶ Sign consent forms

▶ **Then,**

- ▶ SDO overview
- ▶ Method overview
- ▶ Behavior overview

▶ **After that,**

- ▶ Code training clips

▶ **And finally,**

- ▶ Code experimental clips

Training!

(remember to use 0's and 1's when coding)

Post-training

- ▶ Any questions about what we just did?
- ▶ We can't really talk about behaviors or problems once we start the real clips, and you shouldn't talk to each other about the clips. So, let's clarify anything right now before we go forward.
- ▶ Sheet with examples and non-examples on it to refer to.
- ▶ Remember, before each clip, just orient yourself to:
  - ▶ Which clip you're rating
  - ▶ What method you're using
  - ▶ What behaviors you're rating
- ▶ Do your best, don't stress, and take your time.

 Logistics

- ▶ **Breaks**
  - ▶ *Between clips*, take a breath, put last protocol away, grab next one, turn binder page
  - ▶ *Every six clips* (or after an hour of video coding), take a ten-minute break to stretch your legs
  - ▶ *Every eighteen clips*, lunch!
- ▶ **Viewing clips**
  - ▶ As long as you're doing your best, you're fine. Don't worry.
  - ▶ Don't rewind, don't start over (unless you clicked the wrong clip or missed the first few seconds and need to start it from the beginning)
  - ▶ Code the clip to the best of your ability, using the information you have.
  - ▶ Some clips will be harder than others, and that's okay.

 Right now

- ▶ Let's take a 10-minute break
- ▶ When we come back, we'll start coding the experimental clips, and then have lunch at a time we all agree on.

Break!

 What's the plan?

- ▶ **Right now,**
  - ▶ Orient to materials
  - ▶ Sign consent forms
- ▶ **Then,**
  - ▶ SDO overview
  - ▶ Method overview
  - ▶ Behavior overview
- ▶ **After that,**
  - ▶ Code training clips
- ▶ **And finally,**
  - ▶ Code experimental clips

Coding!

(remember to use 0's and 1's when coding)

Appendix I

Analysis Code in R

```
##### Prep data #####

data_raw.df <- read.csv("DataEntry_Cleaned.csv")
data_raw.df <- data_raw.df[,c(1:50)]

summary(data_raw.df)
data_raw.df[is.na(data_raw.df$Int_5_00) == TRUE,]
data_raw.df[is.na(data_raw.df$Int_8_45) == TRUE,]

data.df <- data_raw.df

data.df$ObsInts <- rowSums(data.df[,c(11:50)], na.rm = TRUE)
data.df$TotalInts <- rowSums(!is.na(data.df[,c(11:50)]))

data.df[data.df$TotalInts == 39,]

data.df$Prevalence <- data.df$ObsInts / data.df$TotalInts
data.df$Behavior <- as.factor(data.df$Behavior)
data.df$Rater <- as.factor(data.df$Rater)

##### Write an Excel and TXT file for SPSS #####

#library(xlsx)
#write.xlsx2(data.df, "data_from_R.xlsx", row.names = FALSE)
#library(foreign)
#write.foreign(data.df, "data_from_R.txt", "data_from_R.sps", package = "SPSS")

##### Write tab-delimited files for EduG #####

data_edug.df <- data.df
data_edug.df <- data_edug.df[,c(6,2,8,9,53)]
data_edug.df$Clip_Name <- as.numeric(data_edug.df$Clip_Name)
data_edug.df <- data_edug.df[order(data_edug.df$Clip_Name, data_edug.df$Rater),]

# MTS

write.table(data_edug.df[data_edug.df$Method == "MTS" & data_edug.df$Behavior == "1",5],
            "edugdata/mts1.txt", sep="\t", eol = "\n", row.names = F, col.names = F, append = F)

write.table(data_edug.df[data_edug.df$Method == "MTS" & data_edug.df$Behavior == "3",5],
            "edugdata/mts3.txt", sep="\t", eol = "\n", row.names = F, col.names = F, append = F)

write.table(data_edug.df[data_edug.df$Method == "MTS" & data_edug.df$Behavior == "5",5],
            "edugdata/mts5.txt", sep="\t", eol = "\n", row.names = F, col.names = F, append = F)

write.table(data_edug.df[data_edug.df$Method == "MTS" & data_edug.df$Behavior == "7",5],
            "edugdata/mts7.txt", sep="\t", eol = "\n", row.names = F, col.names = F, append = F)

# PI

write.table(data_edug.df[data_edug.df$Method == "PI" & data_edug.df$Behavior == "1",5],
```

```

"edugdata/pi1.txt", sep="\t", eol = "\t", row.names = F, col.names = F, append = F)

write.table(data_edug.df[data_edug.df$Method == "PI" & data_edug.df$Behavior == "3",5],
  "edugdata/pi3.txt", sep="\t", eol = "\t", row.names = F, col.names = F, append = F)

write.table(data_edug.df[data_edug.df$Method == "PI" & data_edug.df$Behavior == "5",5],
  "edugdata/pi5.txt", sep="\t", eol = "\t", row.names = F, col.names = F, append = F)

write.table(data_edug.df[data_edug.df$Method == "PI" & data_edug.df$Behavior == "7",5],
  "edugdata/pi7.txt", sep="\t", eol = "\t", row.names = F, col.names = F, append = F)

# WI

write.table(data_edug.df[data_edug.df$Method == "WI" & data_edug.df$Behavior == "1",5],
  "edugdata/wi1.txt", sep="\t", eol = "\t", row.names = F, col.names = F, append = F)

write.table(data_edug.df[data_edug.df$Method == "WI" & data_edug.df$Behavior == "3",5],
  "edugdata/wi3.txt", sep="\t", eol = "\t", row.names = F, col.names = F, append = F)

write.table(data_edug.df[data_edug.df$Method == "WI" & data_edug.df$Behavior == "5",5],
  "edugdata/wi5.txt", sep="\t", eol = "\t", row.names = F, col.names = F, append = F)

write.table(data_edug.df[data_edug.df$Method == "WI" & data_edug.df$Behavior == "7",5],
  "edugdata/wi7.txt", sep="\t", eol = "\t", row.names = F, col.names = F, append = F)

##### Preliminary looks #####

library(ggplot2)

ggplot(data = data.df, aes(x = Prevalence)) +
  geom_histogram(binwidth=.05, colour="black", fill="white") +
  facet_grid(Method ~ Behavior)

png(filename = "/Users/austinj/Dropbox/UConn - Dissertation/Dissertation_Manuscript/Figures/DensityPlot.png",
width = 900, height = 550)
ggplot(data = data.df, aes(x = Prevalence, fill = Clip_Name)) +
  geom_density(alpha=.3) +
  facet_grid(Method ~ Behavior) +
  theme_bw()
dev.off()

png(filename = "/Users/austinj/Dropbox/UConn -
Dissertation/Dissertation_Manuscript/Figures/AssignmentPlot.png", width = 700, height = 800)
ggplot(data = data.df, aes(xmin = Rater - .5, xmax = Rater + .5, ymin = Clip_Number - .5, ymax = Clip_Number +
.5)) +
  geom_rect(colour="black", alpha = 0.9, aes(fill=Method_Behavior)) +
  scale_x_discrete(breaks=c(1:10), name = "Rater") +
  scale_y_reverse(breaks=c(seq(1,72,6)), name = "Clip Order") +
  scale_fill_manual(values = c("#EFF3FF", "#BDD7E7", "#6BAED6", "#2171B5",
    "#EDF8E9", "#BAE4B3", "#74C476", "#238B45",
    "#FEE5D9", "#FCAE91", "#FB6A4A", "#CB181D"),
    name = "Method, Behavior",
    breaks = levels(data.df$Method_Behavior),
    labels = c("MTS, 1", "MTS, 3", "MTS, 5", "MTS, 7",

```

```

      "PI, 1", "PI, 3", "PI, 5", "PI, 7",
      "WI, 1", "WI, 3", "WI, 5", "WI, 7")) +
  coord_cartesian(xlim = c(0,11), ylim=c(-1, 74)) +
  theme_bw()
dev.off()

# Just changed aes to Clip_Name, needs to change scale fills and stuff too
ggplot(data = data.df, aes(xmin = Rater - .5, xmax = Rater + .5, ymin = Clip_Number - .5, ymax = Clip_Number +
.5)) +
  geom_rect(colour="black", alpha = 0.5, aes(fill=Clip_Name)) +
  scale_x_discrete(breaks=c(1:10), name = "Rater") +
  scale_y_reverse(breaks=c(seq(1,72,6)), name = "Clip Name") +
  coord_cartesian(xlim = c(0,11), ylim=c(-1, 74)) +
  theme_bw()

##### Descriptives #####

library(plyr)

desc.f <- function(x) {
  m <- format(round(mean(x$Prevalence, na.rm = TRUE), 2), nsmall = 2)
  sdev <- format(round(sd(x$Prevalence, na.rm = TRUE), 2), nsmall = 2)
  return(cbind(m, sdev))
}

descriptives_all.df <- ddply(data.df, c("Clip_Name", "Method", "Behavior"), desc.f)
descriptives_all.df$desc <- paste(descriptives_all.df$m, " (", descriptives_all.df$sdev, ")", sep = "")
descriptives_all.df <- descriptives_all.df[,c(1,2,3,6)]
descriptives_all.df <- reshape(descriptives_all.df, timevar = "Clip_Name", idvar = c("Method", "Behavior"),
  direction = "wide")

descriptives_meth.df <- ddply(data.df, c("Clip_Name", "Method"), desc.f)
descriptives_meth.df$desc <- paste(descriptives_meth.df$m, " (", descriptives_meth.df$sdev, ")", sep = "")
descriptives_meth.df <- descriptives_meth.df[,c(1,2,5)]
descriptives_meth.df <- reshape(descriptives_meth.df, timevar = "Clip_Name", idvar = "Method",
  direction = "wide")

descriptives_complete.df <- ddply(data.df, "Clip_Name", desc.f)
descriptives_complete.df$desc <- paste(descriptives_complete.df$m, " (", descriptives_complete.df$sdev, ")", sep =
  "")
descriptives_complete.df <- descriptives_complete.df[,c(1,4)]
descriptives_complete.df <- as.data.frame(t(descriptives_complete.df))

write.csv(descriptives_all.df, "Results/descriptives_all.csv", row.names = FALSE)
write.csv(descriptives_meth.df, "Results/descriptives_meth.csv", row.names = FALSE)
write.csv(descriptives_complete.df, "Results/descriptives_complete.csv", row.names = FALSE)

# Figure for MTS, PI, WI means and sd's by clip

library(ggplot2)
library(plyr)
descriptives_meth_ggplot.df <- ddply(data.df, c("Clip_Name", "Method"), desc.f)
descriptives_meth_ggplot.df$m <- as.numeric(as.character(descriptives_meth_ggplot.df$m))
descriptives_meth_ggplot.df$sdev <- as.numeric(as.character(descriptives_meth_ggplot.df$sdev))

```

```

descriptives_meth_ggplot.df$Method <- ordered(descriptives_meth_ggplot.df$Method, levels = c("WI", "MTS",
"PI"))

png(filename = "/Users/austinj/Dropbox/UConn -
Dissertation/Dissertation_Manuscript/Figures/Descriptives_MeanBYClip.png", width = 900, height = 550)
ggplot(descriptives_meth_ggplot.df, aes(x=Clip_Name, y=m, fill=Method)) +
  geom_bar(position=position_dodge(), stat="identity") +
  #geom_errorbar(aes(ymin=m-sdev, ymax=m+sdev),
    #width=.2,
    #position=position_dodge(.9),
    #colour = "black") +
  scale_y_continuous(breaks=seq(0, 1, 0.1)) +
  ylab("Mean") +
  xlab("Clip") +
  theme_bw()
dev.off()

# Wilcoxon Signed-Rank Test for non-parametric comparisons between methods

library(exactRankTests)
library(plyr)

wilcoxon_test.df <- wilcox.exact(data.df$Prevalence[data.df$Method == "MTS" & data.df$Clip_Name == "A"],
data.df$Prevalence[data.df$Method == "WI" & data.df$Clip_Name == "A"], paired = TRUE)

wilcoxon.f <- function(x) {
  results_WP <- wilcox.exact(x$Prevalence[x$Method == "WI"], x$Prevalence[x$Method == "PI"], paired =
TRUE)
  results_WM <- wilcox.exact(x$Prevalence[x$Method == "WI"], x$Prevalence[x$Method == "MTS"], paired =
TRUE)
  results_PM <- wilcox.exact(x$Prevalence[x$Method == "PI"], x$Prevalence[x$Method == "MTS"], paired =
TRUE)
  return(cbind(
    WPstat = results_WP$statistic, WPp = results_WP$p.value,
    WMstat = results_WM$statistic, WMp = results_WM$p.value,
    PMstat = results_PM$statistic, PMp = results_PM$p.value))
}

wilcoxon.df <- ddply(data.df, "Clip_Name", wilcoxon.f)

##### Mixed o x r x m x b design with m and b fixed #####
# Some code adapted from gCoef {QME} sample code

library(lme4)
library(plyr)

lmer_output_full <- lmer(data = data.df, formula =
  Prevalence ~ (1 | Clip_Name) + (1 | Rater) + (1 | Method) + (1 | Behavior) +
  (1 | Clip_Name : Rater) + (1 | Clip_Name : Method) + (1 | Clip_Name : Behavior) +
  (1 | Rater : Method) + (1 | Rater : Behavior) + (1 | Method : Behavior) +
  (1 | Clip_Name : Rater : Method) + (1 | Clip_Name : Rater : Behavior) +
  (1 | Clip_Name : Method : Behavior) + (1 | Rater : Method : Behavior),
  REML = TRUE)

lmer_full_varcomp <- ldply(VarCorr(lmer_output_full))
names(lmer_full_varcomp) <- c("Source", "Variance")

```

```

lmer_full_varcomp <- rbind(lmer_full_varcomp, data.frame("Source" = "Residual", "Variance" =
attr(VarCorr(lmer_output_full), "sc") ^ 2))

lmer_full_varcomp$Source <- ordered(lmer_full_varcomp$Source, levels = c(
  "Clip_Name", "Rater", "Method", "Behavior", "Clip_Name:Rater", "Clip_Name:Method",
  "Clip_Name:Behavior", "Rater:Method", "Rater:Behavior", "Method:Behavior",
  "Clip_Name:Rater:Method", "Clip_Name:Rater:Behavior", "Clip_Name:Method:Behavior",
  "Rater:Method:Behavior", "Residual"))
lmer_full_varcomp <- lmer_full_varcomp[order(lmer_full_varcomp$Source),]
lmer_full_varcomp$Percent <- round(lmer_full_varcomp$Variance / sum(lmer_full_varcomp$Variance) * 100, 2)
lmer_full_varcomp$Variance <- round(lmer_full_varcomp$Variance, 3)
lmer_full_varcomp

lmer_full_varcomp_fixed <- data.frame(
  "Source" = c("Clip_Name", "Rater", "Clip_Name:Rater"),
  "Variance" = rbind(
    lmer_full_varcomp[lmer_full_varcomp$Source == "Clip_Name",2]
    + (lmer_full_varcomp[lmer_full_varcomp$Source == "Clip_Name:Method",2] / length(levels(data.df$Method)))
    + (lmer_full_varcomp[lmer_full_varcomp$Source == "Clip_Name:Behavior",2] /
length(levels(data.df$Behavior)))
    ,
    lmer_full_varcomp[lmer_full_varcomp$Source == "Rater",2]
    + (lmer_full_varcomp[lmer_full_varcomp$Source == "Rater:Method",2] / length(levels(data.df$Method)))
    + (lmer_full_varcomp[lmer_full_varcomp$Source == "Rater:Behavior",2] / length(levels(data.df$Behavior)))
    ,
    lmer_full_varcomp[lmer_full_varcomp$Source == "Clip_Name:Rater",2]
    + (lmer_full_varcomp[lmer_full_varcomp$Source == "Clip_Name:Rater:Method",2] /
length(levels(data.df$Method)))
    + (lmer_full_varcomp[lmer_full_varcomp$Source == "Clip_Name:Rater:Behavior",2] /
length(levels(data.df$Behavior)))
    + (lmer_full_varcomp[lmer_full_varcomp$Source == "Residual",2] / length(levels(data.df$Method_Behavior)))
  )
)

lmer_full_varcomp_fixed$Variance <- round(lmer_full_varcomp_fixed$Variance, 3)
lmer_full_varcomp_fixed$Percent <- round(lmer_full_varcomp_fixed$Variance /
sum(lmer_full_varcomp_fixed$Variance) * 100, 2)

# Varcomps for fully random model

fullyrandom.df <- read.csv("data_for_varcompgraph.csv")

fullyrandom.df$VarComp <- ordered(fullyrandom.df$VarComp, levels = c(
  "o","r","m","b","or","om","ob","rm","rb","mb","orm","orb","omb","rmb","ormb,e"
))

png(filename = "/Users/austinj/Dropbox/UConn -
Dissertation/Dissertation_Manuscript/Figures/fullmodel_varcomp.png", width = 900, height = 550)
ggplot(fullyrandom.df, aes(x=VarComp, y=Percentage)) +
  geom_bar(position=position_dodge(), stat="identity") +
  scale_y_continuous(breaks=seq(0, 100, 5)) +
  ylab("Percentage of total variance") +
  xlab("Facet") +
  theme_bw()
dev.off()

```

```
##### Separate o x r designs within each fixed facet #####
```

```
library(plyr)
library(lme4)

gt.f <- function(x) {
  gt_lmer <- lmer(data = x, formula = Prevalence ~ (1 | Clip_Name) + (1 | Rater), REML = TRUE)
  gt_lmer_output <- ldply(VarCorr(gt_lmer))
  names(gt_lmer_output) <- c("Source", "Variance")
  gt_lmer_output <- rbind(gt_lmer_output, data.frame("Source" = "Residual", "Variance" = attr(VarCorr(gt_lmer),
"sc") ^ 2))
  gt_lmer_output$Source <- ordered(gt_lmer_output$Source, levels = c("Clip_Name", "Rater", "Residual"))
  gt_lmer_output <- gt_lmer_output[order(gt_lmer_output$Source),]
  gt_lmer_output$Percent <- round(gt_lmer_output$Variance / sum(gt_lmer_output$Variance) * 100, 2)
  return(gt_lmer_output)
}
```

```
gt_faceted <- ddply(data.df, c("Method", "Behavior"), gt.f)
gt_faceted$Source <- revalue(gt_faceted$Source, c("Clip_Name" = "o", "Rater" = "r", "Residual" = "or,e"))
```

```
library(ggplot2)
png(filename = "/Users/austinj/Dropbox/UConn -
Dissertation/Dissertation_Manuscript/Figures/or_varcomp_percent.png", width = 900, height = 550)
ggplot(data=gt_faceted, aes(x=Source, y=Percent, fill=Source)) +
  geom_bar(stat="identity") +
  geom_text(aes(label = sprintf("%.2f", Percent))) +
  facet_grid(Method ~ Behavior) +
  theme_bw()
dev.off()
```

```
write.csv(gt_faceted, file = "gt_faceted.csv", row.names = FALSE)
```

```
##### Create G and D coefficients #####
```

```
coeffs.f <- function(x) {
  g_rel <- x$Variance[x$Source == "o"] / (x$Variance[x$Source == "o"] + (x$Variance[x$Source == "or,e"]/10))
  d_abs <- x$Variance[x$Source == "o"] / (x$Variance[x$Source == "o"] + (x$Variance[x$Source == "r"]/10) +
(x$Variance[x$Source == "or,e"]/10))
  return(rbind(
    c("g_rel", g_rel),
    c("d_abs", d_abs)))
}
```

```
coeffs.df <- ddply(gt_faceted, c("Method", "Behavior"), coeffs.f)
names(coeffs.df) <- c("Method", "Behavior", "Coeff", "Value")
```

```
coeffs.df$Value <- as.numeric(as.character(coeffs.df$Value))
```

```
# Make coeffs nice for table
```

```
library(reshape2)
coeffs_pretty.df <- dcast(coeffs.df, Method + Behavior ~ Coeff, value.var = "Value")
coeffs_pretty.df <- coeffs_pretty.df[,c(1,2,4,3)]
```

```
write.csv(coeffs_pretty.df, file = "coeffs.csv", row.names = FALSE)
```

```
##### Conduct D Study #####
```

```
dstudy.f <- function(x,y) {
  g_rel <- x$Variance[x$Source == "o"] / (x$Variance[x$Source == "o"] + (x$Variance[x$Source == "or,e"]/y))
  d_abs <- x$Variance[x$Source == "o"] / (x$Variance[x$Source == "o"] + (x$Variance[x$Source == "r"]/y) +
(x$Variance[x$Source == "or,e"]/y))
  return(data.frame(
    "raternumber" = y,
    "g_rel" = g_rel,
    "d_abs" = d_abs))
}
```

```
dstudy.df <- ddply(gt_faceted, c("Method","Behavior"), dstudy.f, y = 1:10)
```

```
write.csv(dstudy.df, file = "dstudy.csv", row.names = FALSE)
```

```
# D study plots
```

```
library(ggplot2)
```

```
png(filename = "/Users/austinj/Dropbox/UConn -
Dissertation/Dissertation_Manuscript/Figures/DStudy_GCoeff.png", width = 900, height = 550)
ggplot(data=dstudy.df, aes(x=raternumber, y=g_rel)) + geom_line() + geom_point() +
  scale_x_continuous(breaks = 1:10) +
  ylim(.8, 1) +
  xlab("Number of Raters") + ylab("G coefficient (relative)") +
  facet_grid (Method ~ Behavior) +
  theme_bw()
dev.off()
```

```
png(filename = "/Users/austinj/Dropbox/UConn -
Dissertation/Dissertation_Manuscript/Figures/DStudy_DCoeff.png", width = 900, height = 550)
ggplot(data=dstudy.df, aes(x=raternumber, y=d_abs)) + geom_line() + geom_point() +
  scale_x_continuous(breaks = 1:10) +
  ylim(.8, 1) +
  xlab("Number of Raters") + ylab("D coefficient (absolute)") +
  facet_grid (Method ~ Behavior) +
  theme_bw()
dev.off()
```

```
##### IOA #####
```

```
### Make random data to test with
```

```
ioa_data.df <- read.csv("Diss_RandomIOA.csv")
rand_ioa_data.df <- as.data.frame(matrix(round(runif(8640,0,1),0), ncol = 40))
names(rand_ioa_data.df) <- names(ioa_data.df)[3:42]
ioa_data_test.df <- cbind(ioa_data.df[,1:2], rand_ioa_data.df)

ioa_data_test.df$ObsInts <- rowSums(ioa_data_test.df[,c(3:42)], na.rm = TRUE)
ioa_data_test.df$TotalInts <- rowSums(!is.na(ioa_data_test.df[,c(3:42)]))
ioa_data_test.df$Prevalence <- ioa_data_test.df$ObsInts / ioa_data_test.df$TotalInts
```

```
### Conduct IOA
```

```
# With real data, replaced "ioa_data_test.df" with correct object
# Prep data
```

```

ioadata1.df <- data.df
ioadata2.df <- read.csv("Diss_ RandomIOA_MCM.csv")

ioadata2.df <- ioadata2.df[,c(1:42)]
ioadata2.df$ObsInts <- rowSums(ioadata2.df[,c(3:42)], na.rm = TRUE)
ioadata2.df$TotalInts <- rowSums(!is.na(ioadata2.df[,c(3:42)]))
ioadata2.df$Prevalence <- ioadata2.df$ObsInts / ioadata2.df$TotalInts

ioadata1.df <- ioadata1.df[,c(2,5,11:53)]
names(ioadata1.df)[2] <- "Protocol"

ioadata1.df$DataType <- 1
ioadata2.df$DataType <- 2
ioadata1.df$UniqueID <- paste(ioadata1.df$Rater, ioadata1.df$Protocol, sep = "_")
ioadata2.df$UniqueID <- paste(ioadata2.df$Rater, ioadata2.df$Protocol, sep = "_")

ioadata1.df <- ioadata1.df[ioadata1.df$UniqueID %in% ioadata2.df$UniqueID,]

ioadata1.df <- ioadata1.df[,c(1:42,45,46)]
ioadata2.df <- ioadata2.df[,c(1:42,45,46)]

ioadata_merged.df <- rbind(ioadata1.df, ioadata2.df)

# Run function

ioatest.f <- function(x) {
  x[1,] == x[2,]
}

library(plyr)
comparison.ioa <- ddply(.data = ioadata_merged.df, .variables = c("Rater", "Protocol"), .fun = ioatest.f)
comparison.ioa_sums <- colSums(comparison.ioa, na.rm = TRUE)
comparison.ioa_byinterval <- sum(comparison.ioa_sums[3:42])/(nrow(comparison.ioa)*40)*100
comparison.ioa_sumspercentage <- as.data.frame(round((comparison.ioa_sums / nrow(comparison.ioa))*100, 2))
comparison.ioa_sumspercentage

#write.csv(ioadata_merged.df, "ioadata_merged.csv", row.names = FALSE)

```