EVALUATING TEACHER IMPLEMENTATION OF DISCONTINUOUS DATA COLLECTION IN THE CLASSROOM

A DISSERTATION

SUBMITTED TO THE FACULTY

OF

THE GRADUATE SCHOOL OF APPLIED AND PROFESSIONAL PSYCHOLOGY

OF

RUTGERS,

THE STATE UNIVERSITY OF NEW JERSEY

BY

SHAWNA R. UEYAMA

IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE

OF

DOCTOR OF PSYCHOLOGY

NEW BRUNSWICK, NEW JERSEY OCTOBER 2017

APPROVED: _____
Kate Fiske Massey, Ph.D.

_____
Robert H. LaRue, Ph.D.

DEAN: _____
Stanley B. Messer, Ph.D.

ABSTRACT

Discontinuous data collection procedures such as momentary time sampling (MTS) and partial interval recording (PIR) provide ABA practitioners with an alternative to tedious and oftentimes unfeasible continuous data collection. Discontinuous data is especially practical for classroom teachers who must collect behavioral data while also implementing instructional protocols. However, the existing literature on MTS and PIR come from simulated or controlled laboratory studies rather than applied settings. Furthermore, most studies focus on methodological error and do not consider human error in discontinuous data collection. The present study compared four discontinuous data collection procedures: 10-s MTS for 10 min, 30-s MTS for 30min, 10-s PIR for 10 min, and 30-s PIR for 30 min in a classroom setting using three teacher-student dyads. This study aimed to identify the procedure that had the least methodological and human error when used by teachers who were collecting duration data on stereotypy. Methodological error was measured by comparing teacher-collected estimates to duration data coded from video. Human error was quantified by calculating teachers' treatment integrity (TI) of an instructional protocol and their interobserver agreement (IOA) for each discontinuous data collection method. In addition, this study compared the social validity of these procedures by examining teacher perceptions and preference. With regards to methodological error, results indicated that 10-s PIR, and especially 30-s PIR, significantly overestimated the occurrence of stereotypy, while both 10-s and 30-s MTS yielded very accurate estimates. All three teachers, however, erroneously perceived PIR to be more accurate than MTS. Results for human error were less conclusive, but indicated that these teachers could multitask while maintaining high TI and IOA. Lastly, findings from the present study suggest that the factors that affect preference are complex and vary across individuals.

ACKNOWLEDGMENTS

To Kate Fiske Massey, thank you for your support in the development, implementation, and write-up of this study. I could not have asked for a better dissertation chair. You gave me the perfect balance of patience, enthusiasm, and encouragement to keep me on track during my very eventful year and a half. To Bob LaRue, thank you for your guidance, kindness, and of course, humor during my time at the DDDC and on this project.

Meredith Bamond and Rob Isenhower, you both have supported me in so many ways throughout this process. I truly appreciate that I could always count on you two. Rose Greenblatt, you did such a wonderful job with the coding. Thank you so much for all of the hard work you put into this study. To the students and aides at the Douglass Developmental Disabilities Center, your willingness to participate in this research made all of this possible. To the classroom staff, especially Jen and Caitlin, thank you for going out of your way to help with scheduling and for always making me feel welcome in your classrooms.

Lastly, to my family, I couldn't have gotten to this point without your support. Thank you for being there for me everyday.

TABLE OF CONTENTS

PAGE

LIST OF FIGURES

Introduction

A core feature of applied behavior analysis (ABA) is the systematic manipulation of independent variables, such as treatment interventions, to measure their effects on dependent variables, such as target behaviors (Peterson, Horner, & Wonderlich, 1982; Wheeler, Baggett, Fox, & Blevins, 2013). When implementing ABA methods with learners with autism, teachers are often expected to precisely and simultaneously implement the treatment plan and collect data on both adaptive and maladaptive behavior. High integrity of the treatment and accuracy of data collection are considered essential in ABA, to provide support for internal and external validity. However, little research has examined the impact of various data collection methods on the accuracy of the data collection system and the treatment integrity (TI) of a simultaneously implemented intervention.

Since human error is unavoidable, it is critical to understand ways in which these errors in teaching and in data collection can be minimized. Data collection literature has focused almost exclusively on methodological limitations (e.g., Ciotti Gardenier, MacDonald, & Green, 2004; Devine, Rapp, Testa, Henrickson & Schnerch, 2011; Green, McCoy, Burns, & Smith, 1982; Hanley, Camillerri, Tiger, & Ingvarsson 2007; Harrop & Daniels, 1986; Murphy & Goodall, 1980; Powell, Martindale, & Kulp, 1975; Powell, Martindale, Kulp, Martindale, & Bauman, 1977; Rapp , Colby-Dirksen, Michalski, Carroll, & Lindenberg, 2008; Wirth, Slaven, & Taylor, 2014) but has done little to understand the role and sources of human error in data collection. Although teachers' mistakes are likely unintentional, especially in demanding contexts such as classrooms, these errors in treatment and data collection have serious implications because they compromise the practitioner's ability to demonstrate functional relationships between independent variables (IVs) and dependent variables (DVs). The current study aims to identify

the discontinuous data collection method that produces the least amount of human and measurement error when implemented by a teacher working with a student with an autism spectrum disorder (ASD). In addition, it will evaluate the effects of discontinuous data collection on TI, to identify a method that not only produces accurate data with high reliability but also enables teachers to provide instruction with high integrity.

Treatment Integrity

Several types of TI are referenced in the literature. For example, TI can measure either a behavioral consultant's adherence to a specific consultation method or a teacher's accuracy of treatment implementation (Noell, 2007). For the purposes of this study, the latter form of TI will be used to indicate the degree to which each specific procedure embedded within an overall intervention is implemented correctly.

TI can be measured through observation, in which a rater records the occurrence or nonoccurrence of treatment components, or by permanent products, where naturally occurring products of the intervention are used as indicators of correct implementation. For example, graded worksheets or a daily log with parent-teacher correspondence are permanent products that could measure TI for teacher feedback and communication with parents respectively. Although permanent products are an efficient method for calculating TI, many procedures do not yield tangible artifacts (Fiske, 2008; Noell, 2007). Direct observation, while time consuming (Gresham, 1989), is often the only viable option for TI measurement. Alternative methods such as self-report questionnaires or interviews that ask teachers to indicate the degree to which they correctly implemented the treatment are not recommended as they are susceptible to biases and do not accurately reflect true levels of TI (Noell, 2007). In sum, direct observation, and when feasible, permanent products are appropriate methods for measuring TI.

TI scores can calculated in several ways. In general, TI represents the percent of opportunities in which the procedure was correctly implemented (Vollmer et al., 2008). In direct observation, one way to calculate TI is to divide the sum of correctly implemented components by the total number of opportunities to implement the components, and multiply the quotient by 100. However, because certain types of errors in TI can impact treatment outcomes differentially (Carroll, Kodak, & Fisher, 2013; Vollmer et al., 2008), researchers often recommend that TI be calculated for each of the individual components to allow for more targeted feedback to instructors (Fiske, 2008; Vollmer et al., 2008). Hagermoser Sanetti et al. (2008) specifically recommend a more conservative measure of TI in which TI is calculated separately for each component of the treatment package by indicating whether each component was implemented correctly on *all* opportunities. Overall TI would then be calculated by dividing the number of components implemented without any errors by the total number of components.

TI data from observations and permanent products allow researchers and practitioners to determine whether treatments are implemented as planned to minimize threats to internal and external validity of experiments (Gresham, 1989; Moncher & Prinz, 1991; Noel, 2007; Perepletchikova & Kazdin, 2005). Internal validity refers to the degree to which the IV can convincingly account for the change in the DV (Mayer, Sulzer-Azaroff, & Wallace, 2014). Studies with strong internal validity are able to assess outcomes to determine the treatment's effectiveness, while those with weak internal validity cannot draw such functional conclusions. For example, if a child's problem behavior decreases after a behavior plan is implemented, it is not possible to attribute these outcomes (DV) to the manipulations made in the intervention (IV) if there are no data supporting the accurate and consistent implementation of the treatment. Similarly, if significant behavior changes do not occur, it is not possible to ascertain whether a

treatment was ineffective or if an effective treatment was implemented with poor integrity (Gresham, 1989; Gresham, Gansle, & Noell, 1993; Peterson et al., 1982; Wilkinson, 2007).

The importance of confirming functional relationships between treatment and behavior outcome cannot be overstated in the field of behavior analysis. In fact, researchers and practitioners must make empirically sound, data-driven decisions to uphold their ethical obligations as set forth by the Professional and Ethical Compliance Code for Behavior Analysts (Behavior Analyst Certification Board, 2016). For instance, Guideline 1.01 states that behavior analysts must rely on scientific knowledge. Guideline 3.01b emphasizes the importance of graphing data to inform behavior-change programs. Without high TI, these requirements cannot be met, as functional relationships are based on assumptions that promote faulty decision-making (Peterson et al., 1982; Vollmer, Sloman, & Pipkin, 2008; Wheeler et al., 2006). Such misguided clinical decisions can have serious implications, as they often inform clinicians on how to approach serious issues including the use of more restrictive behavior management procedures such as punishment, adjustments to medications, and staffing requirements (Vollmer et al., 2008). An assessment that disregards TI data may lead to unnecessary harm to treatment recipients or may prevent them from receiving timely and appropriate care.

Beyond internal validity, measurement of TI is also necessary to ensure external validity, in which conclusions can be generalized to other people, behaviors, or settings (Mayer et al., 2014). In research contexts, treatments are often implemented with homogeneous samples in highly controlled environments. Once treatment efficacy is established with high TI in this setting, the operational definitions for the IV and detailed description of procedures that were used to measure TI can be utilized by researchers or practitioners to replicate the study. The replication allows researchers to evaluate the effectiveness of the methods under new conditions

that may include different instructors, learners, or settings than the original study (Gresham et al., 1993; Wheeler et al., 2006). This replication strengthens external validity of the treatment as it tests the generality of the intervention and may reveal variables that moderate the functional relationship between the treatment and outcome (Perepletchikova & Kazdin, 2005). By knowing more specific details about the generality of the intervention and the populations for whom it is effective, practitioners can select more appropriate treatments for their learners

Although the importance of TI has been recognized, it is often overlooked in research and in practice. This tendency is especially apparent when compared to the standards that are applied to operationalizing and measuring DVs such as behavioral outcome data (Noell et al., 2007; Wilkinson et al., 2007). This discrepancy is also reflected in the literature; for instance, in a review of articles published in the *Journal of Applied Behavior Analysis* (*JABA*) between 1968 and 1980, Peterson et al. (1982) found that only 20% of intervention studies included data on TI. Several more recent reviews of JABA and developmental disabilities journals have found that TI continues to be overlooked, with only 16 to 25% of sampled studies analyzing and reporting TI (Gresham et al., 1993; Gresham, MacMillan, Beebe-Frankenberger, & Bocian, 2000; Progar, Perrin, DiNovi, & Bruce, 2001; Wheeler et al., 2006).

Treatment integrity and treatment efficacy. TI measurement is critical in the practice of ABA not only for internal and external validity but also because TI level may have important implications on treatment efficacy. The body of research regarding the effects of high and low TI on treatment effectiveness, however, is limited and inconsistent.

In a study that systematically varied levels of TI (100%, 50%, 0%) in a three-step prompting procedure for two typically developing children, researchers found that child compliance decreased significantly when TI was reduced from 100% to 50% (Wilder, Atwell, &

Wine, 2006). These findings suggest that high TI was associated with more positive outcomes. Similarly, Carroll et al. (2013) identified the three most commonly erred upon instructional components in discrete trial instruction (DTI), the delivery of controlling prompts, instructions, and contingent rewards, and programmed them to occur during DTI sessions with learners with autism. When TI of these three components was low (33%), learners were unable to acquire skills or did so more slowly than when TI was high (100%), again indicating that TI is positively associated with treatment gains. However, other studies contradict these findings. For instance, Northup, Fisher, Kahng, Harrel, and Kurtz (1997), found outcome effects to be similar when treatment was implemented with 100% integrity and 50% integrity. Another study manipulated a time-out intervention's TI to 25%, 50%, 75%, and 100%, and found similar reductions in aggressive behavior when TI was implemented at 75% and 100% accuracy. At 25% and 50% TI, however, the intervention effects were less robust. These results indicate that although higher TI produces more favorable results, some small reductions to TI may not significantly affect behavior change (Rhymer, Evans-Hampton, McCurdy, & Watson, 2002).

Individual differences may moderate the response to variations in TI, where some students are more or less impacted by deviations from the treatment plan than others (Noell et al., 2007). For example, Vollmer and colleagues (1999) found that, overall, TI of a differential reinforcement of alternative behavior procedure was positively related to behavioral outcome. However, the authors found that some participants in the study showed similar levels of adaptive behavior regardless of the levels of TI. Another study evaluated the effects of constant time delay, in which the TI of the controlling prompt was manipulated in a skill acquisition program (Holcombe, Wolery, & Snyder, 1994). Results were somewhat similar to those of Vollmer et al. (1999) in that some participants showed similar performance regardless of TI. Of the six

participants, four showed acquisition with high and low TI procedures, one only mastered the task in the high TI condition, and one did not acquire the skill in either condition. Lastly, Noell, Gresham, and Gansle (2002) examined acquisition of computer-based arithmetic problems with 33%, 67%, and 100% TI for the prompting procedure. Although the authors found that students performed best with 100% TI, some students' performance was unaffected by significant reductions in TI. Taken together, these results suggest that TI reductions may not always lead to decreases in outcomes for individuals.

Further, some components of a treatment may be highly robust despite reductions to TI; or, high TI of the essential components of the plan may compensate for reductions in the TI of another component. As such, some components may be more important to implement with 100% integrity while others may not (Noell et al., 2007). For example, a component such as extinction, in theory, must be implemented with high integrity, as errors would result in intermittent reinforcement that would interfere with behavior reduction (Vollmer et al., 2008). In contrast, a continuous schedule of praise would be more likely to sustain its treatment effects with some errors in TI. Unfortunately, these ideas remain untested. Carroll et al. (2013) examined the effects of 100% and 67% TI for each of three components, the instruction, controlling prompt, and reinforcement, on skill acquisition in three students with ASD. The results suggest that perfect TI led to faster skill acquisition, but indicate that the effects of TI errors in specific components do not affect all children in the same way. Two participants showed poorest skill acquisition when controlling prompts were omitted and instructions were stated incorrectly, while the third student performed worst when reinforcement was erroneously omitted. With only one study in this area, however, further research is warranted to adequately evaluate the relationship between TI and outcomes.

In sum, a clear consensus regarding the effect of TI on direct outcomes has not yet been made. In the face of such ambiguity, however, behavior analysts must not overlook the importance of measuring TI. Regardless of its effect on outcomes, TI is a crucial tool in the interpretation of findings in research and in practice. High TI is necessary to determine whether a treatment is effective or ineffective, as it indicates that the treatment was implemented as intended. This information on treatment efficacy enables practitioners to more fully interpret findings, make informed decisions regarding the course of treatment, and systematically alter their treatments in a manner that is consistent with best practice.

Treatment integrity and treatment complexity. In addition to the incomplete understanding of effects of high and low TI on treatment outcomes, there are many unanswered questions regarding the factors that affect TI. Currently, the literature in this area is heavily based on consultation methods that seek to improve and maintain TI in classroom interventions. There is ample evidence that suggests that although direct instruction of staff results in accurate implementation of a treatment procedure, TI quickly declines upon training termination (Hagermoser Sanetti, Luiselli, and Handler, 2007; Jones, Wickstrom, & Friman, 1997; Mortenson & Witt, 1998; Noell, LaFleur, Mortenson, Ranier, & LeVelle, 2000; Noell, Witt, Gilbertson, Ranier, & Freeland, 1997; Witt, Noell, LaFleur, and Mortenson, 1997). Fortunately, performance feedback has been identified as an effective method for TI restoration and maintenance (Mortenson & Witt, 1998; Noell et al., 1997; Witt et al., 1997). In performance feedback, a consultant meets with a teacher and reviews objective data on teacher or student performance, provides praise or corrective feedback, and practices and discusses ways to improve TI (Noell, 2007). More specifically, teacher performance data, rather than student data, seems to be associated with greater improvements in TI (Codding et al., 2005; DiGennaro,

Martens, & Kleinmann, 2007) and the depiction of these data as graphs further enhances performance feedback effectiveness (Hagermoser Sanetti et al., 2007).

In contrast to the increasing understanding of consultation components, much less is known about other variables that may affect TI. Research has yet to investigate most environmental or individual factors that impact TI, but within the limited literature, it seems as though treatment acceptability is not necessary for high TI (Noell et al., 2005; Sterling Turner, 2002; Wickstrom, Jones, LaFleur, & Witt, 1998), while high levels of teacher experience (Noell et al., 2000) and poor student behavior (Hansen, Graham, Wolkenstein, & Rohrbach, 1991) are associated with low TI. However, much less is known about other variables that influence TI such as treatment complexity, staffing requirements, resource or time demands, and material requirements. Although these factors have been repeatedly discussed as potential variables that affect TI (Gresham, 1989; Perepletchikova & Kazdin, 2005; Yeaton & Sechrest, 1981) they have yet to be studied. The dearth of research is problematic because in less controlled classroom settings, these environmental factors may have a heightened impact on TI. Furthermore, practitioners must understand factors that most heavily influence TI in these naturalistic settings to best allocate training resources.

One environmental factor that could impact TI is treatment complexity. Treatment complexity is typically understood as the number of components within a specific treatment plan (Yeaton & Sechrest, 1981) and is thought to impact TI. However, treatment complexity can also be thought of as the number of distinct responsibilities the teacher has in addition to administering a treatment plan. Frequently, interventions require practitioners to conduct skill acquisition sessions and collect data for a specific behavior targeted for reduction. To date, little is known about the relationship between the demands of data collection procedures and the TI of

instructional procedures, as studies have not focused on this area. For instance, we do not know whether taking behavioral data while simultaneously implementing intervention may have adverse effects on the TI of the intervention. Interventions described in the research do not allow for such analyses, as one clinician typically implements the intervention while a second observes and collects data on the behaviors of interest (e.g. Codding et al., 2005; DiGennaro et al., 2005; DiGennaro et al., 2007; Noell et al., 2003). Although this preserves the integrity of the intervention and the accuracy of the data, in many school settings teachers often implement treatments and collect data simultaneously. Therefore, a clear understanding of the impact of data collection responsibilities on the TI of an intervention will have utility in advising teachers on how best to balance these responsibilities.

Although it is necessary to understand the effects of concurrent data collection and treatment implementation on TI, it is equally important to determine its impact on data accuracy. With both high TI and accurate data necessary for the determination of functional relationships, researchers and practitioners must consider methods that maximize both components. Thus, in addition to an understanding of how environmental factors such as treatment complexity affect TI and data collection, in-depth knowledge of the impact of various data collection systems on both data accuracy and TI is necessary.

Data Collection

Data collection via direct observation allows practitioners and researchers to quantify behaviors of interest and to monitor them over the course of treatment. These data can be obtained through continuous or discontinuous measurement systems but multiple factors must be considered to determine the most appropriate measurement system for any given behavior. A firm understanding of the system's properties is necessary to select a system that produces data

with minimal error. In doing so, practitioners can collect accurate data, which is a necessary component in the determination of functional control between a treatment and behavior outcomes.

Continuous measurement. In theory, to obtain the most accurate quantification of behavior, practitioners and researchers should use continuous data. Continuous measurement systems include continuous frequency recording (CFR), in which every occurrence of the target behavior during an observation session is tallied, and continuous duration recording (CDR), where the length, usually in seconds, of each bout of the target behavior is summed to yield a total duration. These figures are then divided by the total length of the observation session such that the frequency is expressed as occurrences per unit of time (e.g., seconds or minute) and the total duration is quantified either as a unit measurement (e.g., seconds or minutes) or as a percentage of the session (Fiske & Delmolino, 2012). When implemented accurately, these methods allow observers to measure the true value of the behavior. Such an accurate measurement of the DV allows for precise conclusions to be drawn about the effect of the treatment. In practice, however, these methods are not always feasible and practitioners and even researchers may opt to use a discontinuous measurement system (Fiske & Delmolino, 2012).

Discontinuous measurement. Teachers often juggle teaching with other activities such as data collection when working directly with a student. In these situations, continuous observation may not be practical and discontinuous data collection provides a much-needed alternative. In discontinuous measurement systems, such as time sampling, the observation sessions are divided into a series of equal intervals such as 10 s, the most common interval size used in published research (Fiske & Delmolino, 2012). Discontinuous measurement can be implemented with greater ease than continuous methods, as observers score each interval only

once by recording the occurrence or nonoccurrence of target behaviors during or at a specified moment within these intervals (Cooper et al., 2007).

In one time sampling method, partial interval recording (PIR), the observer records whether the target behavior occurred at any point within the interval. Therefore, intervals with a single occurrence or multiple instances of the behavior are scored identically. In whole interval recording (WIR), the observer records whether the target behavior occurred continuously for the entire duration of the interval. In momentary time sampling (MTS), the observer typically scores the presence or absence of the behavior in the last second of an interval (Powell et al., 1975). With all three methods, the number of intervals in which the behavior met the specified criteria is divided by the total number of intervals, and multiplied by 100 to calculate the percentage of intervals in which the target behavior was scored. Since these methods do not capture every instance or measure the duration of each bout, they are limited as they only provide an estimation of the true value of the behavior.

Despite this limitation, the ease with which discontinuous methods can be implemented makes them popular in both research and applied settings. To illustrate their widespread use, Kelly (1977) found that interval methods were used in over 40% of studies published in *JABA* between 1968 and 1975. More recently, Mudford, Taylor, and Martin (2009) examined *JABA* articles between 1995 and 2005 and found interval methods in 45% of the studies. No study has examined the relative popularity of MTS, PIR, and WIR in these journals, but in a related field of early childhood special education, Lane and Ledford (2014) reviewed four journals and found that 45% of the studies utilized interval systems to measure their dependent variables. Of these studies, 64% used PIR and only 13% and 11% used MTS and WIR, respectively.

Despite their widespread use and their relative ease of implementation, PIR, WIR, and MTS are only appropriate measurement systems if they provide accurate and sensitive estimates of behavior. *Accuracy* is defined as the "extent to which observed values . . . match the true values" of a behavior (Cooper et al., 2007, p. 689). Thus, an accurate measure will have a low absolute error, which is calculated as the difference between the value obtained from the discontinuous method and the value obtained by CDR or CFR.

In addition to the absolute values of behavior, applications of behavioral principles are most often concerned with changes in the DV, usually evidenced by an increase or decrease in behavior, to determine functional control and to evaluate treatment effects. It is therefore critical that measurement systems demonstrate *sensitivity* to changes in behavior. A sensitive measurement system is able to detect increases and decreases in behavior as well as identify when changes have not occurred, thereby avoiding false positive and false negative conclusions. A false positive occurs when a measurement system identifies a change in behavior that is not detected by a continuous measure. A false negative occurs when a measurement fails to detect a change that is found with a continuous measure (Rapp et al., 2008). As with poor TI of an intervention, these errors in data collection are problematic because they provide misleading information about the effects of treatments and environmental factors (IVs) on behaviors (DVs) (Rapp et al., 2008). As a result of these errors, decision-making abilities are impaired. For example, false positives can cause practitioners to continue using ineffective interventions and false negatives can lead to unnecessary discontinuation of effective treatments (Fiske & Delmolino, 2012).

Much of the research on error in discontinuous methods has focused on PIR and MTS, rather than WIR (Murphy & Goodall, 1980; Powell et al., 1975; Wirth et al., 2014). This

underrepresentation of WIR in the literature is likely because this method has fewer advantages to justify its use over continuous or other discontinuous methods. WIR lacks the accuracy and sensitivity to generate meaningful conclusions in data analysis. For example, WIR consistently underestimates duration events (Alvero, Struss, & Rappaport, 2009; Powell, Martindale, & Kulp, 1975; Powell et al., 1977; Wirth et al., 2014). It also produces greater magnitudes of error than MTS (Alvero et al., 2007; Harrop & Daniels, 1986; Wirth et al., 2014) and one study on the measurement of safe and unsafe posture found WIR to be more prone to human error than MTS (Taylor, Skourides, & Alvero, 2012). Furthermore, upon instances of the target behavior, WIR requires constant observation to determine if a behavior spans the entire duration of the interval. As a result, the response effort for WIR implementation can be similar to that of CDR, which produces much more accurate and sensitive data (Taylor et al., 2012). Therefore, the focus of this review will primarily concern the qualities of the remaining two discontinuous measurement systems, MTS and PIR.

Research findings can guide practitioners and researchers to maximize the accuracy and sensitivity of MTS and PIR to best enable practitioners to determine functional relations and make clinical decisions. Factors such as measurement properties (e.g., interval length, use of frequency or duration measurement), sampling error (e.g., duration of observation), and human error (e.g., demand intensity, social validity) may contribute to accuracy and sensitivity of a discontinuous measurement system. The large body of literature examining behavioral, measurement, and sampling error, which together comprise methodological error, has yielded conclusions about the relative strengths and weaknesses of discontinuous measurement methods. Significantly less research has been conducted on human error in discontinuous measurement, although it is an equally important component of measurement error, and is especially

concerning in applied settings in which teachers are unable to attend exclusively to data collection and may be required to engage in other tasks such as provide instruction with high fidelity.

Methodological error in data collection. To minimize error in data collection, it is important to know which method, PIR or MTS, is most appropriate for the given behavior that is to be quantified. However, the relative strengths of these methods vary across situations. When comparing the amount of error generated by discontinuous measurement methods, one must consider several factors, including whether the system is attempting to calculate the behavior's frequency or duration.

Accuracy of absolute frequency and duration estimates. Absolute measures of frequency and duration refer to a measure of one data point in time, rather than the evaluation of change over time. Discontinuous measures such as MTS and PIR are often used to measure these dimensions of behavior, but as a rule of thumb, should not be used to measure absolute frequency. Few studies have evaluated the accuracy of MTS and PIR frequency estimates because they quantify data in terms of the percentage of intervals in which the behavior occurred. In contrast, CFR measures behaviors by the number of occurrences or the number of occurrences per time unit. It is therefore challenging to equate quantitative data in the form of percent occurrence to rates and event counts (Powell & Rockinson, 1978). MTS and PIR are therefore contraindicated in the measurement of absolute frequency, which is more appropriately quantified using CFR.

Although MTS and PIR are inappropriate methods for estimating absolute frequency, they are suitable for estimating absolute duration. However, differences emerge in the accuracy of MTS and PIR estimates of duration when compared to CDR. Studies have consistently found

that MTS yields more accurate estimations of absolute duration than does PIR (Alvero et al., 2007; Harrop & Daniels, 1986; Murphy & Goodall, 1980; Powell et al., 1975; Powell et al., 1977). This finding has been replicated across studies ranging from those that measure simulated stimuli (Wirth et al., 2014), manipulated behavior such as planned sitting or hair twisting, (Powell et al., 1977; Green et al., 1982), to naturally occurring non-clinical behavior such as a secretary's sitting (Powell et al., 1975). Researchers have also replicated these findings using clinically relevant behavior by coding videos of stereotypy in participants with ASD (Ciotti Gardenier et al., 2004; Murphy & Goodall, 1980) and on-task behavior in children with emotional and behavioral disorders (Gunter et al., 2003). However, these studies did not code behavior in vivo, as it is often collected in naturalistic settings such as schools. Hanley et al. (2008) and Saudargas and Zanolli (1990) filled this gap in the literature when they found similar results when socially significant behavioral data were collected in vivo by trained observers. Only Delmolino, Fiske, and Dackis (2008) have extended the current literature to examine discontinuous data collection by teachers who were simultaneously engaged in instructional tasks. The teachers collected in vivo stereotypy data, which showed that once again, that MTS was more accurate than PIR when measuring duration events.

From simulation research to naturalistic studies of discontinuous data collection, results consistently indicate that PIR overestimates durations, and yields larger error margins, while MTS under- and overestimates durations but does so with smaller error margins (Ciotti Gardenier et al., 2004; Green et al., 1982; Harrop & Daniels, 1986; Murphy & Goodall, 1980; Powell et al., 1975; Powell et al., 1977; Rapp et al., 2008). Furthermore, in some cases, MTS can maintain accuracy with interval lengths much larger than 10 s. For instance, intervals as long as 120 s were shown to be within 5% of the estimates generated with 5-s MTS (Hanley et al.,

2007). Gunter et al. (2003) also demonstrated the accuracy of 2-min MTS intervals for measuring on-task behavior, as evidenced by the similar data paths created with a continuous measure. PIR has never demonstrated such flexibility in the literature. MTS' ability to accurately estimate durations across a wide range of interval lengths makes it a potentially optimal method for absolute duration measurement in classroom settings. In these settings, teachers may not have the ability to implement a demanding data collection procedure when preoccupied with the implementation of a treatment plan. However, in most cases, teachers are not so concerned with absolute estimates of behavior, but seek to determine whether a student's behavior is changing as a result of a treatment plan. For these purposes, it is necessary to understand the ways MTS and PIR vary in their ability to detect changes in behavior frequency or duration.

Sensitivity to frequency and duration changes. Given the tendency for PIR to overestimate and MTS to under and overestimate absolute durations, there has been some concern regarding the utility of these discontinuous methods in detecting changes in the duration or frequency of behaviors. Since applications of behavioral principles are often concerned with changes in the dependent variable, usually evidenced by an increase or decrease in behavior, it is essential that these measurement systems can determine functional control and evaluate treatment effects. Thus, a measurement system's ability to demonstrate *sensitivity* to changes in behavior is critical. Sensitivity to change will be affected by a variety of factors, such as interval length and observation length.

Interval length. To ensure adequate sensitivity, Cooper et al. (2007) recommend the use of MTS intervals less than 120 s. Other researchers have cautioned against MTS intervals over 60 s (Brittle & Repp, 1984) and 4 min (Gunter et al., 2003). Powell et al. (1975) recommend the use of PIR intervals under 80 s and Jacobsen (1982) suggested the use of an interval under 5 min.

The extensive research to date on the effect of interval duration has shown that such one-dimensional generalizations may not be appropriate, as interval length interacts with other factors to yield different degrees of error in discontinuous methods (Wirth et al., 2014). However, Saudargas and Zanolli (1990) correctly suggested that, "as the interval gets longer . . . similarity between the data recorded and the actual occurrence of the behavior probably decreases" (p. 123). In other words, when all else is held constant, shorter intervals yield data with less error than longer intervals. This relationship has been replicated extensively with both PIR and MTS (Devine et al., 2011; Hanley et al., 2007; Powell et al., 1975; Powell et al., 1977; Rapp et al., 2008). More specifically, Wirth et al. (2014) clarified that error decreases as the interval length approaches and becomes shorter than the duration of the behavior bout. To illustrate, if each instance of a child's stereotypy occurs for approximately 5 s, an interval length of 5 s or less would yield little error. Additionally, if each instance of another behavior lasted 60 s, a much longer interval of up to 60 s would also produce little error. It is notable, however, that even when using optimal interval lengths, MTS and PIR yield differing conclusions regarding frequency or duration change due to the inherent properties of their measurement system. It is therefore important for practitioners and researchers to know the strengths and weaknesses of each method and to understand the impact of interval sizes, session lengths, and predicted direction of behavior change on the utility of MTS and PIR.

MTS and PIR are often used to measure changes in frequency events within various experimental designs. Early studies on this emerging research area evaluated the sensitivity of MTS and PIR in an ABAB reversal design for the treatment of self-injurious behavior. Findings suggested that 10-s PIR and 10-s MTS both generated the same conclusions as those obtained by CFR. The authors did, however, note that PIR was slightly more sensitive to changes in

frequency (Meany-Daboul, Roscoe, Bourret, & Ahearn, 2007). Rapp et al. (2008) conducted a series of simulation studies in which the size of behavior change, cumulative duration and frequency of behavior were manipulated. Results indicated that 10-s PIR detected the most changes in frequency events, including 80% of the small changes, when using a 10-min observation sample. 20-s PIR, however, only detected approximately 30% of frequency changes. No MTS interval size could detect 80% or more of the changes in frequency. The evidence suggests that 10-s PIR is the most sensitive to frequency changes, but also demonstrates that PIR sensitivity to frequency quickly diminishes when the interval length increases.

MTS and PIR are also used to detect duration changes. Although 10-s PIR was found to be more sensitive to changes in frequency, 10-s MTS has been consistently more sensitive to changes in duration in reversal and multi-element designs (Meany-Daboul et al., 2007; Rapp, Colby, Vollmer, Roane, Lomas, & Britton, 2007). To illustrate, in a simulation study, Rapp et al. (2008) found that 10-s MTS intervals were able to detect 70% of small changes, and 80% of small, medium, and large duration changes when using 10-min observation sessions. Ten-second PIR was only able to detect about 50% of these changes (Rapp et al., 2008). Notably, in comparison to PIR which lost its sensitivity to frequency changes when interval length increased from 10-s to 20-s, MTS maintained some of its sensitivity to duration changes when using intervals up to 30s, where it detected most moderate to large duration changes (Rapp et al., 2008).

Observation length. The literature thus far has suggested that 10-s PIR and 10-s MTS are most sensitive to changes in frequency and duration respectively, when using 10-min observation sessions. In practice, however, 10-s intervals are not always feasible for instructors who must attend to many other ongoing duties. Fortunately, decreasing sampling error by lengthening

observation sessions could potentially offset the increased measurement error associated with lengthened intervals. Until recently, however, observation length has received very little attention, but findings suggest that longer observation periods are associated with less sampling error (Mudford, Beale, & Singh, 1990; Tiger et al., 2013). Despite these promising findings, most investigations of MTS and PIR sensitivity seem to have arbitrarily elected to use 10-min observation sessions (e.g., Meany-Daboul et al., 2007; Rapp et al., 2008). It is likely, however, that teachers would benefit from more options in discontinuous data collection, such as increased interval lengths, which may be less intrusive and therefore more preferable for a busy teacher.

To explore the possibility of lengthening intervals by increasing observation durations, these two factors were systematically manipulated to evaluate their effects on MTS and PIR sensitivity to detecting changes in frequency and duration (Devine et al., 2011). The study replicated Rapp et al.'s (2008) finding that 10-s MTS was most sensitive to changes in duration across session lengths. In addition, authors found that although 30-s MTS was unable to detect sufficient duration changes with 10-min observations, upon lengthening the observation session to 30 or 60 min, 30-s MTS detected over 80% of the changes identified by CDR. Consistent with Rapp et al. (2008), 10-s PIR was found to be the most sensitive measurement of frequency across session lengths, however 30-s MTS for 30- or 60-min sessions also demonstrated high sensitivity to frequency changes (Devine et al., 2011). In contrast, 20- and 30-s PIR intervals were unable to demonstrate sufficient sensitivity to frequency changes, even with longer sessions. Taken together, these findings indicate that with MTS, but not PIR, interval lengths can be extended without compromising sensitivity if session durations are lengthened. Furthermore, by extending the length of MTS intervals and sessions to 30-s and 30 or more minutes, respectively, it is possible to also detect frequency changes that were previously undetected with MTS. One

limitation to this study is that it only examined three observation lengths, with a maximum of 60 min, which may limit the generality of the conclusions to classroom settings where teachers may collect data throughout the school day.

Summary of recommendations. Upon considering the existing literature on the effects of interval length and observation duration on PIR and MTS accuracy and sensitivity, several recommendations can be made. Ten-second PIR provides the most accurate estimate of absolute frequency and 10-s MTS yields the most accurate estimate of duration. Since practitioners are typically more interested in detecting changes in behavior, the more relevant recommendations are those regarding sensitivity. The literature to date has demonstrated that 10-s PIR for 10 minutes or 30-s MTS for 30 minutes or more are most sensitive to changes in frequency (Rapp et al., 2008; Devine et al., 2011). For changes in duration, 10-s MTS for 10 min or 30-s MTS for 30 min are recommended (Devine et al., 2011; Rapp et al., 2007; Rapp et al., 2008). Currently, there are no data to suggest the superiority of one option over the other. Lastly, PIR is not recommended for use with duration events, as it consistently overestimates durations with large magnitudes of error.

These suggestions provide a basic guideline for the use of discontinuous measurement and provide a much-needed resource to assist practitioners to make empirically supported decisions when designing their discontinuous measurement system. Despite the extensive research that led to these guidelines, it must be emphasized that these recommendations are based on studies that did not observe socially significant behaviors in vivo, but instead coded simulated data or reanalyzed existing data sets (Devine et al., 2011; Rapp et al., 2008). Additionally, trained researchers rather than teachers conducted observations. Furthermore, these studies focused heavily on measurement error and sampling error, without accounting for human

error. The generality of these findings to applied settings such as classrooms, where teachers collect data, often on multiple behaviors, with few resources and several ongoing demands is unknown. In such settings, human error may be greater, and may render otherwise sensitive measurement systems unsuitable to such contexts.

No study to date has compared the amount of human error across various discontinuous measurement methods (PIR or MTS), interval lengths, and session durations. Such research would further inform practitioners who are deciding between methodologically equivalent methods (e.g., 10-s MTS for 10 minutes and 30-s MTS for 30 minutes to detect changes in duration events) and would shed light on the way different levels of human error arise across methods when used in applied settings. Moreover, it is unclear if data collection methods such as 10-s MTS for 10 min or 30-s MTS for 30 min would differ in their effect on a practitioner's ability to successfully achieve other responsibilities such as maintaining high TI. To address these gaps in the current knowledge base, research must begin to assess the human error in data collected by teachers.

Human error in data collection. Human error is introduced when a data collector incorrectly codes for the presence or absence of a behavior (Delmolino et al., 2008). Although it is often neglected in the literature, human error in data collection impacts the interpretation of functional relations in the same way as methodological error. Due to the limited number of studies that directly examine human error, however, our understanding of this type of error with MTS and PIR in applied contexts is incomplete. Such a limitation to current literature is problematic, because high human error can potentially yield data that are not accurate or sensitive and mislead a practitioner. In the few areas that have been researched, authors have found different levels of human error in PIR, MTS, and WIR and have investigated the effects of

demand intensity on human error. Additionally, research has investigated the relationship between social validity and human error.

To best determine the utility of PIR and MTS in a classroom setting, one must know if they produce different amounts of human error. Most studies to date, however, have been conducted in contrived settings. Taylor et al. (2012) compared human error in WIR and MTS when measuring different types of sitting posture, and found MTS to yield less human error. In another study, undergraduate psychology students coded MTS and PIR data from videos. The participants' MTS interobserver agreement (IOA) scores, based on mean kappa scores, were more accurate than their PIR IOA scores (Murphy & Harrop, 1994) but it is important to note that this study was conducted in a contrived laboratory environment. In a more naturalistic setting, Delmolino et al. (2008) compared teacher data, interval-by-interval, to those collected from video recordings of the session. This approach is favorable as it determines agreement and measures how closely the obtained data resemble the correct value that is presumed to have no human error. Teacher IOA was low overall but more so for MTS (72.1%) than PIR (78.5%). These results indicate that teachers may have more difficulty collecting accurate data than is typically observed in laboratory settings. Unfortunately, little research has been conducted to help us better understand this finding. Furthermore, since the two contrasting findings that have not yet been replicated, it is difficult to determine if MTS and PIR generate different amounts of human error.

Demand intensity. One factor that contributes to human error, demand intensity, is influenced by interval length. Although shorter interval lengths yield more methodologically accurate discontinuous data when observation length is held constant, these efforts to minimize methodological error may increase demands on the observer and cause more human error (Fiske

& Delmolino, 2012). For example, Hanley et al. (2007) demonstrated that human error was greater when two trained observers used shorter intervals to code on-task and in-zone behavior for multiple preschoolers. Interval-by-interval IOA for in-zone behavior with 30-s MTS intervals was 71.7%, while the same behavior at 60, 90, and 120-s had over 96% IOA. Although the observation procedures employed by the observers in this study differed from the standard MTS procedure, these results are troubling, as 30-s and even 10-s intervals are most often recommended in the existing literature. With high human error associated with short intervals, the accuracy and sensitivity of the measurement system may be compromised. Although there is no research on these specific interval lengths and durations, it is plausible that when measuring frequency events, the suggested 10-s PIR for 10 minutes may yield more human error than its methodologically equivalent 30-s MTS for 30 min. Similarly, for duration estimates, the recommended 10-s MTS for 10 min may yield more error than the other choice, 30-s MTS for 30 min. Further research on the effect of interval length on human error is warranted to understand the specifics of these recommendations.

Along with shorter intervals, two other factors that may increase demand intensity are higher levels of target behavior and co-occurring distractions. Mintz (2011) compared MTS data collected by college undergraduates who coded "nail biting" behavior from a video while varying behavior level and interval size. Results suggested that human error was not affected by level of behavior (20, 50, or 80% of session) or interval size (1, 5, 10, or 15 min). Human error rates were below 20% for all parameter combinations except for 10-min intervals with high rates of behavior, where error approached 30%. To address limits of the generality of these findings to classrooms and other naturalistic settings, the author also examined human error under distracted conditions, where the participants were asked to score math worksheets while collecting MTS

data. Findings indicated that under these conditions error was even lower (consistently under 11%) across all intervals and levels of problem behavior, suggesting that distractions did not affect accuracy of data collection. Additionally, unlike Hanley et al. (2007), this study did not show that shorter intervals led to greater human error. This inconsistency with Hanley et al.'s (2007) study cannot be fully interpreted due to differences in observation procedures and behavior rates. It is especially noteworthy that the interval lengths in this study were significantly longer than those used in most discontinuous data collection research, which rarely exceed 2 min. Furthermore, the findings are based on overall IOA, where total percent occurrence of behavior obtained by the observer and a standard rating are compared. This method is less rigorous than the recommended interval-by-interval IOA used in other studies (e.g., Hanley et al., 2007; Delmolino et al., 2008), and may explain the low overall human error and lack of variability in error across conditions.

In addition to interval size, behavior level, and simultaneous distractions, the number of target behaviors to code during each interval may influence human error. Currently, however, little research supports this hypothesis. Research findings on the relationship between number of target behaviors and human error have been mixed. When undergraduate students collected 15-s PIR data on three, six, and nine behaviors, their IOA decreased as the number of target behaviors increased (Dorsey, Nelson, and Hayes, 1986). It is notable, however, that the lowest IOA (kappa = .62) when measuring nine behaviors, was still above the threshold for adequate IOA (kappa > .60). In contrast, 10-s MTS and 10-s PIR IOA from another group of undergraduate observers did not vary as the number of target behaviors ranged from one to three. Although there was no relationship between target observation quantity and human error, the results demonstrated a clear pattern of MTS yielding higher kappa scores than PIR across all observation conditions;

MTS IOA consistently exceeded .70 while PIR IOA remained below .60 (Murphy & Harrop, 1994). The divergent findings for the relationship between the number of behaviors observed and human error may be explained by the different comparisons made in the respective studies. Perhaps a range of one to three behaviors was too small to produce variability in human error that was apparent when three, six, and nine behaviors were measured. Given the inconclusive findings regarding the effect of quantity of target behaviors on human error in MTS and PIR, further investigation in this area is warranted.

Social validity. A discontinuous measurement system's social validity, as evidenced by teacher preferences and perceptions, may also influence human error. Although it is unclear if factors related to social validity directly affect human error in data collection, TI literature suggests that this may not be the case, as acceptability had little effect on human error (Noell et al., 2005; Sterling Turner, 2002; Wickstrom et al., 1998). Data collection IOA may similarly be unaffected by perceptions of a measurement system's acceptability, however, these factors may affect error in other ways. Teachers may perceive certain data collection methods as more acceptable, accurate, enjoyable, and easy to implement, and may use those perceptions to guide their choice of data collection system. Since the preferred data collection systems are not necessarily those that yield the most accurate data, perceptions may lead teachers to use inappropriate methods that yield data with more methodological error. Although much more research is needed to fully understand the extent to which this problem exists and its potential remedies, a few studies have begun to investigate staff perceptions and preferences for various data collection procedures and parameters.

Staff preference for various MTS interval lengths (60-s, 90-s, 120-s) was examined with a procedure similar to a multiple stimulus without replacement preference assessment (MSWO)

(DeLeon and Iwata, 1996; Hanley et al., 2007). Observers chose one of three data sheets, each corresponding to the interval length on the first day and were asked to choose from the two remaining sheets on the following day. Both observers demonstrated a preference for 90-s MTS by selecting it first. On the following day, one observer selected the sheet for 60-s MTS and the other chose 120-s. The results from this preference assessment also matched the observers' verbal reports that they preferred the 90-s MTS, as it was the most "comfortable." Observers further elaborated that the 60-s interval required more vigilance while the 120-s interval was "boring," likely because of the prolonged waiting period during data collection.

More recently, Kolt and Rapp (2014) used a concurrent chains design to evaluate therapists' relative preferences for 10-s and 60-s MTS and PIR for measuring duration events. Therapists were presented with four colored data collection sheets, with each color corresponding to the data collection procedure and were asked to pick the one they wanted to use. Twelve trials were conducted for each of the eight therapists who then subsequently used the selected method to collect data from a prerecorded 6-min video in which an actor engaged in the target behavior. Six of eight therapists chose 60-s MTS over the other three methods. In both Hanley et al. (2007) and Kolt and Rapp's (2014) studies, the observers were not engaged in any other demands while coding, so it is possible that their preferences for interval duration may not be representative of those held by practitioners who must concurrently teach, manage problem behavior, and collect data.

In addition to methods that evaluate the choice of practitioners in data collection methods, self-report questionnaires about perceptions of the methods provide some rationale for the choices practitioners make. For example, questionnaires provided to teachers who conducted in vivo MTS and PIR data collection on stereotypy using interval lengths ranging from 10-s to

60-s found differences in their perceptions of these systems (Fiske & Delmolino, 2008). Overall, on a scale of 1 to 5, teachers rated MTS as easier to implement (M = 3.7) than PIR (M = 3.3), likely because MTS requires only one second of surveillance at the last second of each interval, while PIR requires ongoing observation of students' behavior during each interval until the behavior occurs. Despite different ratings, the study did not examine if these perceptions were associated with preferences in any way. Interestingly, although MTS was rated as easier to implement, some teachers described more frustration with this system, as they were concerned that the behavior was not being captured during the last second of the interval. Additionally, teachers rated PIR as more accurate (M = 4.0) in capturing student behavior than MTS (M = 3.1). Although this difference only approached significance (p =. 06), eight out of nine teachers rated PIR as more accurate. The authors suggest that this perceived inaccuracy of MTS and the associated frustration may explain the observed biases within instances of human error; 70% of the MTS IOA errors were caused by teacher's scoring an occurrence of stereotypy when it had not occurred during that second. Skepticism towards MTS and the unsatisfactory feeling associated with "missing" a behavior may contribute to the perception that MTS is less acceptable and make teachers less willing to use MTS in the classroom (Sterling-Turner et al., 2002). This possibility is especially problematic because even with a lower average IOA, MTS data produced much more accurate data than PIR in this study.

Similarly, in Kolt and Rapp's (2014) study, a follow-up survey was administered to therapists who collected data using 10-s and 60-s MTS and PIR. This measure also indicated that data collectors perceived MTS to be easier to implement than PIR. More specifically, observers reported that 60-s MTS was easiest to implement, as it was less stressful than 10-s MTS. Ten-second PIR was reported as the least preferred method, and observers described it as requiring

"too much attention or effort." However, in line with Fiske and Delmolino's (2008) finding that teachers rated PIR as more accurate than MTS, 10-s PIR was perceived to yield the most accurate data.

Based on previous research findings, observers and teachers have preferences and perceptions about interval lengths and measurement systems. Continued research is needed to better understand how the perceptions that MTS is more frustrating and inaccurate or reports that PIR is more difficult to implement, influence human error and willingness to use such methods. Furthermore, additional issues such as observer drift and observer biases (Kazdin, 1977) may also be of relevance in these contexts. With less controlled environments and few IOA checks, teachers may be less inclined to maintain precise data collection.

Rationale for the Present Study

TI and data collection are the two main factors that allow practitioners and researchers to identify functional relationships between treatments and outcomes. To make these interpretations and to use them to inform clinical decisions, however, there must be evidence that the treatment is implemented as planned and the data collected is accurate. Thus, high TI and high IOA with a methodologically sound data collection system are both necessary.

This goal to maximize the precision of treatment implementation and data collection is challenging due to the limitations in the current literature base. Little is known about contextual factors that impact TI such as treatment complexity and time demands. Similarly, the data collection literature is imbalanced and has significant gaps; it has thoroughly described sources of methodological error, but has overlooked an equally important component, human error. What is known about data collection, however, suggests that the optimal methods for discontinuous data collection are demanding, as they commonly recommend the use of 10- or 30-s intervals. In

the absence of conclusive evidence regarding the effect of data collection complexity on IOA, it is reasonable to hypothesize that more demanding PIR or MTS procedures that aim to minimize methodological error may result in greater human error as evidenced by lower IOA. In addition, we can predict that complex data collection procedures may also influence TI, and conversely, that concurrent implementation of treatment plans may reduce IOA. To best understand how these two key factors interact, as they likely do in practice, research must begin to study TI and IOA from teachers who administer treatments and collect data concurrently.

The current study seeks to replicate Delmolino et al.'s (2008) study on the accuracy and agreement of stereotypy data collected by teachers who are simultaneously teaching learners with ASD. To extend this research, however, the optimal intervals and observation lengths for duration events such as stereotypy, as suggested by existing literature, will be used. Both 10-s MTS for 10-min observations (Devine et al., 2011; Rapp et al., 2008) and 30-s MTS for 30-min (Devine et al., 2011) have been identified as appropriate measurement methods for collecting samples representative of duration data. Data collected with these methods will be compared to 10-s PIR for 10-min and 30-s PIR for 30-min observations, which are less sensitive and therefore not recommended for use with duration events (Rapp et al., 2008). In addition, teacher perceptions and preference for data collection procedures will be assessed. Lastly, this research aims to extend the literature by investigating the impact of these data collection procedures on the TI of classroom teaching procedures.

Hypotheses

Teacher IOA. Teachers' average reliability (IOA), as compared to discontinuous data collected via video, will be highest for 30-s PIR for 30 min, followed by 30-s MTS for 30 min, and 10-s PIR for 10 min. IOA will be lowest for 10-s MTS for 10 min. Data collected with all

four methods is expected to fall below the accepted 80%.

Teacher measurement error. The average measurement error of teachers' data, as compared to relative duration, will increase from lowest to highest in the following order: 30-s MTS for 30 min, 10-s MTS for 10 min, 10-s PIR for 10 min, and for 30-s PIR for 30 min.

Teacher treatment integrity. Discontinuous data collection will reduce teacher's TI in teaching relative to teaching only sessions during which teachers are not required to collect discontinuous data. Furthermore, 30-s MTS will be associated with the smallest decrease in TI relative to baseline, while 10-s PIR will be associated with the largest decrease in TI.

Teacher ratings. On average, teachers are expected to rate 30-s MTS for 30 min as easiest to implement, followed by 30-s PIR for 30 min, and 10-s MTS for 10 min. They will rate 10-s PIR for 10 min as the most difficult to implement. Teachers will report that 10-s PIR for 10 min yields the most accurate data, followed by 30-s PIR for 30 min. Ten-second MTS for 10 min will be rated third most accurate and 30-s MTS for 30 min will be rated as the least accurate measurement of stereotypy, as measured by mean scores. Both types of questionnaires, the Teacher Perceptions Form (Appendix D) and the Teacher Perceptions Ranking Form (Appendix E) will support these findings. Additionally, the Teacher Perceptions Ranking Form will indicate that teachers rank 30-s MTS for 30 min as most preferred, 30-s PIR for 30 min as second most preferred, followed by 10-s MTS for 10 min. Teachers will rate 10-s PIR for 10 min as least preferred.

Teacher preferences. Teachers' preferences, as indicated by the percent of opportunities in which they selected a data collection method, are expected to match those of their self-report. Teachers are expected to base their preferences on perceived ease. The procedures will be most to least preferred in the following order: 30-s MTS for 30 min, 30-s PIR for 30 min, 10-s MTS

for 10 min, and 10-s PIR for 10 min.

Method

Participants

Participants were three student-teacher dyads. The students carried a diagnosis of ASD and attended the Douglass Developmental Disabilities Center (DDDC). All three students engaged in one or more forms of vocal or motor stereotypy that matched the definitions provided by Ciotti Gardenier et al. (2004; Appendix A) and demonstrated the ability to receptively identify or match stimuli from an array. Student A was a male aged 16 years and 6 months, Student B was a male aged 14 years and 9 months, and Student C was a female aged 14 years and 7 months. All teacher participants were DDDC instructional staff who received job training in skill acquisition programming and data collection. All teacher participants had prior experience running skill acquisition programs and collecting behavioral data for their students. Teacher A was a 26-year old female who had 1.5 years of experience working with individuals with autism. Teacher B was a 24 year-old female who had 4 years of experience. Teacher C was a 25 year-old female with 1.5 years of experience.

Setting

The DDDC is a day program for individuals with ASD and provides 25+ hours of ABA intervention per week for each enrolled student. All sessions for each student took place in the student's classroom or another common work-area (e.g. life skills room) at the DDDC during regular school hours. Sessions occurred in a one-on-one instructional format, using stimulus materials that matched the student's current skill acquisition or maintenance programs. During sessions, other students and teachers were present in the room and engaged in classroom instruction.

Procedure

This study (1) evaluated discontinuous data collected by teachers of students with ASD to identify the procedure that produced the least amount of human error and measurement error, (2) measured teacher perceptions and preferences for the different discontinuous data collection procedures, and (3) examined the effects of simultaneous data collection on TI to determine the method of data collection that yielded the least amount of human error in TI.

Training. Prior to the first session, teachers were given a brief written description of MTS and PIR procedures and a description of the essential components of teaching (Carroll et al., 2013, Appendix B). Teachers also received a written operational definition for the student's specific stereotypy that was created in collaboration with the teacher. The PI reviewed these procedures with the teacher and answered their questions. At the end of the training session, teachers were given a brief questionnaire (Appendix C) that asked staff to use a 5-point Likert scale to rate how important each of the following factors were when selecting a data collection procedure: ease of implementation, accuracy of method, and length of procedure (i.e., observation duration).

Teacher data collection. Sessions were run up to three times per week, depending on the students' schedules and classroom staff availability. In each session, one to three conditions were run such that total session duration never exceeded 60 min per day. Each condition was separated by a brief 1- to 5-min break in which students A and B used their iPad and student C helped clean up session materials.

At the start of each session, the PI randomly selected the discontinuous data collection procedure to be used (10-s MTS for 10-min, 30-s MTS for 30-min, 10-s PIR for 10-min, 30-s PIR for 30-min, or no discontinuous data collection). In the four discontinuous data collection

conditions, teachers were required to use the discontinuous data collection procedure while implementing receptive identification or matching tasks from the student's current skill acquisition or maintenance programming. In the teaching only condition, where no discontinuous data collection was required, teachers only implemented skill acquisition or maintenance programs. If no such programs existed, a program and the accompanying materials were developed in collaboration with the classroom staff.

Before each discontinuous data collection session, a data sheet and a timer (GYMBOSS Interval Timer & Stop Watch) were attached to a clipboard. A small video camera was attached to the teacher's clipboard to videotape the teacher's data collection. A handheld video camera was also used to record the teacher and student, ensuring that all instances of stereotypy were captured. At the beginning of each session, the PI started the cameras and said, "Start" as the timer began. The teachers immediately began to implement the teaching plan and collect data. Teachers were encouraged to use the same program for as much of the session as possible but if necessary, were allowed to use other programs after the first ten teaching trials had been completed. During sessions, teachers provided breaks in a manner consistent with their student's existing classroom behavioral plan. Teachers collected discontinuous data on stereotypy during this time.

During these sessions, teachers used data sheets that included 60 intervals, the exact number needed for each observation with each method. For PIR data collection, the timer was set for 10 or 30 s. The timer sounded at the end of the 10 s or 30 s and the teacher recorded whether stereotypy occurred at any point during the interval by marking a plus (+) or minus (–) on the behavioral data sheet in the corresponding interval. Timers reset automatically at the end of each interval. The teacher repeated the data collection procedure until all intervals on the data sheet

were marked.

For MTS data collection, the timer was set for 10 or 30 s. The timer sounded for 1 s at the end of the interval and the teacher recorded whether stereotypy occurred when the timer sounded by marking a plus (+) or minus (–) on the data sheet in the corresponding interval. The timer reset automatically and teachers repeated the data collection procedure until all intervals on the data sheet were marked.

In addition, for all discontinuous data collection sessions, another data sheet was provided for teachers to collect trial-by-trial data for the student's correct and incorrect responses on the skill acquisition program. If prompts were used, teachers specified the type of prompt (physical, partial physical, model, gesture, or verbal).

The same procedures were used for the teaching only condition with the exception that discontinuous data were not collected during these sessions and the session duration lasted ten teaching trials, rather than 10 or 30 min. Since no discontinuous data were collected, there was no need to use a clipboard, timer, and camera for these data collection purposes. The PI still used a handheld video camera to record the teacher and student and provided the data sheet for skill acquisition data. When the PI said, "Start" teachers immediately began to implement the receptive identification or matching program and began to collect trial-by-trial data on correct and incorrect student responses, noting the prompt levels used when applicable. As in the discontinuous data collection sessions, teachers were encouraged to use the same program throughout the session and gave breaks to students in a manner consistent with their existing behavior plan. After completing ten trials of the program, the teacher informed the PI that the session was over.

Teacher perceptions. At the end of each discontinuous data collection session, teachers

filled out a two-item Teacher Perceptions Form (Appendix D). Ratings of perceived ease of data collection, accuracy of data collection method, and accuracy of teaching procedure when using 10-s MTS for 10 min, 30-s MTS for 30 min, 10-s PIR for 10 min, or 30-s PIR for 30 min were obtained using a 5-point Likert scale. This form was adapted from Fiske and Delmolino's (2008) "Social Validity Form."

After all sessions were completed, teachers were given the Teacher Perceptions Ranking Form (Appendix E) which asked teachers to make direct comparisons between the four methods by ranking them from 1 to 4 on preference, perceived ease, and perceived accuracy. Teachers were asked to briefly describe their reason for each ranking.

Preference assessment. The MSWO method used by Hanley et al. (2008) to determine teacher preference for data collection methods was used in the current study. Following all data collection sessions, the teachers were asked to take data for another series. Teachers were presented with four data sheets, each clearly labeled with either "10-s MTS for 10 min", "10-s PIR for 10 min", "30-s MTS for 30 min." or "30-s PIR for 30 min." The teacher was prompted to "Pick the one you want to use for this session." The teacher conducted a session using that data collection method and that sheet was marked with a "1." The next day, the remaining three data sheets were presented to the teacher in the same way. The teacher was prompted to choose a data collection procedure and they conducted a session using that method. A "2" was marked on that sheet. The same procedure was repeated on the third day to determine the third most preferred procedure. Teachers were not required to conduct a session with the final data sheet, and it was assumed that it was the least preferred method of data collection. A "4" was marked on the data sheet for this method.

Dependent Variables and IOA

One dependent variable for this study was the *IOA for teacher-collected discontinuous data*. The PI coded interval-by-interval data on stereotypy from all videos of the discontinuous data collection sessions. The PI used the same data collection procedures used by the teacher. The sound of the timer in the video signaled the end of each interval. These data were compared to the teacher-collected data, by reviewing the data sheets as well as the video of the data sheet to further ensure that the teacher marked the correct intervals and to determine whether any intervals were skipped. When teachers missed an interval and did not self-correct, the PI adjusted the teacher's subsequent data entries so that they corresponded to the correct interval on the data sheet. Such adjustments were made by shifting all subsequent "+" or "-" markings to the next interval. This was the only type of adjustment required in the present study, as there were no instances in which teachers double-coded intervals or missed multiple intervals in a row. Adjustments were made three times for Teacher A, never for Teacher B, and once for Teacher C.

Using the adjusted data sheets, exact IOA was calculated for each session by dividing the number of intervals in which the teacher and PI had agreements by the total number of intervals (60). This figure was multiplied by 100 to yield a percentage.

For 30% of all videotaped sessions, an undergraduate researcher used the videos to code interval-by-interval stereotypy data using the same procedure as described above. Exact IOA was calculated using data from the PI and researcher to determine agreement. For Teacher A's videos, average IOA was 88.9% (range 81.7 to 100%), for Teacher B's videos, average IOA was 89.6% (range 80.0 to 98.3%), and for Teacher C's videos, average IOA was 90.83% (range 85.0 to 98.3%).

To calculate *measurement error*, the PI coded continuous data of stereotypy from videos

of all discontinuous data collection sessions, using the procedure described by Ciotti Gardenier et al. (2004). One-second intervals were used to code for the presence or absence of stereotypy using the same operational definitions as those used by teachers. The number of 1-s intervals during which stereotypy occurred were divided by the total number of intervals in the observation (600 for 10-min sessions and 1800 for 30-min sessions) to yield a measure of relative duration. Relative duration was reported as the percent of seconds during which stereotypy occurred. The relative duration of stereotypy obtained by the PI using permanent product was compared to the percent of intervals obtained by teachers using discontinuous measurement methods. Measurement error was calculated as the absolute difference between the relative duration (from continuous data) and the percent of intervals from the discontinuous data (MTS or PIR). For example, if continuous data from permanent product determined relative duration of stereotypy as 50% of the session and the teacher's MTS or PIR data indicated that stereotypy occurred during 60% of intervals, the absolute measurement error for that method was equal 10 percentage points.

For 30% of all videotaped sessions, a research assistant recorded data on the relative duration of stereotypy using the same procedure as described above. Data from both observers were compared to calculate exact IOA for relative duration. For Teacher A's videos, average IOA was 90.9% (range 81.2 to 97.1%), for Teacher B's videos, average IOA was 94.2 (range 86.7 to 99.5%), and for Teacher C's videos, average IOA was 93% (range 87.8 to 96.2).

To calculate *treatment integrity* of teacher's instructional sessions, the PI watched the videos to code the first ten trials of all discontinuous data collection sessions and teaching only sessions. For each of the ten trials, the PI marked whether the teacher correctly or incorrectly implemented each of the applicable components including, (1) establishing ready behavior, (2)

securing attention, (3) providing a clear instruction, (4) presenting the instruction once, (5) giving praise contingent on a correct response, (6) using a controlling prompt, and (7) responding to problem behavior by not attending, removing demands, and/or blocking problem behavior (Carroll et al., 2013; Appendix B). TI was calculated by dividing the sum of correctly implemented components by the total number of opportunities to implement the components, and multiplying the quotient by 100.

For 30% of sessions, a research assistant independently scored the teacher's TI using the above method. To obtain exact IOA, comparisons were made between the PI and research assistant for each component for each trial; the number of components where both raters agreed were divided by the total number of components (70). An agreement was scored if both raters recorded correct implementation, incorrect implementation, or marked "not applicable," for the same component. For Teacher A's videos average IOA was 98.1% (range 95.7 to 100%), for Teacher B's videos, average IOA was 97.1% (range 91.4 to100%), and for Teacher C's videos, average IOA was 98.8% (range 95.7 to 100%).

Design

A multielement design was used to compare exact IOA for teacher-collected discontinuous data and to evaluate TI across conditions. Ten-second MTS for 10 min, 10-s PIR for 10 min, 30-s MTS for 30 min, 30-s PIR for 30 min, and teaching only sessions were randomized across conditions.

For exact IOA, teacher IOA from the four discontinuous data collection conditions were graphed and compared. Data from the teaching only condition was not included in this analysis. For teacher TI, data from all five conditions (four discontinuous data collection conditions and one teaching only condition) were graphed to determine the effects of the four discontinuous data

collection procedures on TI.

Results

Teacher IOA

As shown in Figure 1, teacher IOA was high and exceeded 80% for 53 of the 54 sessions. Teacher A's data indicated that 30-s MTS yielded the highest IOA while 10-s PIR yielded the lowest. Similarly, Teacher B's IOA was lowest in the 10-s PIR condition and about equal across the other three data collection methods. Teacher C's IOA data indicate that the percent agreement for MTS was higher than that of PIR, where 30-s and 10-s PIR yielded the lowest and second lowest IOAs respectively. Although no consistent or clear differentiation of data was observed in the graphs, 10-s PIR resulted in the lowest IOA in two of the three participants and 30-s MTS yielded or tied for the highest IOA for all three participants.

Measurement Error

Across participants, 10-s MTS and 30-s MTS produced less measurement error than PIR methods (Figure 2). For Teachers A and B, 30-s MTS produced the most accurate estimate of stereotypy duration (average absolute error = 3.41 and 2.56, respectively). For Teacher C, 10-s MTS produced the most accurate estimate (average absolute error = 1.87). For all three teachers, the differences between the average measurement error from 10-s and 30-s MTS were small, with the largest difference at 4 percentage points. Measurement error was much larger with 10-s PIR (average absolute error range = 18.15 to 30.13 percentage points). Measurement error was greatest and most variable with 30-s PIR (average absolute error range = 39.14 to 60.44). Results also show that 10-s MTS and 30-s MTS over and underestimated by small margins, and 10-s and 30-s PIR overestimated by large margins (Figure 3). The original hypothesis was partially supported, as MTS outperformed PIR, and 10-s PIR outperformed 30-s PIR in terms of

producing the lowest levels of measurement error. However, 30-s MTS, and not 10-s MTS, was the most accurate method as it produced the lowest measurement error in two of three participants.

Treatment Integrity

Figure 4 depicts TI across three teachers. For all participants, TI was highest in the no data collection condition, in which they implemented a skill acquisition protocol without collecting discontinuous data. Although TI was slightly lower in the various MTS and PIR data collection conditions, none of the procedures compromised TI to an unacceptable level; with the exception of one session, Teacher A's first session, the treatment integrity for all sessions was greater than 90%. For Teacher A, TI was consistently high and ranged from 94-100% with the exception of the first session, 10-s MTS, in which TI was only 78%. Among the data collection conditions, TI was highest in the 30-s MTS condition and lowest in the 10-s PIR condition. For Teacher B, TI ranged from 90.6-100%. TI was similar for 10-s PIR, 30-s PIR, and 10-s MTS. TI was lowest in the 30-s MTS condition, but still exceeded 90% for all sessions. Teacher C's TI ranged from 94-100%, and was 100% for sessions in which no stereotypy data were collected. Of the discontinuous data collection procedures, TI was highest with 30-s PIR and lowest with 30-s MTS. Overall, for two of three participants, the 30-s MTS condition resulted in the lowest TI but for the third participant this same condition yielded the highest TI.

Perceptions of Importance, Accuracy, and Ease

The initial questionnaire (Appendix C) given to each teacher prior to their first session indicated that, in general, ease, accuracy, and duration of procedure (i.e., observation length) were considered important when choosing data collection procedures. As shown in Figure 5, the lowest rating was Teacher A's rating for the importance of the duration of a procedure (rating =

3). Notably, Teacher B rated all three factors as "extremely important" (ratings = 5), and although accuracy and ease were considered equally important for Teacher A and B, Teacher C was the only one to rate the accuracy of the procedure as more important than its ease.

Figure 6a depicts the teachers' perceptions of ease across discontinuous data procedures following implementation of the procedures. As hypothesized, Teachers A and B perceived 30-s MTS as easiest to implement, followed by 30-s PIR and 10-s MTS. Ten-second PIR was perceived as the most difficult to implement. Contrary to the hypothesis, Teacher C rated the methods similarly on ease of implementation (range = 3-4.5) but overall perceived PIR to be easier to implement than MTS.

Figure 6b depicts the teachers' perceptions of accuracy of each discontinuous data procedures. Across participants, PIR was perceived to be more accurate than MTS, but 10-s intervals were not perceived to be more accurate than 30-s intervals. Teacher A and C rated 30-s PIR as most accurate. Teacher B perceived all methods except 30-s MTS as accurate. Similarly, Teacher A also perceived 30-s MTS as least accurate, while Teacher C perceived 10-s MTS as least accurate.

Preference Assessment

The results of the MSWO, shown in Figure 7, indicated that preferences varied across participants. In the MSWO, Teacher A and B chose 30-s MTS first as their most preferred data collection method, 10-s MTS second, 10-s PIR third, leaving 30-s PIR as least preferred. In contrast, Teacher C preferred 30-s PIR most, followed by 10-s PIR, 30-s MTS, and 10-s MTS.

Discussion

This study compared the implementation of 30-s MTS, 30-s PIR, 10-s MTS, and 10-s PIR by teachers in a naturalistic setting. It sought to identify the discontinuous data collection

method with the least measurement error when used by teachers who were simultaneously implementing a skill acquisition procedure. The present study also aimed to identify the method that produced the least human error, as indicated by IOA and TI, and examined teacher perceptions and preferences for the four different procedures.

Measurement Error

Across the three teachers, MTS provided more accurate estimates of absolute duration than PIR. This finding is consistent with those past studies that identified MTS as the most accurate discontinuous data collection procedure for duration measurement in more contrived settings (Alvero et al., 2007; Harrop & Daniels, 1986; Murphy & Goodall, 1980; Powell et al., 1975; Powell et al., 1977). Furthermore, this study replicated Delmolino et al.'s (2008) finding that MTS was more accurate than PIR when used by teachers collecting in vivo stereotypy data. Their finding that MTS methods produced less measurement error even in instances when MTS had higher human error than PIR was also replicated. For example, for Teacher A, the IOA for 30-s PIR was higher than the IOA for 10-s MTS, yet the absolute measurement error was less for 10-s MTS (measurement error = 4.06 percentage points) than for 30-s PIR (measurement error = 39.86 percentage points). Consistent with past literature, results indicated that PIR overestimates with large error margins while MTS under and overestimates but with smaller error margins (Ciotti Gardenier et al., 2004; Green et al., 1982; Harrop & Daniels, 1986; Murphy & Goodall, 1980; Powell et al., 1975; Powell et al., 1977; Rapp et al., 2008).

MTS was also more flexible than PIR; a change from 10-s to 30-s interval length did not result in large increases in measurement error for MTS. In fact, measurement error was lower with 30-s MTS for 30 min than 10-s MTS for 10 min. In contrast, measurement error of 30-s PIR was up to 140% greater than that of 10-s PIR. Although the present study only tested two interval

lengths that differed by only 20 s, this finding is consistent with Hanley et al.'s (2007) finding that behavior estimates obtained by 120-s MTS intervals were within 5% of those obtained by 5-s MTS intervals.

With regards to methodological error, findings from this investigation suggest that either 10-s MTS for 10 min or 30-s MTS for 30-min be used when collecting discontinuous data for absolute duration of stereotypy.

Teacher IOA

Across participants, teachers were able to collect MTS data with higher IOA when compared to PIR. For all three participants, the poorest IOA occurred when using a PIR procedure, thus the results from this study suggest that MTS may be less susceptible to human error.

The results from this study also expand on Dorsey et al. (1986) and Hanley et al.'s (2007) findings that higher demand intensity of the data collection method was associated with lower IOA. For the present study, when comparing the same discontinuous data collection procedure, the 30-s interval had higher IOA than the 10-s interval in three of six comparisons. In contrast, the 10-s interval had higher IOA in only one comparison. No detectable differences were observed between the IOAs in the remaining two comparisons. These findings suggest that teachers may have had greater difficulty coding behavioral data with more frequent and shorter intervals.

However, no evidence emerged to indicate that multitasking teachers in a naturalistic setting had any more difficulty collecting data with high IOA than did researchers who collected data in controlled or less demanding settings in other studies. The IOA for all teachers in the current study was high and exceeded 80% with all four data collection procedures. Teachers in

the current study were able to conduct a skill acquisition protocol and reliably collect data with interval lengths as short as 10-s. These findings differ from those from Delmolino et al.'s (2008) naturalistic study that found teacher IOA to fall below 80% (average MTS IOA = 72.1% and average PIR IOA = 78.5%). One explanation for this difference is that the teachers' discontinuous data collection in the current study was video recorded so that adjustments could be made to correct for skipped or double-coded intervals. Additionally, teacher IOA may have been higher in this investigation because the data sheets included a column that indicated the number of intervals remaining in the session. The interval timer similarly indicated the number of intervals remaining, so teachers likely self-corrected and tracked their place on the data sheet more easily than in Delmolino et al.'s (2008) study.

Treatment Integrity

The lower TI in data collection conditions as compared to teaching only conditions indicates that treatment complexity does increase human error. Since no study to date has looked at the effect of simultaneous data collection on TI, this finding is an important addition to the sparse literature on this topic. The decrease in TI associated with concurrent data collection, however, was small. Only Teacher A's first session had a TI below 90%, likely due to her lack of familiarity with the new data sheets and specific teaching procedure. Furthermore, no clear evidence emerged that one type of data collection procedure compromised TI more than another.

These findings are encouraging because they suggest that teachers can carry out complex data collection procedures while maintaining an acceptable level of TI in their teaching. We cannot, however, conclude whether the small reductions in TI affected student outcomes. Nevertheless, given what research says about the relation between high TI and internal and external validity, the high TI across conditions in the current study suggests that teachers in a

classroom can implement instructional programs with high enough fidelity to interpret their functional relationships and/or test their generality through replication.

Perceived Accuracy and Ease

Overall, teachers were more likely to incorrectly perceive 10-s and 30-s PIR as more accurate than 10-s and 30-s MTS. All three teachers rated 30-s PIR as the most accurate (or just as accurate as other procedures), when in fact it had the most measurement error. In contrast, 30-s MTS was perceived as the least accurate when averaged across three participants, yet it produced estimates that were very close to the actual duration of stereotypy. One possible explanation for teachers' perceptions is that they may have felt that each instance of behavior was counted using PIR, as they marked a "+" in each interval when behaviors occurred. In contrast, with MTS, they marked a "-" when a behavior occurred if it did not coincide with the 1-s beep of the timer signaling the end of the interval. Teacher reports supported this idea as one teacher said, "PIR seems to capture more behavior than MTS." Additionally, as suggested by Fiske and Delmolino (2008), teachers may have experienced a feeling of dissatisfaction when they "missed" (e.g., it occurred just before the timer sounded) a behavior when using MTS, and especially 30-s MTS. This possible explanation is consistent with the patterns in perceived accuracy; with 30-s PIR, the method perceived as most accurate, teachers are most likely to "catch" an instance of behavior within each interval and for 30-s MTS, the method perceived as least accurate, they would be least likely to do so.

When comparing 30-s and 10-s intervals while controlling for the participant and discontinuous data collection procedures (i.e. comparing Teacher A's 30-s MTS to Teacher A's 10-s MTS and Teacher A's 30-s PIR to Teacher A's 10-s PIR), on five of six comparisons 30-s intervals were rated as easier to implement than 10-s intervals. Teacher A explained that the

longer intervals were easier because she "only had to take data every 30 seconds vs. 10 seconds." Likely, the less frequent interruptions allowed her to focus more on the various tasks in which she was engaged. The perceived ease, however, did not necessarily predict actual teacher performance as indicated by their IOA or TI. For example, Teacher C rated both 10-s and 30-s MTS as the most difficult to implement but both procedures had the highest IOA. In contrast, Teacher A rated 10-s PIR as the most difficult to implement and it corresponded with the poorest IOA and poorest TI. These inconsistencies between perceptions and the actual outcomes suggest that teachers might choose a suboptimal discontinuous data collection procedure if their preferences were based solely on perceptions of ease and accuracy.

Teacher Preference

The MSWO results indicated that MTS was more preferred than was PIR in two of the three cases. Teachers A and B, whose rankings were identical, both preferred 30-s MTS the most and perceived it as the easiest to implement, but as the least accurate. This finding seems consistent with Kolt and Rapp's (2014) results that revealed that most therapists preferred 60-s intervals over 10-s intervals primarily due to perceived ease, rather than accuracy. A closer look, however, reveals that perceived ease was not consistently associated with preference. Teacher A and B's last MSWO choice, 30-s PIR, was also rated highly in both ease and accuracy. Therefore, other perceptions likely affected teachers' preference; although Teacher A perceived 30-s PIR as easy, she described 30-s PIR as a "marathon," hinting to the possibility that longer session durations could also affect preferences in some cases. For instance, using 30-s PIR, teachers had to monitor the target behavior for the entirety of every interval, but for 30-s MTS they were only required to attend to the behavior for 1 s at the termination of an interval. In both cases, teachers recorded data for 30 min but the task may have become more tiring or boring

with 30-s PIR than with 30-s MTS.

In contrast to Teachers A and B, Teacher C preferred PIR to MTS. When asked to explain her preferences, she said that she preferred PIR because it seemed to capture the most stereotypy. Interestingly, for the initial questionnaire that used a 5-point Likert scale (Appendix C), Teacher C was the only participant who rated accuracy (rating = 5) as more important than ease (rating = 4) or session duration (rating = 4) when selecting a data collection system. Teachers A and B rated both ease and accuracy as "extremely important" (rating = 5). These findings suggest that the factors that affect preference are complex and vary across individuals.

The present study did not identify a pattern in the way perceptions, as indicated on the Teacher Perceptions Form (Appendix D), affected preference but it demonstrated that perceptions of ease and accuracy were not associated with human error. This finding is consistent with the existing literature that indicates that acceptability, a type of perception, does not affect human error (Noell et al., 2005; Sterling-Turner, 2002; Wickstrom et al., 1998).

The teacher participants in this study performed the various tasks with very low human error, yet all of them had erroneous perceptions of the accuracy of MTS and PIR. Despite the high skill level of these instructors, their common misconception that PIR produced more accurate estimates of behavior duration than MTS could result in another form of human error, as they could choose data collection procedures that are prone to high measurement error. For example, Teacher C chose 30-s PIR first in the MSWO and said, anecdotally, that she would have chosen this method outside of this research context, as she believed it was the most accurate. Further education to correct misconceptions about MTS and PIR could help shift preferences to more accurate methods.

Limitations and Future Directions

The present study found important evidence that 10-s and 30-s MTS, but not PIR, provided accurate duration estimates of stereotypy and demonstrated that teachers can implement instructional programs and collect discontinuous data while maintaining high TI in instruction and high IOA in data collection. Additionally, it revealed the tendency for teachers to incorrectly perceive PIR as more accurate than MTS. Certain limitations, however, require that readers interpret these and other findings with caution.

This study included only DDDC aides who were college graduates in their mid-20s with over 1 year of experience. Further investigation is necessary to determine whether results from this study, particularly the high IOA and TI, would generalize to ABA practitioners from the broader community. The DDDC aides may differ from most ABA practitioners, as they receive rigorous and ongoing staff training, have daily contact with a large number of peers and supervisors with BCBAs, and work in a setting with a relative abundance of staffing, time, and material resources. Future studies could test the generality of the results by replicating the study with more typical ABA practitioners in other settings, or by evaluating these procedures with newly hired staff at the DDDC.

The results are also tempered by a small sample size of three teacher-student dyads, as well as a small sample of data: Many results were based on only three data points per condition. The limited data and their lack of differentiation limited the ability to draw strong conclusions, especially regarding teacher IOA, TI, and perceptions. Future research could address this weakness by recruiting and evaluating a larger number of teacher-student dyads and by running more sessions to obtain clearer differentiation between conditions.

Although the study aimed to maximize the external validity of the results by using a more

naturalistic environment than previous research, limitations still emerged in the generality of the findings due to the specifics of this research protocol. For example, teachers were videotaped and knew that the primary investigator would examine their data, and thus the findings may not generalize easily to settings where teacher performance is not evaluated regularly. Also, the receptive language and matching tasks used with students in this evaluation were mostly mastered skills the learners previously had in their repertoire, and were answered correctly by the students on the vast majority of trials. The tasks were also done while students were seated and required limited materials. Further research is necessary to determine whether the findings of this study can be replicated with new programs or demanding multi-step tasks.

Another limitation of the study was that teachers were only required to collect data on stereotypy, in addition to data on the teaching task. The generality of these findings to other behaviors, especially those with much longer or shorter bout lengths, remains untested. Also, although each student engaged in more than one topography of stereotypy, they were not coded separately. Future studies could manipulate the numbers of behaviors teachers are required to code and examine subsequent effects on their IOA and TI. Since the current literature is sparse and inconclusive (Dorsey et al., 1986; Murphy & Harrop, 1994), such work may aid in clarifying this relationship to inform best practice in data collection.

Furthermore, while this study assessed the accuracy of 10-s and 30-s MTS and PIR, future research should examine the sensitivity of these methods when detecting changes in behavior in a naturalistic setting. This evaluation is an essential next step in the literature, as most practitioners are concerned about increases or decreases in behavior frequency and duration rather than isolated data points. Sensitivity research in a naturalistic setting is an especially exciting area for future inquiry because, despite its superiority in estimating absolute duration,

MTS is not necessarily better than PIR in detecting changes in behavior frequency or duration. For example, existing research indicates that 10-s PIR for 10 min is as sensitive as 30-s MTS for 30 min, and is more sensitive than 10-s MTS for 10 min when measuring changes in duration events (Devine et al., 2011).

Lastly, future research must look at other dimensions of teachers' perception of data collection methods beyond ease and accuracy. Researchers should ask instructional staff for more detailed explanations regarding their preferences. These studies should investigate teachers' experiences with data collection systems, such as their level of comfort or boredom when collecting data as examined in Hanley et al. (2007) and level of frustration as suggested by Fiske and Delmolino (2008). Additionally, preference assessments should also be included in these studies to better understand the relationship between teacher perceptions and preferences for data collection procedures. This research would add to the limited literature on social validity of discontinuous data collection and begin to explain how various perceptions affect choice making.

Teaching Implications

Despite several limitations, the results of this study provide some guidelines for teachers collecting discontinuous data. The present study demonstrated that multitasking teachers can collect discontinuous data with intervals as short as 10 s without compromising TI and IOA to unacceptable levels. The findings also suggest that 10-s and 30-s PIR yield high measurement error and should generally not be used to estimate duration of behavior. Taken together, these findings indicate that teachers may choose between 10-s MTS for 10 min and 30-s MTS for 30 min when collecting data on duration events in the classroom. Lastly, teachers could benefit from further education regarding discontinuous data collection procedures to correct erroneous

perceptions that PIR is more accurate than MTS when used for estimating duration of a behavior. Since the responses of one of the teachers in the current study indicates that perceptions of a data system's accuracy can influence preference, such staff training may increase the likelihood that staff will choose an appropriate data collection procedure.

By following these recommendations, instructional staff can implement treatments with high fidelity and collect data that are accurate and reliable so that functional relationships detected between treatments (IVs) and behavior data (DVs) can be interpreted with confidence. Teachers can use these discontinuous data to inform treatment planning; they can determine when to terminate or modify ineffective interventions, and when to continue effective ones. In doing so, teachers can avoid unnecessary harm to their learners and improve students' outcomes in a timely and meaningful manner.

References

Alvero, A. M., Struss, K., & Rappaport, E. (2008). Measuring safety performance: A comparison of whole, partial, and momentary time-sampling recording methods. *Journal of Organizational Behavior Management*, *27*(4), 1-28.

Behavior Analysis Certification Board (BACB). (2016). Professional and Ethical Compliance Code for Behavior Analysts. Retrieved February 2, 2016, from http://bacb.com/wp-content/uploads/2016/01/160120-compliance-code-english.pdf

Brittle, 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– 488.

Carroll, R. A., Kodak, T., & Fisher, W. W. (2013). An evaluation of programmed treatment-integrity errors during discrete-trial instruction. *Journal of Applied Behavior Analysis*, *46*(2), 379-394.

Ciotti Gardenier, N., MacDonald, R., & Green, G. (2004). Comparison of direct observational methods for measuring stereotypic behavior in children with autism spectrum disorders. *Research in Developmental Disabilities*, *25*(2), 99-118.

Codding, R. S., Feinberg, A. B., Dunn, E. K., & Pace, G. M. (2005). Effects of immediate performance feedback on implementation of behavior support plans. *Journal of Applied Behavior Analysis*, *38*(2), 205-219.

Cooper, J. O., Heron, T. E., & Heward, W. L. (2007). *Applied behavior analysis*, 2nd ed. Upper Saddle River, N.J.: Pearson Prentice Hall.

DeLeon, I. G., & Iwata, B. A. (1996). Evaluation of a multiple-stimulus presentation format for assessing reinforcer preferences. *Journal of Applied Behavior Analysis*, *29*(4), 519-533.

Devine, S. L., Rapp, J. T., Testa, J. R., Henrickson, M. L., & Schnerch, G. (2011). Detecting

changes in simulated events using partial-interval recording and momentary time sampling III: Evaluating sensitivity as a function of session length. *Behavioral Interventions, 26,* 103–124.

DiGennaro, F. D., Martens, B. K., & Kleinmann, A. E. (2007). A comparison of performance feedback procedures on teachers' treatment implementation integrity and students' inappropriate behavior in special education classrooms. *Journal of Applied Behavior Analysis*, *40*(3), 447-461.

DiGennaro, F. D., Martens, B. K., & McIntyre, L. L. (2005). Increasing treatment integrity through negative reinforcement: Effects on teacher and student behavior. *School Psychology Review*.

Dorsey, B. L., Nelson, R. O., & Hayes, S. C. (1986). The effects of code complexity and of behavioral frequency on observer accuracy and interobserver agreement. *Behavioral Assessment*, *8*(4), 349-363.

Fiske, K. E. (2008). Treatment integrity of school-based behavior analytic interventions: A review of the research. *Behavior Analysis in Practice*, *1*(2), 19.

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

Fiske, K. E., & Delmolino, L. (2008, May). Teacher perceptions of and accuracy in data collection using momentary time sampling and partial interval recording. In K. Fiske (Chair), *The Assessment and Application of Momentary Time Sampling and Partial Interval Recording in Classroom Settings*. Symposium presented at annual meeting of Association for Behavior Analysis. Chicago, IL.

Delmolino, L., Fiske, K. E., & Dackis, M. (2008, May). Comparing the use of momentary time

sampling and partial interval recording in stereotypy data collection. In K. Fiske (Chair), *The Assessment and Application of Momentary Time Sampling and Partial Interval Recording in Classroom Settings*. Symposium presented at annual meeting of Association for Behavior Analysis. Chicago, IL.

Green, S. B., McCoy, J. F., Burns, K. P., & Smith, A. C. (1982). Accuracy of observational data with whole interval, partial interval, and momentary time-sampling recording techniques. *Journal of Behavioral Assessment*, *4*(2), 103-118.

Gresham, F. M. (1989). Assessment of treatment integrity in school consultation and prereferral intervention. *School Psychology Review*, 18, 37-50.

Gresham, F. M., Gansle, K. A., & Noell, G. H. (1993). Treatment integrity in applied behavior analysis with children. *Journal of Applied Behavior Analysis*, *26*(2), 257-263.

Gresham, F. M., MacMillan, D. L., Beebe-Frankenberger, M. E., & Bocian, K. M. (2000). Treatment integrity in learning disabilities intervention research: Do we really know how treatments are implemented?. *Learning Disabilities Research & Practice*, *15*(4), 198-205.

Gunter, P. L., Venn, M. L., Patrick, J., Miller, K. A., & Kelly, L. (2003). Efficacy if using momentary time samples to determine on-task behavior of students with Emotional/Behavioral disorders. *Education & Treatment of Children, 26*(4), 400-412.

Hagermoser Sanetti, L. M., Luiselli, J. K., & Handler, M. W. (2007). Effects of verbal and graphic performance feedback on behavior support plan implementation in a public elementary school. *Behavior Modification*, *31*(4), 454-465.

Hanley, G. P., Cammilleri, A. P., Tiger, J. H., & Ingvarsson, E. T. (2007). A method for describing preschoolers' activity preferences. *Journal of Applied Behavior Analysis*, *40*(4), 603–618.

Hansen, W. B., Graham, J. W., Wolkenstein, B. H., & Rohrbach, L. A. (1991). Program integrity as a moderator of prevention program effectiveness: Results for fifth-grade students in the adolescent alcohol prevention trial. *Journal of Studies on Alcohol and Drugs*, *52*(6), 568.

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.

Holcombe, A., Wolery, M., & Snyder, E. (1994). Effects of two levels of procedural fidelity with constant time delay on children's learning. Journal of Behavioral Education, 4, 49–73.

Jacobsen, N. K., (1982). Temporal and procedural influences on activity estimated by time-sampling. *Journal of Wildlife Management, 46*(2), 1982.

Jones, K. M., Wickstrom, K. F., & Friman, P. C. (1997). The effects of observational feedback on treatment integrity in school-based behavioral consultation. *School Psychology Quarterly, 12*(4), 316.

Kazdin, A. E. (1977). Artifact, bias, and complexity of assessment: The ABCs of reliability. *Journal of Applied Behavior Analysis*, *10*(1), 141-150.

Kelly, M. B. (1977). A review of the observational data- collection and reliability procedures reported in the Journal of Applied Behavior Analysis. *Journal of Applied Behavior Analysis, 10*, 97–101.

Kolt, L. D., & Rapp, J. T. (2014). Assessment of therapists' preferences for discontinuous measurement systems. *Behavioral Interventions*, *29*(4), 304-314.

Lane, J. D., & Ledford, J. R. (2014). Using Interval-Based Systems to Measure Behavior in Early Childhood Special Education and Early Intervention. *Topics in Early Childhood Special Education*, 0271121414524063.

Mayer, R., Sulzer-Azaroff, B., & Wallace, M. (2014). *Behavior analysis for lasting change* (3rd ed.). Cornwall-on-Hudson, NY: Sloan Publishing.

Meany-Daboul, M. G., Roscoe, E. M., Bourret, J. C., & Ahearn, W. H. (2007). A comparison of momentary time sampling and partial-interval recording for evaluating functional relations. *Journal of Applied Behavior Analysis, 40,* 501–514.

Mintz J. C. (2011). An analysis of the methodological and human error within momentary time sampling data collection (Unpublished doctoral dissertation). Louisiana State University, Baton Rouge, Louisiana.

Moncher, F. J., & Prinz, F. J. (1991). Treatment fidelity in outcome studies. Clinical Psychology Review, 11, 247–266.

Mortenson, B. P., & Witt, J. C. (1998). The use of weekly performance feedback to increase teacher implementation of a prereferral academic intervention. *School Psychology Review,*

Mudford, O. C., Beale, I. L., & Singh, N. N. (1990). The representativeness of observational samples of different durations. *Journal of Applied Behavior Analysis, 23*, 323–331.

Mudford, O. C., Taylor, S. A., & Martin, N. T. (2009). Continuous recording and inter-observer agreement algorithms reported in the *Journal of Applied Behavior Analysis* (1995–2005). *Journal of Applied Behavior Analysis, 42,* 165–169.

Murphy, G., & Goodall, E. (1980). Measurement error in direct observations: A comparison of common recording methods. *Behaviour Research and Therapy, 18,* 147–150.

Murphy, M. J., & Harrop, A. (1994). Observer error in the use of momentary time sampling and partial interval recording. *British Journal of Psychology, 85,* 169–17

Noell, G. H. (2007). Research examining the relationships among consultation process, treatment

integrity, and outcomes. In W. P. Erchul & S. M. Sheridan (Eds.), Handbook of research in school consultation: Empirical foundations for the field (pp.315-334). Mahwah, NJ: Erlbaum.

Noell, G. H., Duhon, G. J., Gatti, S. L., & Connell, J. E. (2002). Consultation, follow-up, and implementation of behavior management interventions in general education. *School Psychology Review, 31*(2), 217-234.

Noell, G. H., Gresham, F. M., & Gansle, K. A. (2002). Does treatment integrity matter? A preliminary investigation of instructional implementation and mathematics performance. *Journal of Behavioral Education*, *11*(1), 51-67.

Noell, G. H., Witt, J. C., Gilbertson, D. N., Ranier, D. D., & Freeland, J. T. (1997). Increasing teacher intervention implementation in general education settings through consultation and performance feedback. *School Psychology Quarterly, 12*(1), 77.

Noell, G. H., Witt, J. C., Slider, N. J., Connell, J. E., Gatti, S. L., Williams, K. L., . . . Duhon, G. J. (2005). Treatment implementation following behavioral consultation in schools: A comparison of three follow-up strategies. *School Psychology Review,*

Noell, G., Witt, J., LaFleur, L., Mortenson, B., Ranier, D., & LeVelle, J. (2000). A comparison of two follow-up strategies to increase teacher intervention implementation in general education following consultation. *Journal of Applied Behavior Analysis, 33*(1), 271-284.

Northup, J., Fisher, W., Kahng, S., Harrel, B., & Kurtz, P. (1997). An assessment of the necessary strength of behavioral treatments for severe behavior problems. *Journal of Developmental and Physical Disabilities, 9*, 1–16.

Perepletchikova, F., & Kazdin, A. E. (2005). Treatment integrity and therapeutic change: Issues and research recommendations. *Clinical Psychology: Science and Practice*, *12*(4), 365-

383.

Peterson, L., Horner, A., & Wonderlich, S. (1982). The integrity of independent variables in behavior analysis. *Journal of Applied Behavior Analysis, 15,* 477–492.

Powell, J., Martindale, A., & Kulp, S. (1975). An evaluation of time-sample measures of behavior. *Journal of Applied Behavior Analysis, 8,* 463–469.

Powell, J., Martindale, B., Kulp, S., Martindale, A., & Bauman, R. (1977). Taking a closer look: Time sampling and measurement error. *Journal of Applied Behavior Analysis, 10,* 325–332.

Powell, J., & Rockinson, R. (1978). On the inability of interval time sampling to reflect frequency of occurrence data. *Journal of Applied Behavior Analysis, 11*, 531-532.

Progar, P. R., Perrin, F. A., DiNovi, B. J., Bruce, S. S., & NeuroHealth, B. (2001). Treatment integrity: Some persistent concerns and some new perspectives. *A Context for Science with a Commitment for Behavior Change*, 28.

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

Rapp, J. T., Colby-Dirksen, A. M., Michalski, D. N., Carroll, R. A., & Lindenberg, A. M. (2008). Detecting changes in simu- lated events using partial-interval recording and momentary time sampling. *Behavioral Interventions, 23,* 237–269.

Rapp, J. T., Colby, A. M., Vollmer, T. R., Roane, H. S., Lomas, J., & Britton, L. N. (2007). Interval recording for duration events: A re-evaluation. *Behavioral Interventions, 22,* 319–345.

Repp, A. C., Roberts, D. M., Slack, D. J., Repp, C. F., & Berkler, M. S. (1976). A comparison of frequency, interval, and time- sampling methods of data collection. *Journal of Applied Behavior Analysis, 9,* 501–508.

Rhymer, K. N., Evans-Hampton, T. N., McCurdy, M., & Watson, T. S. (2002). Effects of varying levels of treatment integrity on toddler aggressive behavior. *Special Services in the Schools*, *18*(1-2), 75-82.

Saudargas, R. A., & Zanolli, K. (1990). Momentary time sampling as an estimate of percentage time: A field validation. *Journal of Applied Behavior Analysis, 23,* 533–537.

Sterling-Turner, H. E., Watson, T. S., & Moore, J. W. (2002). The effects of direct training and treatment integrity on treatment outcomes in school consultation. *School Psychology Quarterly*, *17*(1), 47.

Sterling-Turner, H. E., Watson, T. S., Wildmon, M., Watkins, C., & Little, E. (2001). Investigating the relationship between training type and treatment integrity. *School Psychology Quarterly, 16*(1), 56.

Taylor, M. A., Skourides, A., & Alvero, A. M. (2012). Observer error when measuring safety-related behavior: Momentary time sampling versus whole interval recording. *Journal of Organizational Behavior Management, 32,* 307–319.

Tiger, J. H., Miller, S. J., Mevers, J. L., Mintz, J. C., Scheithauer, M. C., & Alvarez, J. (2013). On the representativeness of behavior observation samples in classrooms. *Journal of Applied Behavior Analysis*, *46*(2), 424-435.

Vollmer, T. R., Roane, H. S., Ringdahl, J. E., & Marcus, B. A. (1999). Evaluating treatment challenges with differential reinforcement of alternative behavior. *Journal of Applied Behavior Analysis, 32*, 9–23.

Vollmer, T. R., Sloman, K. N., & Pipkin, C. S. P. (2008). Practical implications of data reliability and treatment integrity monitoring. *Behavior Analysis in Practice*, *1*(2), 4.

Wheeler, J. J., Baggett, B. A., Fox, J., & Blevins, L. (2006). Treatment Integrity A Review of Intervention Studies Conducted With Children With Autism. *Focus on Autism and Other Developmental Disabilities*, *21*(1), 45-54.

Wickstrom, K. F., Jones, K. M., LaFleur, L. H., & Witt, J. C. (1998). An analysis of treatment integrity in school-based behavioral consultation. *School Psychology Quarterly*, *13*(2), 141.

Wilder, D. A., Atwell, J., & Wine, B. (2006). The effects of varying levels of treatment integrity on child compliance during treatment with a three-step prompting procedure. *Journal of Applied Behavior Analysis, 39,* 369–373.

Wilkinson, L. A. (2007). Assessing treatment integrity in behavioral consultation. *International Journal of Behavioral Consultation and Therapy*, *3*, 420-432.

Wirth, O., Slaven, J., & Taylor, M. A. (2014). Interval sampling methods and measurement error: A computer simulation. *Journal of Applied Behavior Analysis*, *47*(1), 83-100.

Witt, J. C., Noell, G. H., LaFleur, L. H., & Mortenson, B. P. (1997). Teacher use of interventions in general education settings: Measurement and analysis of the independent variable. *Journal of Applied Behavior Analysis, 30*(4), 693-696.

Yeaton, W. H., & Sechrest, L. (1981). Critical dimensions in the choice and maintenance of successful treatments: strength, integrity, and effectiveness. Journal of Consulting and Clinical Psychology, 49(2), 156.

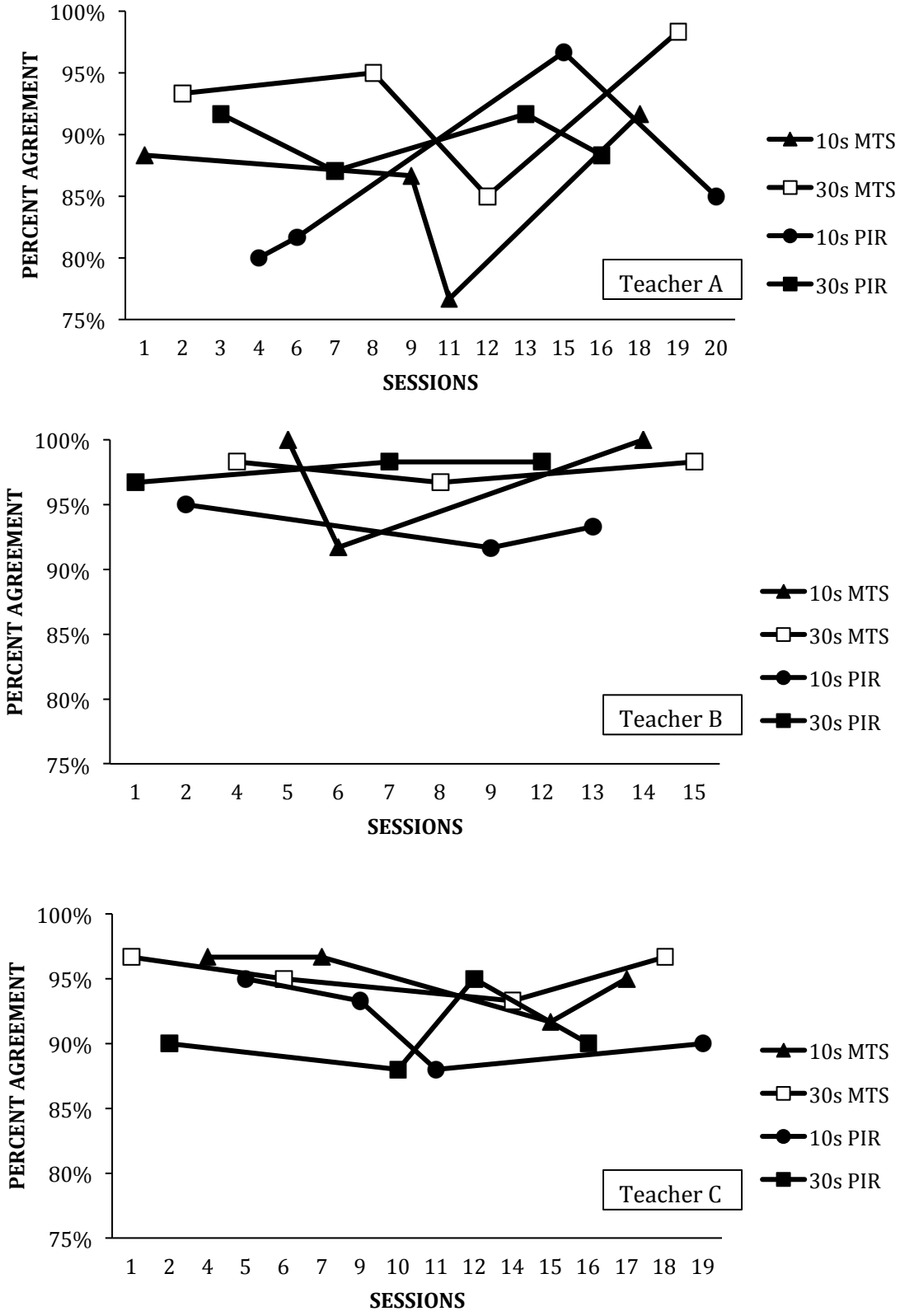

Figure 1. Teacher A (top), Teacher B (middle), and Teacher C's (bottom) IOA across discontinuous data collection procedures.

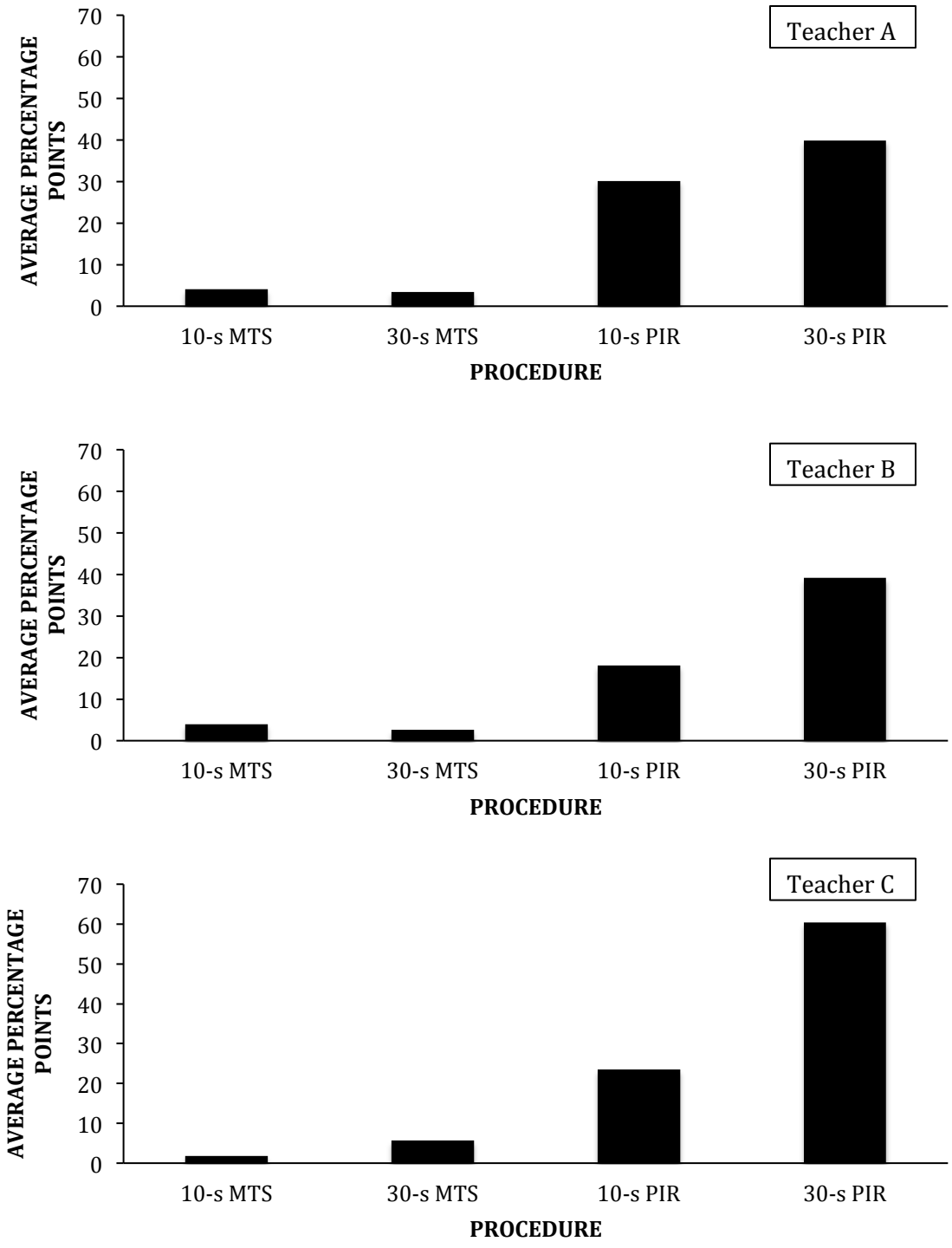

Figure 2. Teacher A (top), Teacher B (middle), and Teacher C's (bottom) average measurement error across discontinuous data procedures. Measurement error represented as the average absolute number of percentage points from the true value

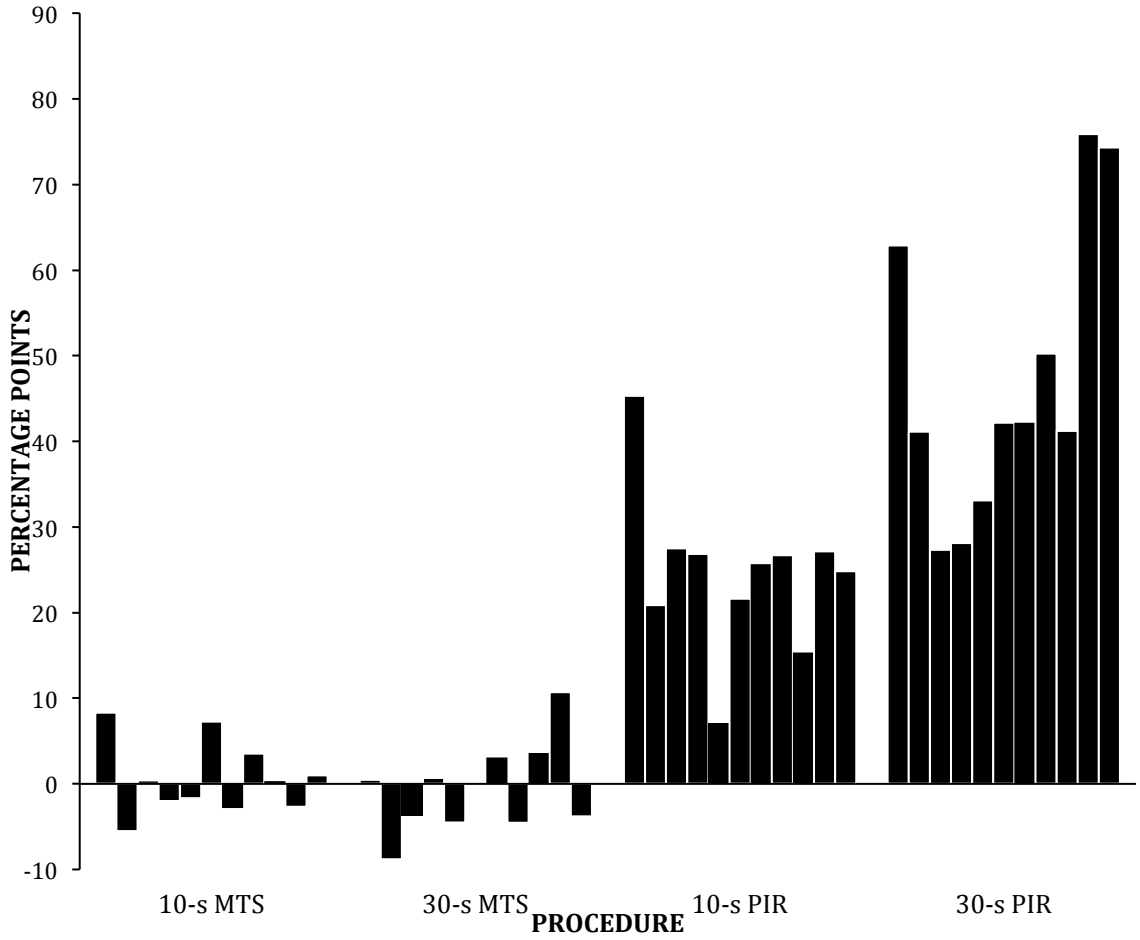

Figure 3. Measurement error across discontinuous data procedures. Measurement error represented as difference in percentage points between discontinuous estimate and true value.

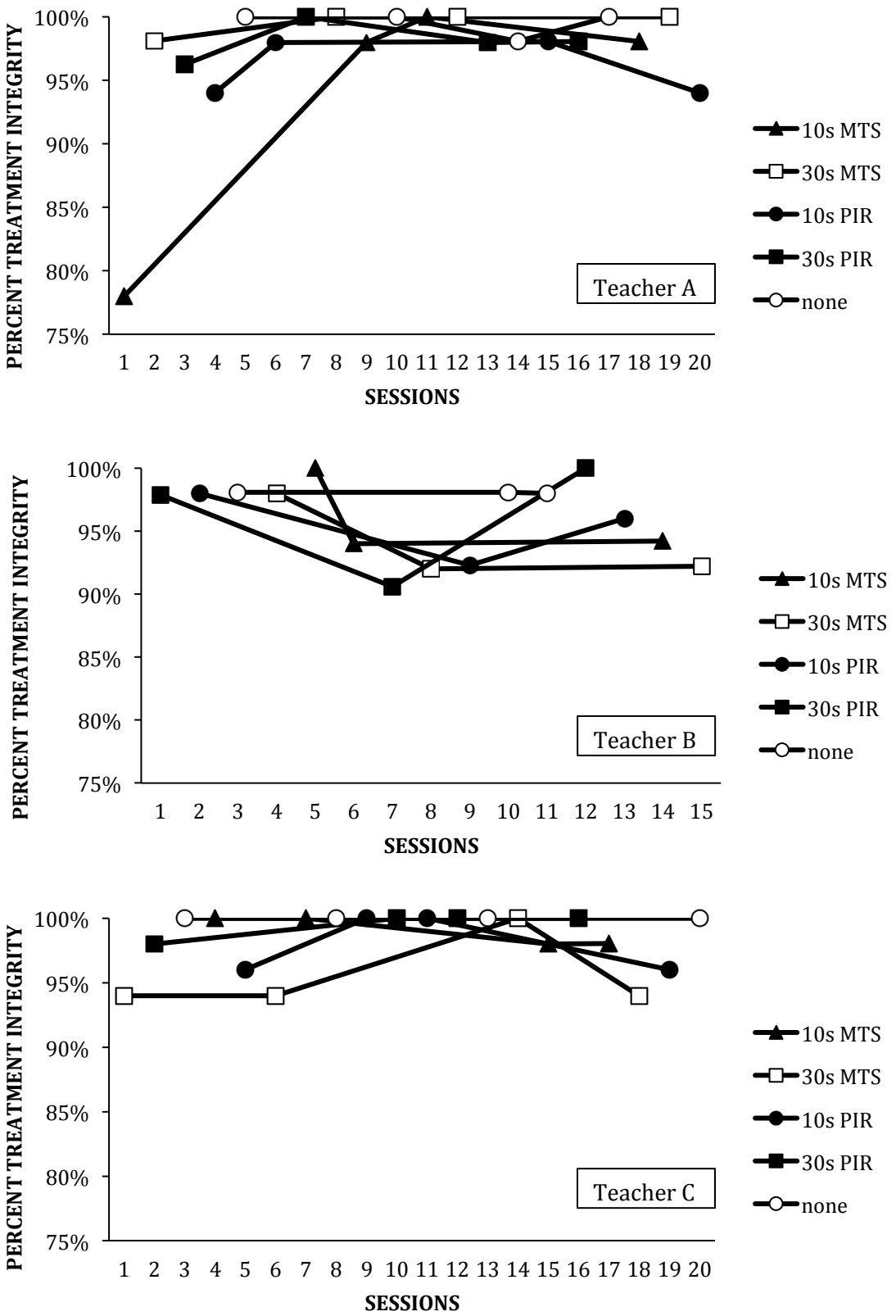

Figure 4. Teacher A (top), Teacher B (middle), and Teacher C's (bottom) treatment integrity across discontinuous data collection procedures.

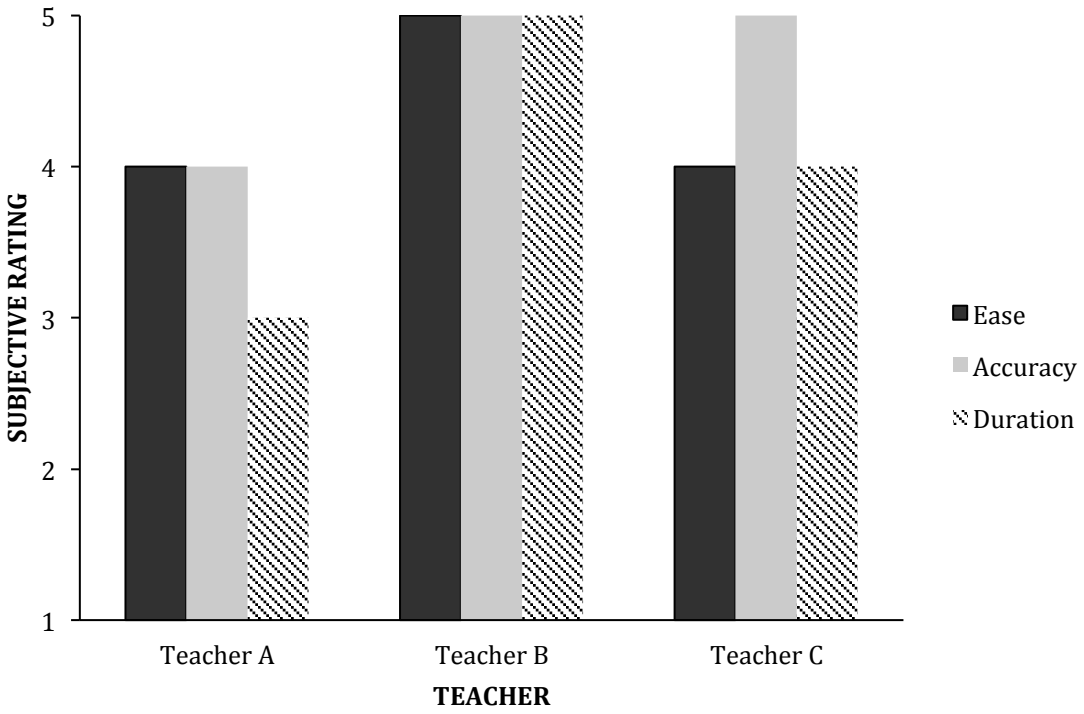

Figure 5. Teacher rating of importance of ease, accuracy, and duration when choosing a data collection procedure.

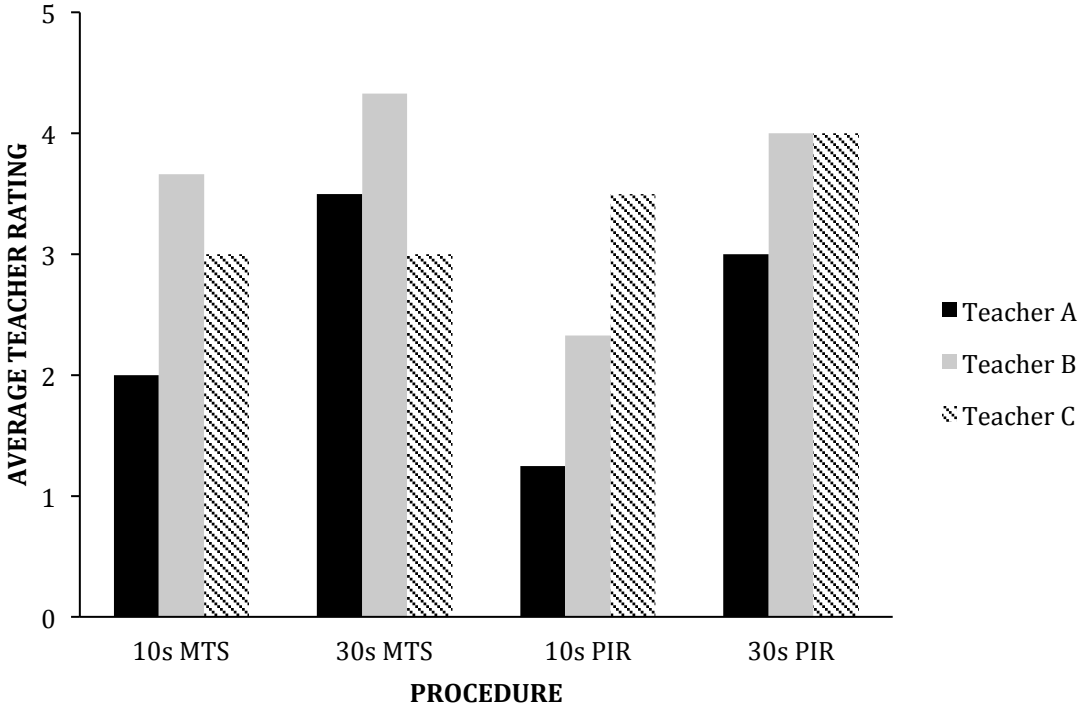

Figure 6a. Average teacher rating of perceived ease for each discontinuous data collection procedure.

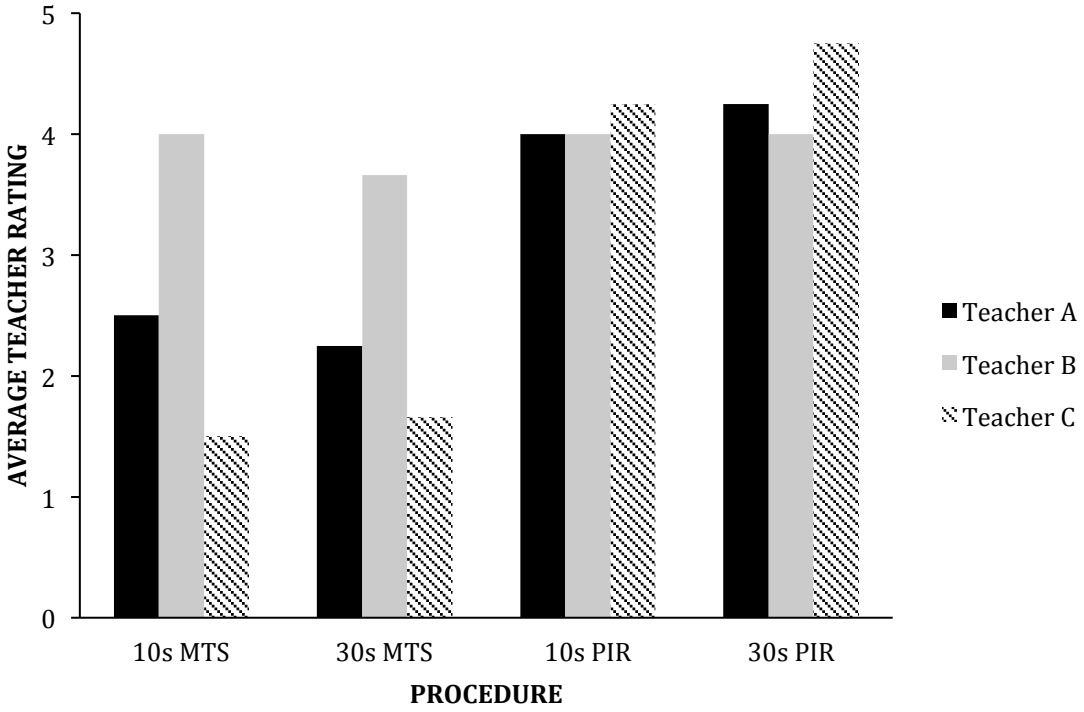

Figure 6b. Average teacher rating of perceived accuracy for each discontinuous data collection procedure.

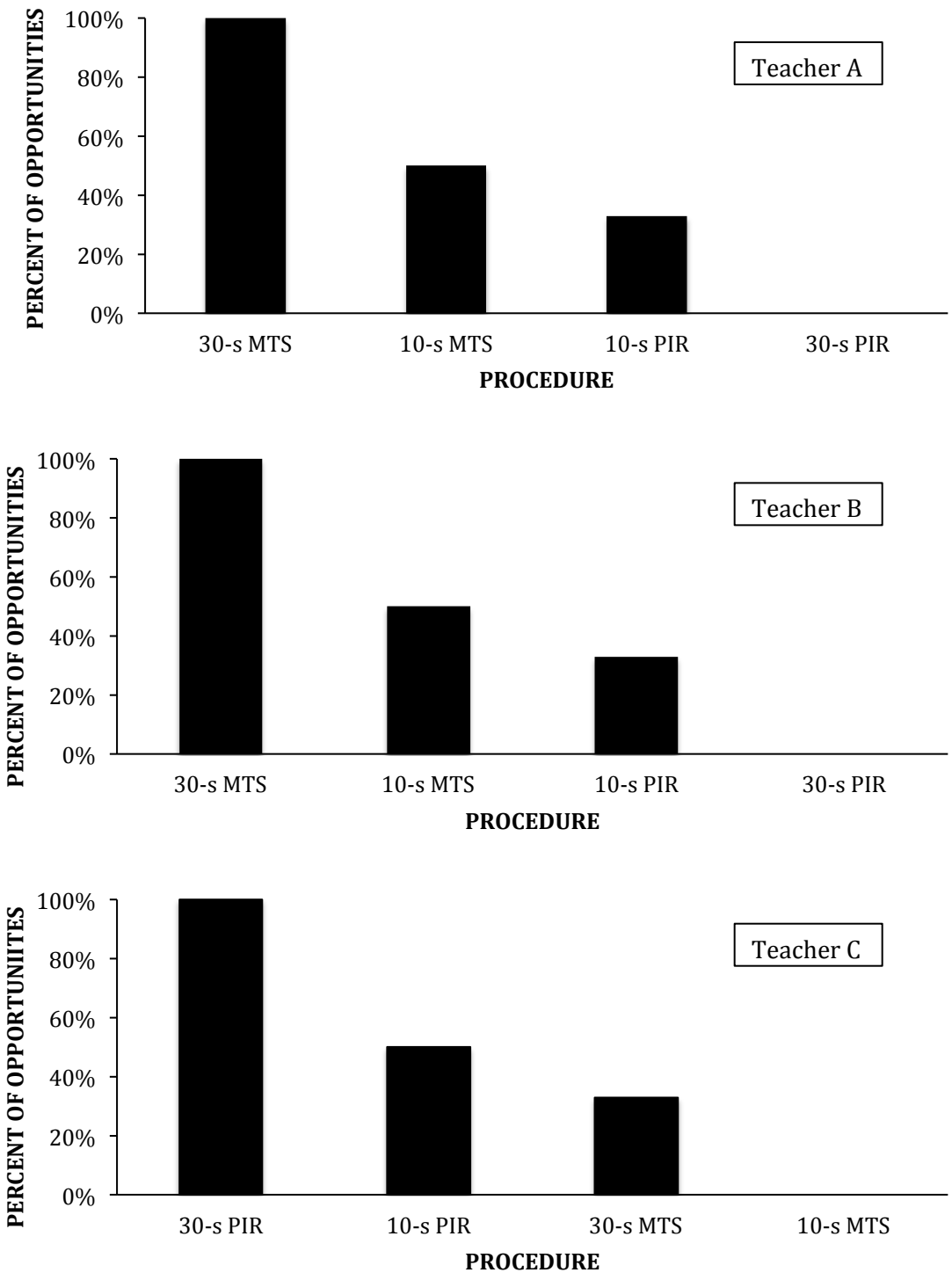

Figure 7. Teacher A (top), Teacher B (middle), and Teacher C's (bottom) preferences for discontinuous data collection procedures. MSWO ranking adjusted to percent of opportunities in each procedure was selected.

Appendix A

Operational Definition for Motor and Vocal Stereotypy (from Ciotti Gardenier et al., 2004)

Stereotypy: Responses that have no apparent function and are not teacher directed.
Examples include:
rocking or swaying of torso, head, or body (full motion down and up or left and right);

- vocalizations that are not recognizable words (in normal conversational tone and volume) and are not in direct response (within 5 s) to teacher request for vocal response (or vocalizations that are recognizable words but are not directed toward another individual (student's body or face is not oriented toward another person));
- hand flapping or other non-functional hand movements;
- non-functional rotation of hand (more than 90 degrees) with or without materials;
- positioning hands in front of face or over ears;
- finger flicking;
- spinning objects;
- addition of objects to a line (2 or more) objects;
- licking, mouthing, or smelling objects, people or surfaces;
- manipulation of objects in a manner not appropriate to materials, not including throwing;
- non-functional closing or squinting of eyes;
- non-contextual laughing or giggling (not in response to interaction with materials or interaction with another person);
- non-functional movement of any or all body parts or objects, including jumping when paired with screaming;
- pressing or rubbing fingers or whole hand against surface or body parts;
- tapping objects.

Non-examples include:

- ''walking'' toys (e.g., cars, stuffed animals, dolls);
- whining—high pitched prolonged vocalization;
- crying;
- screaming, vocalizations above normal conversation level;
- laughing in response to tickling or joke;
- student rocking in one direction and teacher redirecting back;
- movements generated from an unobservable body part, i.e., legs wiggling but view on tape is from waist up;
- smiling that does not produce an audible sound;
- wiping face or mouth;
- incorrect responses to teacher direction (note that this is specific to the direction, e.g., only incorrect motor responses to cues meant to set the occasion for a motor response and incorrect vocal responses to cues meant to set the occasion for a vocal response are considered non-examples);
- approximations of word or request;
- rubbing eyes;
- leaning on forearm or fist;

- tapping anywhere on teacher's body to get attention;
- immediate echolalia: words identical to those spoken by another person.

Appendix B

Operational Definitions of Teacher Responses when Teaching (from Carroll et al., 2013)

- Establish Ready Behavior: Teacher waits to present an instruction until the student does not engage in disruptive movements of the limbs and is oriented toward the teacher (e.g., shoulders facing the teacher) for a minimum of 1s.
- Secure Attention: Teacher requires the student to look (prompted or unprompted) at training materials before presenting the instruction.
- Use Clear Instructions: Teacher presents an instruction that is concise, clearly specifies the target behavior, and does not include unnecessary words.
- Present the Instruction Once: Teacher does not repeat an instruction, with the same or similar wording, in the absence of a controlling prompt following an error or no response from the child (i.e., an error is not scored if the instruction is repeated while the teacher is delivering a controlling prompt).
- Praise contingent on correct responses: Praise is delivered within 5s of a correct unprompted or prompted response.
- Tangible or edible item contingent on correct response: A preferred tangible or edible item is delivered within 5s of a correct unprompted or prompted response.*
- Use Controlling Prompt: A prompt that evokes a correct response is provided within 10s of an instruction following no response or within 3s following an incorrect response.
- Ignores, continues with current demands, or blocks problem behavior: Teacher does not provide verbal or physical attention, minimizes facial expression following problem behavior, and continues with the current trial. If it is necessary to block dangerous behavior, the teacher rearranges the environment or uses the minimum amount of physical interaction necessary to keep the student safe.

*excluded in present study

Appendix C

Considerations for Data Collection Form

How important are the following factors when choosing a data collection procedure?

1. Ease of procedure

1	2	3	4	5
Not important At all				Extremely important

2. Accuracy of procedure

1	2	3	4	5
Not important At all				Extremely important

3. Duration of procedure (i.e., observation length)

1	2	3	4	5
Not important At all				Extremely important

Appendix D

Teacher Perceptions Form (adapted from Fiske & Delmolino, 2008)

1. Overall, how easy was it to collect data using [10-s MTS, 10-s PIR, 30-s MTS, or 30-s PIR] during the [10 or 30]-minute period?

1	2	3	4	5
Not easy at all				Extremely easy

2. How accurately do you think this data collection method reflects the student's overall stereotypy during this [10 or 30]-minute period?

1	2	3	4	5
Not accurate at all				Extremely accurate

Appendix E

Teacher Perceptions Ranking Form

1. Which method did you prefer? Rank them from 1 to 4 (1= most preferred, 4= least preferred).

10-s MTS for 10 minutes _____

10-s PIR for 10 minutes _____

30-s MTS for 30 minutes _____

30-s PIR for 30 minutes _____

Briefly explain why:

2. Which method was easiest? Rank them from 1 to 4 (1= easiest, 4= hardest).

10-s MTS for 10 minutes _____

10-s PIR for 10 minutes _____

30-s MTS for 30 minutes _____

30-s PIR for 30 minutes _____

Briefly explain why:

3. Which method gave the most accurate data for stereotypy? Rank them from 1 to 4 (1= most accurate, 4= least accurate).

10-s MTS for 10 minutes _____

10-s PIR for 10 minutes _____

30-s MTS for 30 minutes _____

30-s PIR for 30 minutes _____

Briefly explain why: