An Investigation of How Variables Impact Accuracy of Data Collection

By

Jessica Riley
M.A., University of Kansas, Lawrence, 2021
B.A., University of Kansas, 2016

Submitted to the graduate degree program in Applied Behavior Analysis and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Master of Arts.

Chair: Thomas Zane, Ph.D., BCBA-D

Edward K. Morris, Ph.D.

Richard Kubina, Ph.D., BCBA-D

Date Defended: 11 March, 2021

The thesis committee for Jessica Riley certifies that this is the approved version of the following thesis:

APA 7th Edition Master's Thesis

_____

Co-Chair: Edward K. Morris, Ph.D.

_____

Co-Chair: Richard Kubina, Ph.D., BCBA-D

Date Approved: 11 March, 2021

Abstract

Assessment of accuracy should be conducted ongoing throughout research and intervention. This study aimed to identify how the number and frequency of behaviors effected measurement accuracy. Participants were students attending a four-year institution in the Mid-West. Two participants were undergraduate students with no prior data collection training and one participant was a graduate student with three years of experience as a Registered Behavior Technician. Participants were trained to identify and collect data on nine target behaviors. Participants watched and recorded target behaviors for six, 10-min. videos. The predetermined behavior and frequency occurrence assigned to each of the six, videos were: 1) three behaviors occurring two times each, 2) three behaviors occurring five times each, 3) six behaviors occurring two times each, 4) six behaviors occurring five times each, 5) nine behaviors occurring two times each, and 6) nine behaviors occurring five times each. Accuracy was identified by comparing participant data to true value observer data agreed upon by two experienced observers. Results indicated the number of behaviors being recorded had no effects on accuracy of data collection. During the initial and maintenance assessments, two participants measured behavior with lower accuracy when measuring conditions with five occurrences than when measuring two occurrences per behavior. One participant measured with lower accuracy when measuring two occurrences during the initial assessment, but measured less accurately when measuring five occurrences during the maintenance assessment. This indicated the frequency of behavioral occurrences may have had an effect measurement accuracy.

Keywords: accuracy, true value, measurement, number of behaviors, frequency

Acknowledgments

I would like to thank Thomas Zane, Ph.D., BCBA-D for guiding me through the process of my graduate program and assisting in the growth and development of my professional knowledge and abilities. I am truly grateful for all that you have taught me.

I would also like to thank Edward K. Morris, Ph.D. and Richard Kubina, Ph.D., BCBA-D for committing to this project regardless of the extended time required to finish it.

Table of Contents

List of Figures

List of Tables

An Investigation of How Variables Impact Accuracy of Data Collection

Experimental and applied behavioral analysis rely on consistently measuring observable behaviors (Dempsey, Iwata, Fritz, & Rolider, 2012; Farkas & Tharp, 1980; Farmer & Nelson-Gray, 1990; Fiske & Delmolino, 2012; 1990; Kostewicz, King, Datchuk, Brennan, & Casey, 2016; LeBlanc, Raetz, Sellers, & Carr, 2016; Lipinsky & Nelson, 1974; Ueyama, 2017) and demonstrating consistently demonstrates controlled changes in behaviors across settings and conditions (Fiske & Delmolino, 2012; Kostewicz et al., 2016; LeBlanc, Lund, Kooken, Lund, & Fisher, 2019). To ensure consistent measurement of observable behavior and demonstration of consistently controlled changes in behaviors, direct observation methods are used as the foundation to the fields' primary methodologies (Fiske & Delmolino, 2012; Kostewicz et al., 2016).

Direct observation is a method in psychology that looks at the actual occurrence of behavior and measuring it directly (LeBlanc et al., 2016; Repp, Nieminen, Olinger, & Brusca, 1988). Observable behaviors are typically measured as they occur in the environment (Farkas & Tharp, 1980; Repp et al., 1988). Direct observation methods are used to record the actual occurrences of behavior (Dempsey et al., 2012; Miltenberger & Weil, 2013). These methods can also be used to identify other environmental variables that affect behaviors (Dempsey et al., 2012; Cooper, Heron, & Heward, 2007b). Direct observation methods have been used to effectively evaluate ongoing behavioral processes and confirm suspected variables affecting behavior (Dempsey et al., 2012; Miltenberger & Weil, 2013; Van Houten, Axelrod, Bailey, Favell, Fox, & Iwata, 1988).

Measurement systems that are used for data collection can be classified as continuous and discontinuous measurement. Continuous measurement systems record every response in an observation interval (LeBlanc et al., 2019; Cummings & Carr, 2009), allowing the observer to see and analyze every instance of behavior across all response opportunities to offer a more complete record of behavior (Cummings & Carr, 2009). Discontinuous measurement systems, in contrast, are not meant to capture every occurrence of the target response. This system collects data samples during an observation by dividing an observation into intervals and scoring behavior based on if the target behavior occurred or not, not an exact count (Fiske & Delmolino, 2012). Direct observation and the systems that are used to collect data may assist observers with defining behavior based on variables affecting behavior in their natural environment (Machado, Luczynski, & Hood, 2019; Miltenberger & Weil, 2013) and providing more meaningful and valid data (LeBlanc et al., 2016). To ensure the efficacy of direct observation methods, some critical dimensions must be discussed: (1) accuracy, (2) true value. (3) treatment integrity, and (4) reliability.

Accuracy

All areas of behavior analysis emphasize on the importance of accuracy (Miltenberger & Weil, 2013), the extent to which the datum collected match the true value of an event given the applicable variables and conditions (Cooper, Heron, & Heward, 2007a; Johnston & Pennypacker, 1980; Kazdin, 1977; Kostewicz et al., 2016; Rapp, Carroll, Stangeland, Swanson, & Higgins, 2011). Accuracy is assessed by comparing the data that are collected by observers to the true value or “gold standard” of an event to ensure that the data being collected represent the target events that actually occurred (Cooper et al., 2007a; Johnston & Pennypacker, 1980; Smith, Lambert, & Moore, 2013; Wu, Whiteside, & Neighbors, 2007).

True Value

The true value is a referred to measurement of the actual occurrence of the target behavior that is less susceptible to the errors that may occur with other measurements used to collect the data of the target behavior in question. To ensure that it is the true value, the researcher must make sure that all potential sources of error were removed in the data collection process (Cooper et al., 2007a; Johnston & Pennypacker, 1980; Kostewicz et al., 2016). The true value is often identified by having multiple experienced observers observe the behavior event in question and agreeing upon the measurement of the behavior (Smith et al., 2013; Wu et al., 2007). Having accurate data assists with establishing validity.

Treatment Integrity

Another concern of accurate data collection is treatment integrity, the degree to which independent variables are being implemented as required (Vollmer, Sloman, & St Peter Pipkin, 2008). Treatment integrity assessments on the implementation of procedures may also include the implementation of data collection procedures. Integrity checks may provide additional information when dealing with complex behaviors that may be affected by multiple variables (Repp et al., 1988). With the implementation of multiple assessments such as accuracy, integrity, and reliability checks, incorrect data collection procedures and incorrect implementation could be caught early on and corrected with feedback and additional training (Repp et al., 1988; Vollmer et al., 2008).

Reliability

Another factor that must be assessed is reliability. Reliability is when a procedure can be consistently replicated under the same conditions and consistently produce the same results (Cooper et al., 2007a; Johnston & Pennypacker, 1980; Kazdin, 1977; Repp et al., 1988).

Reliability may be assessed by comparing data collected by multiple observers observing the same behavioral events, also known as interobserver agreement (Mash & McElwee, 1974; Kazdin 1977). If observers recorded similar values for the same behavioral events that meet a specific criterion (typically around 80% agreement or greater), then the results may suggest that agreement is high (Kazdin, 1977).

There are many different formulas used to assess reliability. Each formula may be affected by behavior in different ways (Repp et al., 1988). Some examples of these formulas include, but are not limited to, total count, total duration, interval-by-interval, scored-interval, and unscored-interval agreement.

Total Count

Total count interobserver agreement reliability is calculated by dividing the smaller frequency count of one observer by the larger frequency count of another observer, and then multiplying by 100 (Cooper et al., 2007a; Lipinsky & Nelson, 1974, Rolider, Iwata, & Bullock, 2012). This form of interobserver agreement is considered to be the most simplistic formula used for event recording (Cooper et al., 2007a). This formula is typically used for behavior that occurs at lower rates (Lipinsky & Nelson, 1974). However, high agreement does not mean that the observers were recording the same instances of behavior (Cooper et al., 2007a, Rolider et al., 2012). Often times this formula may overestimate how many times behavior was actually agreed upon.

Total Duration

Total duration agreement is calculated by taking the shorter total duration of an observer, dividing it by the larger total duration, and then multiplying by 100 (Cooper et al., 2007a;

Lipinsky & Nelson, 1974). However, this formula may not ensure that observers recorded the same durations for the same occurrences of behavior and overestimates agreement.

Interval-by-Interval

Interval-by-interval agreement is calculated by dividing the number of intervals agreed upon by the total number of intervals, and multiplying by 100. This formula often overestimates actual agreement for low and high rates of behaviors due to random or accidental agreement scores on intervals (Cooper et al., 2007a, Rolider et al., 2012).

Scored-Interval

The scored-interval formula is calculated by dividing the number of intervals of which behavior occurred that were in agreement, by the total number of intervals in which one or both observers scored the occurrence of behavior, and multiplying by 100. This method is best used for low rates of behavior and tends to overestimate high rates of behaviors (Cooper et al., 2007a; Repp et al., 1988).

Unscored Interval

On the contrary, the unscored-interval agreement formula is calculated by dividing the number of intervals of which both observers agreed that behavior did not occur divided by the total number of intervals in which one or both of the observers recorded the nonoccurrence of behavior, and multiplying by 100. This has been considered a more stringent assessment for high rates of behavior, but it overinflates low rates of behavior (Cooper et al., 2007a; Repp et al., 1988).

Exact-Agreement

Exact-agreement is calculated dividing the number of intervals in which both observers recorded the same number of occurrences of behaviors in an interval and dividing by the total

number of intervals. This is the one of the most conservative and stringent reliability measurements. If there is any disagreement between observers during an interval, then it is counted as a complete disagreement for that interval (Rolider et al., 2012).

Assessment of Behavior Measurement

Although assessment of accuracy should be conducted ongoing throughout research and intervention, reliability measurements using interobserver agreement are often used in replacement of accuracy measurements (Cooper et al., 2007a; Kazdin, 1977; Kostewicz et al., 2016) due to the mistaken assumption that accuracy and reliability assessments are the same (Kostewicz et al., 2016). Agreement is not the same as accuracy (Kazdin, 1977; Repp et al., 1988; Vollmer et al., 2008). An example of when agreement can be high, but accuracy is low is when observers independently record behaviors that are in agreement with other observers results, but all of the observers are incorrectly collecting data, therefore making the data inaccurate regardless of agreement (Dorsey, Nelson, & Hayes 1986; Lipinsky & Nelson, 1974; Repp et al., 1988; Smith et al., 2013; Vollmer et al., 2008; Wu et al., 2007). On the contrary, interobserver agreement could be low due to one observer accurately recording data and another observer inaccurately collecting data (Haaf, Brewster, de Saint Victor, & Smith, 1989; Kazdin, 1977). Another instance when agreement may be inaccurately high or low is when observers record the same instance of behavior but at slightly different times. An example of this occurrence would be if one observer records the target behavior at the end of an interval and another observer records the same occurrence of behavior at the beginning of the next interval (Vollmer et al., 2008). These examples display how the use of reliability assessments cannot be used in place of accuracy assessments to display the accuracy of data collection.

The difference between accuracy assessments and interobserver agreement is the extent to which the data (true value or other observer) that is compared to the observer's data actually reflects that actual occurrence of behavioral events (Kazdin, 1977). Literature suggested that accuracy should be assessed by identifying predetermined, true standards and comparing observer data to this true standard (Johnson & Pennypacker, 1980; Kazdin, 1977; Smith et al., 2013). There are multiple suggested methods to identify the true standard. These methods included selecting experienced observers to agree upon a set number of behaviors that have occurred and by creating videos and/or other materials that followed a precise script that only displayed a predetermined set of behavioral occurrences (Kazdin, 1977; Smith et al., 2013). For example, Smith and colleagues (2013) discussed literature that outlined the establishment of the "gold standard" of measuring accuracy. The gold standard method was a method that established what the true value or the actual measurement of the occurrence of behavior was. This method has been done by using electromechanical recording, using scripts acted out by confederates, repetitive viewing of records until multiple observers agree on a true value of behavior, and/or using outside expert observers to review records and provide the true value.

Threats to Accuracy and Reliability

There are many threats that could negatively impact the accuracy and reliability of data collection. Vollmer and colleagues (2008) suggested that the threats to reliability could be considered as either errors of omission (when the observer fails to record the occurrence of an event) or errors of commission (when the observer records a behavioral event that did not occur or records one behavioral event as another behavioral event).

Mash and McElwee (1974) suggested that accuracy was a function of three factors. These three factors included recording procedures (the number of behaviors being recorded, definitions,

device used for recording, and measurement procedures), observer characteristics (age, sex, and prior experience) and the conditions of observation (number of subjects, number of other behaviors occurring, frequency and rate of target behavior, and temporal sequencing of behavior). Cooper and colleagues (2007a), Farkas and Tharp, (1980), Kazdin, (1977), Kostewicz and colleagues (2016), Lipinski and Nelson, (1974), Smith and colleagues (2013), and Vollmer and colleagues, (2008) have elaborated on many other threats in addition to those discussed by Mash and McElwee (1974). Some of the other factors that the other literature expanded on included other observer variables (such as observer reactivity and observer drift), observer training variables (such as observer expectancies, feedback, materials used for training, and complexity of training), and other observation conditions (such devices used for measurement recording, interval duration length, and session duration length).

Definitions of Behavior

Accuracy and reliability may be affected when observers count other events that do not fit the definition of behavior or withhold counting events that do fit the definition of the behavior. This may occur when observers drift from the definition and record behaviors that do not fit the predetermined description of the behavior (discussed in more depth later when discussing observer drift). However, other causes of these phenomena could potentially be due to vague behavior definitions and other issues with definition topographies (Kazdin, 1977; Johnston & Pennypacker, 1980; Smith et al., 2013).

Smith and colleagues (2013) assessed definition topographies, assessing how different characteristics of target behavior descriptions affected the accuracy and reliability of data collection. For the first trial, eighteen undergraduate students observed a 14-minute video and recorded self-injurious behavior that met criteria for a definition that included subjective words

such as “forceful swing” and “hit”. For the second trial, the students observed the same 14-minute video and recorded self-injurious behavior that met criteria for a definition that avoided subjective words and words that could not be physically observed. The students’ data were then compared to a permanent record of when data occurred that was agreed upon by two experienced observers. The results indicated that the second definition, that used specific words lacking in subjectivity and only had words that were physically observable, produced more accurate data collection, and the behavior of the observers was affected by the characteristics of the definition.

Kazdin (1977) recommended that team trainings should continuously occur throughout treatment to ensure that all individuals delivering services agree on the definitions and standards. Although this could potentially be expensive and time consuming, it may help maintain high levels of accurate data collection, reliability, and treatment integrity.

Observation Conditions

Literature has indicated that previous training on identification and recording of behaviors that occurred in specific behavior sequencing compared to behavioral events that occurred randomly, may affect the accuracy of behavioral identification and data collection, and reliability (Kazdin, 1977). For example, Mash and McElwee (1974) investigated how behavior predictability and history of observation of predictable behaviors affected accuracy. They did not find a difference in accuracy between more and less predictable behavioral events; however, they found a correlation between the observer’s history of recording behavioral events that were more predictable compared to less predictable. Their results suggested that those who have a history with recording behaviors that are less predictable engaged in recording behaviors with higher accuracy.

Temporal sequencing has also been investigated by addressing the rate of responding and response distribution characteristics that effect reliability scores. For example, Rolider and colleagues (2012) investigated how three different characteristics of high-rate responding influenced different methods of reliability scoring. The three characteristics investigated were increases in response rate, irregular inter-response times that create response bursts, and high rates of behavior that occur at the end of intervals. The methods of reliability scoring that were tested were total agreement, interval agreement, proportional agreement, and exact agreement. The results indicated that the rate of responding and response bursts did not affect the results of agreement scores across all four methods. The results also suggested that high rates of responding that occurred at the end of the intervals did have significant effects on agreement scores when using interval, proportional, and exact agreement methods of reliability. However, these results may have been due to arbitrary rates of responding being chosen for the session. Kazdin (1977) suggested that stimuli used in training should vary in predictability of behavioral events to ensure that observers gain the ability to accurately record behavioral events as they occur and can adjust to changes in the complexity of behavioral events, not based off of prediction of behavioral sequencing.

Measurement Systems

Another threat to accuracy that has been frequently investigated was the type of measurement system used to measure behavior. Research indicated that continuous measurement procedures were more accurate (Fiske & Delmolino, 2012; Giunta-Fede, Reeve, De Bar, Vladescu, & Reeve, 2016; Johnston & Pennypacker, 1980). However, continuous measurement procedures may be more difficult to record if there is not a clear beginning and end to the behavior. In addition, it may require more of the observer's attention if there are multiple

behaviors that need to be recorded, or if other responses need to be recorded during other activities (Cummings & Carr, 2009; Fiske & Delmolino, 2012; Johnston & Pennypacker, 1980). This could lead to a reduction in accurate data collection due to unobserved events when the observer's attention is focused on recording other behaviors or responses, or when the observer is recording high-rates of behavior and cannot record the data quick enough (Kazdin, 1977; Madsen, Peck, & Valdovinos 2016).

In an effort to address these issues, many behavior analysts may use discontinuous measurement procedures such as partial interval recording, whole interval recording, and momentary time-sampling (Cummings & Carr, 2009; Fiske & Delmolino, 2012). Each of the discontinuous measurement systems have certain conditions under which they are better to be used to accurately capture the occurrence of behavior (LeBlanc et al., 2019). Partial interval procedures are typically best for use with behaviors that occur more frequently, but these procedures may overestimate the occurrence of behavior and potentially reduce the accuracy (Fiske & Delimolino, 2012; Repp et al., 1988). Whole interval procedures may be best when behaviors occur less frequently, but may reduce accuracy by underestimating the occurrence of behavior. They also may be affected by the duration of intervals (Fiske & Delimolino, 2012; Repp, Roberts, Slack, Repp, & Berkler, 1976; Repp et al., 1988). Momentary time-sampling procedures may be best if it is not necessary to decrease behaviors to levels of zero, but they should not be used if behaviors infrequently occur or if behaviors are short in duration (Fiske & Delmolino, 2012). If these measurement systems are not used under the specified conditions, they may not provide accurate data results (Fiske & Delimolino, 2012).

Many researchers have assessed how accurate different data collection procedures were and whether or not they should be used given certain response variables and complexity of

observational systems. For example, Repp and colleagues (1976) compared time-sampling, interval recording, and frequency measurement procedures when used to measure behaviors of different rates of responding (high, medium, or low) and different patterns of responding (constant or bursts). Observers recorded the occurrence of pen deflections created by electromechanical equipment using each of the measurement systems under high, medium, and low rates of responding. To identify participants' accuracy of data collection, the participants' data were compared to the permanent product created by the electromechanical equipment. The results indicated that time-sampling methods did not accurately represent environmental events of behavior, and that interval recording did accurately record low and moderate rates of responding, but did not accurately represent high rates of responding. Based off of these results, Repp and colleagues (1976) suggested that other studies that have used time-sampling methods may be questionable due to potential misrepresentation of environmental events of behaviors. Repp and colleagues (1976) also suggested that studies which have used short 10-second intervals for interval recording in baseline measurements for high-rate behaviors may have greater differences between baseline and intervention measurements than what was reported due to a lack of instances of high-rate behaviors occurring within the short interval.

Other areas in which measurement procedures have been assessed is the effectiveness of representing actual occurrences of behaviors when teaching new behaviors. Giunta-Fede and colleagues (2016) used an adapted alternating treatments design to compare continuous and discontinuous measurement. The three conditions that were tested included a continuous count measurement, first trial probe data, and probe data taken every 5th session. They assessed how measurement procedures effectively represented the actual occurrence of behavior by analyzing how long it took individuals to master target skills and how well the skills maintained when

using each measurement procedure. The results indicated that there were minimal differences between measurement procedures and their effects representing the occurrence of target behaviors during teaching procedures, but continuous measurement still was more sensitive to changes in behavior and was most conservative.

LeBlanc and colleagues (2016), acknowledged the need for a standardized decision-making model for selection of measurement procedures. LeBlanc and colleagues recognized how not every measurement procedure was appropriate for all situations due to factors such as behavior frequency, discrete occurrences, etc. They included measurement procedures such as event recording, duration, latency, intensity, permanent product, partial interval recording, and momentary time sampling. They formulated questions that may lead practitioners to the appropriate (i.e., most accurate) measurement procedures based on the strengths and limitations of each method. These questions were based off of the different strengths and weaknesses of variables that effect each measurement system. The questions addressed the behavior in regards to the behavior's topography, the ability of observers to record the behavior, the resources available to record the behavior, and other potential variables that could affect the integrity of measurement systems. Although this standardized decision-making model was not tested in this article, it may be used as a guiding tool to assist practitioners in selecting a measurement procedure that may potentially record more accurate data.

Recommendations addressing measurement procedures that impact accuracy of data collection suggest that an observer should understand what the effects of each recording procedure are and how they impact accuracy of data collection. For example, partial-interval methods may overestimate the occurrence of behavior, where whole interval methods may underestimate the occurrence of behavior (Repp et al., 1988). Momentary time sampling

methods on average produce more accurate occurrences of behavior (Repp et al., 1988). If using a measurement procedure that uses interval, smaller observation intervals should be used as they may produce more accurate data than larger intervals (Repp et al., 1988).

Data Recording Materials

Some literature indicates that current technology may assist with increasing accuracy and reliability while also reducing the time spent collecting data. For example, Tarbox, Wilke, Findel-Pyles, Bergstorm, and Granpeesheh (2010) used an alternating treatment design to compare the accuracy and time spent collecting data using a program called mTrial with the traditional method of pen-and-paper data. Tarbox and colleagues used three dependent variables that consisted of measuring the percentage of correct data collected by the therapist, the duration, in seconds, of time spent taking data per session, and duration (seconds) of time spent graphing the data on a line graph per session. They also had four independent variables consisting of collection of data for DTT programming using the pen-and-paper method, collection of data for DTT programming using mTrial, graphing using pen and paper, and graphing using mTrial. Participants consisted of four children between the ages of three and five that were diagnosed with autism and clients received home-based behavioral services from a local provider. Sessions also occurred in the participants' homes by their regular therapists (who were taught to use mTrial in a one-on-one, 1-hour long training session prior to the start of the study). Programs chosen for sessions consisted of three programs that were chosen out of each participant's individual BIP that fit the criteria of a minimum of 10 trials per session and was taught using DTT. The results of this study indicated that pen-and-paper methods were more accurate and took less of the therapists' time to collect data during session. However, the results also indicated that pen-and-paper methods were more time consuming when graphing. One limitation to this

study was the lack of data collected using mTrial. Less than 10 data points were collected for most sessions.

There have been many overviews and detailed task analyses on how to customize data collection sheets for discrete trial training (Dixon, 2003) and functional analyses (Jackson & Dixon, 2007) using programs such as the Pocket PC or "Xcode". However, there are no other empirical studies evaluating the impact of the use of technology for data collection on accuracy of measurement.

Observer Training

A lack of or incomplete observer training is also a potential threat to accuracy. If an observer does not know how to use the data collection sheet, has a lack of awareness about correct behavior definitions or environmental factors that affect behaviors, then they may not collect accurate data or record data with high reliability compared to those who are familiar and well trained on these variables (Kazdin, 1977; Madsen, et al., 2016; Repp et al., 1988; Vollmer et al., 2008). If data collection systems are too complex, observers also may need more extensive training to ensure that accuracy of data collected are not compromised due to a lack of knowledge or skills, or due to other distractors that may prevent the provider's ability to attend to all target behaviors, immediately record behaviors, and/or to complete other job responsibilities (Kazdin, 1977; Madsen et al., 2016).

Dempsey and colleagues (2012) assessed how training impacted data collection agreement and accuracy when training observers using video and using in-vivo sessions. Participants enrolled in an undergraduate laboratory course across three semesters were assigned to either receive in vivo training or video training each semester. All participants received group instruction on behavior identification and data collection procedures. Those who received in-vivo

training took data with an experienced observer on live 10-minute sessions until they could record data with 90% agreement across three sessions and two clients. Those who received video training recorded behaviors for six, scripted video segments that increased with complexity with each video segment. To set the true value for the videos used throughout this study, two experienced observers collected data of each behavioral occurrence with 100% agreement on each video. During video training, Participants' data were compared to the true value and had to record behavior with 90% agreement across all six videos before moving on. Dempsey and colleagues then delivered a post-test in which all participants watched multiple videos and compared the participants' data to the data collected by two experienced observers whom scored 100% agreement with each other. Dempsey and colleagues found that in-vivo and video training both resulted in high agreement, but those who received both trainings continued to have improved agreement scores when measuring behaviors during in-vivo sessions 1-month after the post-training test. They also suggested that in-vivo training required less preparation than video training, but video training may be more controlled and may be more useful when training on a variety of situations or different complexities of behavior.

Dempsey and colleagues described their results in terms of agreement with the experienced observers, but once, during the discussion of their results, referred to the agreement scores indicating a reflection of accuracy. Prior to the study, they evaluated the experienced observers' ability to consistently meet a 90% agreement score when measuring behaviors in-vivo. Experienced observers were also required to record data with 100% agreement for each of the videos used in the study. The experienced observers meeting 100% agreement for the videos could meet criteria for a true value score for accuracy measurements. However, when comparing participants' data to those of one of the experienced observers for in-vivo sessions, it may be

difficult to indicate these scores as accuracy measurements given the potential environmental events (described in depth below: observer drift, observer reactivity, etc.) that could have affected the experienced observers' ability to accurately score data. Perhaps if other procedures were used, then they could have described the results in terms of accuracy instead of agreement. Suggestions for procedural changes include: (1) experienced observers collecting data on video recordings of the in-vivo sessions (2) frequent agreement checks between experienced observers throughout the study to ensure that drift and other variables (described below) have not occurred (Kazdin, 1977). Reis, Wine, and Brutzman (2013) assessed the effectiveness of training using videos and immediate performance feedback procedures after recording data. Participants scored three videos varying in length in random order. Their data were then compared to the true value which was identified by the agreement of experienced observers. Reis and colleagues found that the simulated session videos, task clarification, and feedback procedures successfully increased data recording accuracy among direct-care staff for adults in a residential facility. High accuracy of data collection also maintained across time after the removal of the training procedure

Jerome, Kaplan, and Sturmey (2014) investigated in-service training with performance feedback. They used a multiple-baseline across participants design to display the effects of receiving instructions, in-service training, in-service training with feedback. Direct-care staff, of adults with intellectual disabilities, collected data under three different conditions: instructions on collection of behavior data, in-service that elaborated on definitions and importance of accurate recording training, and in-service training with feedback on data collection performance. To assess accuracy, Jerome and colleagues used experienced observers, who consistently recorded data with 90% agreement prior to the study, to record data during in-vivo sessions as their true value data. The data collected by direct-care staff was compared to the data

collected by the two experienced observers after each session. The results indicated that in-service training alone was a good method to increase accuracy of data collection. However, when in-service was combined with performance feedback, accuracy of data collection further increased. Like Dempsey and colleagues (2012), Jerome and colleagues (2014) could have implemented additional procedures (as described previously) to ensure the experienced observers maintained accurate data recording throughout the study.

Although continuous feedback has been shown to be an effective training component, if not implemented correctly, feedback could potentially have some unwanted effects on the accuracy and reliability of data collection. Feedback that is given to observers can change the expectancies that an observer may have in regards to a target behavior (Kazdin, 1977; Lipinski & Nelson, 1974; O'Leary, Kent & Kanowitz, 1975). This could decrease accuracy of data collection and could decrease reliability when compared to observers with effective training. For example, O'Leary and colleagues (1975) suggested that expectancies, in combination with other feedback, could change observer behavior. In their study, O'Leary and colleagues showed observers video tapes that had the same amount of disruptive behavior during baseline as it did during the treatment condition. During baseline, instructions that the behavior would decrease were given, but no feedback was given during behavior recording. During the treatment phase, instructions that the behavior would decrease were given, positive comments were delivered when the observers recorded reductions in behavior, and negative comments were provided when there was an increase in behavior or a lack of behavior change. The participants' data were compared to three criterion observers, who had over a year of experience recording the target behavior and had scored with 80% interobserver agreement for the occurrence of behavior for the videos used. O'Leary and colleagues found that expectancy alone in baseline was not enough

to change observer behavior, and participants recorded behavior with higher agreement with the actual occurrence of the event. However, when expectancies and feedback were both delivered, the observer behavior changed to recording data that depicted the expectancy of behavior change more than depicting the actual occurrence of the event.

To increase data collection accuracy and reliability, literature has suggested training observers systematically (Repp et al., 1988). For example, Madsen and colleagues (2016) recommended training staff by providing performance feedback and written descriptions of behaviors. Due to successful applications and replications in research, Madsen and colleagues also recommended using training procedures such as behavior skills training, modeling, roleplaying, video modeling, and combinations of video training and in-vivo training. Kazdin (1977) and Vollmer and colleagues (2008) also recommended using multiple stimuli across multiple environments during the training process to increase generalization to more natural settings. Kazdin (1977) recommended new observers should occasionally be brought in for behavior data collection purposes, and that videos of sessions should be taken and scored in random order. Kazdin (1977) and Vollmer and colleagues (2008) both recommended that feedback should only be in the form of positive or corrective feedback for accuracy of data collection and treatment implementation. Repp and colleagues (1988) suggested that extensive observer practice and training should be ongoing.

Observer Reactivity

Other literature has indicated that observers who are informed (or assumed) that reliability checks were being conducted were more likely to record target behaviors with higher accuracy than when they were not informed. (Lipinski & Nelson, 1974; Repp et al., 1988). Repp and colleagues (1988) also suggested that other observers may even record data with lower

accuracy when being observed. Reactivity to reliability and accuracy checks could affect the extent to which a treatment or study could be replicated or generalized (Lipinsky & Nelson, 1974; Repp et al., 1988)

When reliability assessments are being conducted, Kazdin (1977), Repp and colleagues (1988), and Vollmer and colleagues (2008) recommended that conditions should be standardized when the assessments are not being conducted in order to capture the true reliability or accuracy of session data. Repp and colleagues (1988) suggested that an adaptation period should be allotted for both the subject and the observer. This could allow the subject to adapt to the setting and the presence of the additional observer, and for the observer to have time to engage in data collection behaviors as they adapt to the presence of the other observer prior to assessment. To allow for the observer to adapt to other observers within sessions, Repp and colleagues (1988) recommended making observations systematic and frequent. Unobtrusive observations may be conducted (Kazdin, 1977; Repp et al., 1988). Repp and colleagues (1988) suggested using a contract that describes the terms of observation that gives permission to make the observer blind to observations schedules in order to reduce reactivity and increase the level of reliability and accuracy assessments. Other recommendations included the use of other trained observers to collect data on the same session via one-way mirrors and or videos of the session. The use of multiple observers may help ensure that observers do not learn idiosyncratic behaviors of other observers (Kazdin, 1977).

Observer Drift

Observer drift is when observers change how they record data of behavioral events over time, even when the definitions do not change (Kazdin, 1977; Repp et al., 1988). The occurrence of observer drift may make it difficult to compare data across observers and/or conditions to

compare the effectiveness of the procedures due to the recorded responses no longer being the same response (Repp et al., 1988). This phenomenon could occur in one observer or across multiple observers. When this occurs across observers, interobserver may remain high, but accuracy could be reduced. Observers who collect behavior data under different conditions may record behavior differently due to stimulus control of the setting or experimental condition. This may result in inaccurate data collection and low observer agreement (Kazdin, 1977).

The literature recommended that ongoing training should occur with all observers from all conditions as a group. This would allow for the observers to discuss definitions, and to practice data collection procedures (Kazdin, 1977). Other recommendations also included videotaping sessions and recording data on them in random order after sessions and bringing in other newly trained observers to assess accuracy (Kazdin, 1977). Vollmer and colleagues (2008) and Repp and colleagues (1988) also suggested that ongoing training and continuous monitoring of observers would assist with the reduction of observer drift.

Complexity of the Behavior Code

In addition, the complexity of the behavior code has also been suggested to be a threat to accuracy (Cooper et al., 2007a; Dorsey et al., 1986; Mash & McElwee, 1974; Murphy & Harrop, 1994; Repp et al., 1988; Ueyama, 2017; Vollmer et al., 2008). The complexity of the behavior code can include many different variables such as behavior programs that require multiple steps, or a requirement to collect data on many different behaviors or environmental variables (Kazdin, 1977; Vollmer et al., 2008). However, the number of behaviors being recorded and their influence on accuracy of data collection are important variables that have had limited investigation, but have been stated multiple times throughout literature (Ueyama, 2017). Cooper and colleagues (2007a) did not offer any references to support the statement. Kazdin (1977)

made this claim and only presented one study as support of this claim. Although there have been many claims that the number behaviors affect accurate data collection (Cooper et al, 2007a; Kazdin, 1977; Vollmer et al., 2008) the limited amount of research shows how empirical evidence may still be needed to display those effects.

Research that has systematically tested these variables also tested for other variables such as the number of behaviors and the previous experience of the observer. For example, Mash and McElwee (1974) conducted a study with 48 participants, all varying in age and background, to assess how the number of behaviors and experience of the observers effected the accuracy of data collection. There were eight conditions that were tested: participants that recorded four or eight different behaviors that occurred in predictable behavior sequences, recorded four or eight behaviors that occurred in unpredictable behavior sequences, participants that were trained to record four or eight behaviors of predictable sequencing recorded the same number of behaviors but with unpredictable behavior sequencing, and participants that were trained to record four or eight behaviors of unpredictable sequencing recorded the same number of behaviors, but with predictable behavior sequencing. Data collected by participants was compared to a true value that was pre-written in a script and recorded in audio-tapes prior to conducting the study. Results indicated that those who recorded four behaviors had higher accuracy scores across all five trials than those who recorded eight behaviors. Results also indicated that participants who had prior training on predictable behaviors reported less accurate measures in relation to their own performance when measuring behaviors that were less predictable.

Dorsey and colleagues (1986) assessed how the number of behaviors and the frequency in which they occurred effected thirty-six undergraduate students enrolled in an introductory level psychology course. Dorsey and colleagues used three, 15-minute videos, with a female and

male confederate, to study the accuracy of nine different behaviors under nine different conditions with different frequencies of behavior. Dorsey and colleagues also assessed how accuracy measurements and interobserver agreement measurements differed when assessing the occurrence of behaviors under these different conditions. The nine different conditions included recording three target behaviors with a set number of occurrences (3, 15, and 33) within the session, recording six target behaviors with a set number of occurrences (3, 15, and 33) within the session, and recording nine target behaviors with a set number of occurrences (3, 15, and 33) within the session. The true value measurement, in which participants' data were compared to, was a prewritten script that was acted out by confederates using video recording. Dorsey and colleagues did not describe integrity checks conducted to ensure that the pre-planned behaviors only occurred as written in the script. The results indicated that accuracy scores were lower when the number of behaviors being measured was higher, and the accuracy scores were higher when the number of behaviors was lower. The results also indicated that the relationship between accuracy and agreement was not affected by the number of behaviors being recorded, but instead was affected by the frequency of occurrences within a session. However, they also reported that the two different methods used to calculate the results had different results for the effects that frequency of behavior had on accuracy and interobserver agreement. The methods used to calculate the results were the Kappa method which measured variance of accurate scoring, percent of agreement across all intervals, and a scored-interval method. When using kappa variance methods, the results indicated that both accuracy and interobserver agreement were higher when the frequency of target behaviors being observed was lower. This result was in agreement of the results found in the study conducted by Mash and McElwee (1974).

Farmer and Nelson-Gray (1990) used a “mixed factorial” design to assess how accurately individuals took data using both estimation and counting procedures. The study consisted of eighteen different independent variables that included recording procedures using estimation under nine conditions: (1) recording one target behavior with four occurrences per behavior, (2) recording one target behavior with twelve occurrences per behavior, (3) recording one target behavior with twenty occurrences per behavior, (4) recording two target behaviors with four occurrences per behavior, (5) recording two target behaviors with twelve occurrences per behavior, (6) recording two target behaviors with twenty occurrences per behavior, (7) recording three target behaviors with four occurrences per behavior, (8) recording three target behaviors with twelve occurrences per behavior, (9) recording three target behaviors with twenty occurrences per behavior. Recording procedures using counting were placed under the same conditions, making a total of eighteen conditions. The participants, who were not provided prior training by the experimenters, viewed a video for each condition of one confederate on talking on the phone. Participants’ data were compared to a pre-determined frequency of occurrences. The results indicated that estimation procedures were less accurate than counting, more repetitions within a session were correlated with a higher degree of measurement accuracy than those with fewer within-session repetitions, and that there were not any identifiable correlations between the number of behaviors recorded (up to three behaviors) and the accuracy of data collection (Farmer & Nelson-Gray, 1990). These results indicated that three target behaviors may be an acceptable number of behaviors to record without jeopardizing accuracy of data collection.

Murphy and Harrop (1994) investigated how the number of behaviors being recorded, partial interval recording procedures, and momentary-time sampling procedures effected accuracy of data collection of studying behaviors. The six independent variables included in this

study were partial interval recording for one behavior during session, partial interval recording for two behaviors during session, partial interval recording for three behaviors during session, momentary time sampling recording for one behavior during session, momentary time sampling recording for two behaviors, and momentary time sampling recording for three behaviors. Participants' results were compared to a pre-determined criterion written into the video scripts. Murphy and Harrop found that momentary-time sampling procedures were more stringent when recording behavior and recorded measurement with higher accuracy that partial interval recording procedures. Results indicated that the recording procedures did have an effect on the accuracy of measurement. However, the number of behaviors being recorded did not affect the accuracy of data collection.

La France, Heisel, and Beatty (2007) investigated how session duration and number of behaviors being observed effected the accuracy of data collection on nonverbal cues. La France and colleagues randomly assigned 112 undergraduate students from different communication courses to four different independent variables that included: (1) recording two behaviors during a 2-minute session, (2) recording two behaviors during a 10-minute session, (3) recording eight behaviors during a 2-minute session, and (4) recording eight behaviors during a 10-minutesession. Participants' data were compared to a pre-determined criterion of behavioral occurrences. The study results replicated the previous findings and indicated that increases in the number of behaviors being recorded decreased accuracy of behavior data collection, and that session duration did not have an effect on the accuracy of data collection.

Social Significance

Assessing accuracy and reliability informs if target behaviors are correctly being measured across settings, observers, subjects, and stimuli. By ensuring the accuracy of data collection,

behavior analysts may make data-based decisions when determining intervention success or needed changes for interventions (Jerome et al., 2014; Johnston & Pennypacker, 1980; Vollmer et al, 2008). Literature indicated that accurate data collection was an important factor due to the decisions that are based off of the data collection (Haaf et al., 1989; Johnston & Pennypacker, 1980; Vollmer et al, 2008). If agreement data were used in replacement of accuracy, it is possible that discrepancies in behavior identification and recording may not be identifiable and decisions for behavior change could be inaccurate and unsuccessful (LeBlanc et al., 2016; Repp et al., 1988).

Van Houten and colleagues (1988) suggested that it is the right of the clients and the ethical obligation of behavior analysts to provide clients with effective and ongoing behavioral assessment using these methods and to make decisions for scientifically validated treatments based off of the data collected through the use of these methods. The decisions outlined in research that are often based off of data collected included the use of restrictive or non-restrictive treatment procedures, the type of staffing needed, and potential placement of the participant, etc. (Vollmer et al, 2008). However, in order to make appropriate data-based decisions for variable changes such as treatments, it is important that the data collected are accurate (Haaf et al., 1989; Johnston & Pennypacker, 1980; Vollmer et al, 2008). Inaccurate data could potentially lead to dangerous decisions and outcomes.

These socially significant issues and concerns regarding decision-based data are applicable to many different subjects in multiple different areas. Behavior analytic methods and procedures have been emphasized and applied across multiple areas such as the field of education (Martens, Daly, & Ardoin, 2015), obesity (Hausman & Kahng, 2015), gambling (Daar & Dixon, 2015), drug addiction (Dallery, Defulio, & Meredith, 2015), health and fitness

(Normand, Dallery, & Ong, 2015) and organizational behavior management (Sigurdson & McGee, 2015) among many other fields.

In addition to those fields, behavior analytic methods have been well known for being applied to treating individuals diagnosed with intellectual disabilities. Problem behaviors are common in those who are diagnosed with intellectual disabilities (Jang, Dixon, Tarbox, & Granpeesheh, 2011). The more severe the intellectual disability is, the more likely an individual is to engage in a variety of challenging behaviors from tantrums, self-injury, and aggression to stereotypy (Jang et al., 2011). These behaviors and much that they entail, such as assessment of functions and a variety of interventions, have become frequently documented throughout research (Madsen et al., 2016). However, during assessments of behavior, multiple behaviors are often identified for behavior change (Cooper et al., 2007b).

When multiple behaviors are identified for behavior change, behavior analysts have to prioritize the behaviors which they target. This may be essential given that majority of a provider's attention may be required to monitor the occurrence of problem behaviors and record behavior occurrence (Lipinski & Nelson, 1974) and that the number of behaviors being observed could affect the accuracy of data collection (Cooper et al., 2007a).

To prioritize which target behaviors should be targeted for reduction, Cooper and colleagues (2007b) suggested nine different questions that should be asked in the decision-making progress. These questions address if the behavior or lack of behavior posed any danger, how often the behavior occurred or would occur once learned, how long the behavior or lack of behavior had been a problem, if the behavior would produce more reinforcement, the importance of the reduction or increase of behavior to future skill development and functioning, what reinforcement would be produced for others, and the success and cost of changing the behavior

(Cooper et al., 2007b). Cooper and colleagues (2007a) stated that the number of behaviors being observed effected the accuracy and reliability of data collection, however, the nine questions that were suggested by Cooper and colleagues (2007b) failed to give information in regards to how many behaviors should be targeted at one time.

Bosch and Fuqua (2001) also suggested a need for a systematic method for prioritizing problem behaviors. Bosch and Fuqua suggested that the criteria be based on access to new variables (such as reinforcers, contingencies, and environments), the social validity, generalization, ability to compete with inappropriate behaviors, and how other individuals were affected by the behavior change. Other suggestions for prioritizing behavior also included considering the number of the criteria that were met. However, like Cooper and colleagues (2007b), Bosch and Fuqua (2001) failed to identify an appropriate number of behaviors to target.

Although research has suggested that choosing target behaviors should be done systematically, there is still a common lack of consideration to how many behaviors should be targeted for behavior change. However, multiple different sources that assessed efficacy of behavioral treatment suggested that the number of behaviors among multiple other factors could affect the accuracy and reliability of data collection (Farkas & Tharp, 1980; Kostewicz et al., 2016; Lipinski & Nelson, 1974; Smith et al., 2013; Vollmer et al., 2008).

Issues in accuracy of data collection and measurement is important in all behavior analytic settings including research and applied settings. These issues should be addressed by extending literature on the different variables that could affect the accuracy in order to minimize risks of making inappropriate or ineffective data-based decisions, to extend the understanding of the role accuracy plays on effective behavioral interventions and treatment integrity, and to provide information to potentially aide in the establishment of systematic procedures that can be

used to assist behavior analysts in identifying the number of behaviors that can be targeted for intervention while maintaining accuracy, reliability, and high treatment integrity.

Previous research conducted parametric analyses with statistical analyses using group designs (i.e., Murphy & Harrop, 1994; La France et al., 2007) or a combination of within-subject and between-subject designs (Dorsey et al., 1986; Farmer & Nelson-Gray, 1990). To extend the research on the variables that impact accuracy of data collection and measurement, the current study focused on evaluating how the number of the behaviors and their frequencies influenced accuracy data of individual participants using a single-subjects design. The research question was, how did the number of the behaviors and their frequencies influence the accuracy of data collection for individual participants using a single-subjects design?

Method

Participants and Setting

Participants 1 and 2 were two undergraduate students taking an *Introduction to Applied Behavior Analysis* course at a four-year institution in the Midwest, and Participant 3 was a student enrolled in a graduate-level Applied Behavior Analysis program at the same institution. Participants 1 and 2 had no formal college-level training on measurement techniques. Participant 3 had three years of experience as a Registered Behavioral Technician working with individuals with disabilities.

Participants were recruited for the study through emails that was sent to all students enrolled in an introductory-level Applied Behavior Analysis course, and to first-year students in the graduate-level Applied Behavior Analysis program. Of those emails, seven individuals at the undergraduate level replied to the initial email, displaying interest in participating. Of those eight individuals, three of the potential participants did not respond to the follow-up email to set appointment times to discuss the details of the study. Two additional follow-up emails were sent in regards to setting up meeting times, but no potential participants responded. Two potential participants set up times in which to schedule an initial meeting to discuss study requirements, but one of the potential participants chose not to participate due to the estimated duration required to participate in the study, and the other potential participant did not attend the initial meeting. A follow-up email, in regards to rescheduling, was sent to that potential participant, but she did not respond to that email.

The participants met via Zoom for three different meetings: (1) A formal interview, (2) the initial assessment, and (3) the maintenance assessment. At the onset of the study, participants were given a formal interview that included discussion of criteria for participation, study

information, review of the consent form and expectations, contact information, and their consent to send them the official consent form and data sheets. All of their questions in regards to participation and expectations were also answered at this time. The formal interview, on average took approximately 31-minutes for each of the three participants. In the consent form and during the formal interviews, the researcher gave the participants general knowledge about the purpose of the study, informing them that the researcher was investigating different variables that may affect accuracy of data collection, but the exact variables were not mentioned. The participants were also offered up to $25.00 in Amazon gift cards for, a $10.00 Amazon e-gift card upon completion of the initial assessment, and a $15.00 Amazon e-gift card upon completion of the maintenance assessment.

During the initial meeting, participants were given a list of criteria that were required to be met when establishing their home Zoom environment to participate in the study's procedures. The criteria included: (1) no use of cellphones during sessions unless it is during the offered break times, (2) cellphones had to be set to silent, (3)all other electronics (i.e., radios, television, etc.) had to be turned off during the duration of sessions, (4) no pets or other individuals in the room, (5) video and audio had to be turned on during the duration of the sessions, (6) participants had to be in the camera view at all times during the sessions except during allotted breaks, (7) participants had to be wearing appropriate attire (i.e., pants/shorts and a shirt that covered cleavage), and (8) participants were not to engage in other activities unrelated to the study (i.e. laundry, reading a book, games on their phone, etc.). If these environmental arrangements did occur, participants were given a single warning that if it occurred again, they would be released from the study without the opportunity for receiving the gift card(s).

After being approved to participate, two of the participants were mailed a package that contained a copy of the consent form, data collection sheets for the initial and maintenance assessments, training manual, a blank envelope labeled, “Session 1” for the initial assessment data sheets, a blank envelope labeled, “Session 2” for the maintenance assessment data sheets, and a larger prepaid blank envelope to return all data sheets and the consent form to the researcher. The researcher contacted the individual participants via email to confirm they had received the package prior to the first session. The third participant did not live in an area of which easily received mail; thus, this participant requested to receive the data sheets, consent forms, and training manual via email, and be allowed to send PDF copies back to the experimenter researcher. Due to the individual circumstances, the participant was allowed to print the documents off and email copies back at the end of each session. Given issues with mail delivery at the time of the study, each participant held up their data sheets and consent forms for the experimenter researcher to photograph, in the event that data sheets were delayed or lost in the mail.

Dependent Variables

The dependent variable was the percent accuracy of the behavior counts by the participants compared to the true occurrences (true value) of behaviors as determined by the researcher. Participants collected data on a total of up to nine different problem behaviors (see Table 1) that occurred at different frequencies depending on the condition (see descriptions of conditions below). For each condition, the total occurrences of target behavior collected by each participant was then compared to the total true occurrences of target behaviors. The percent of accuracy was recorded by using the exact count-per-interval method described below.

The participants collected frequency data on nine different behaviors. The behaviors included kicking, hitting others, grabbing, pushing, hitting self, head banging, dropping, throwing items, and swiping items (see Table 1). These behaviors were chosen for the current study due to the ease of capturing and their discrete definitions.

To assess percent of accuracy, each participants' data were compared to the true value observers' data using exact count-per-interval measurements for each behavior across each condition. Exact count-per-interval measurements were calculated for each target behavior by adding up the number of intervals of which the observer's data were 100% in agreement with the true value observers' data, divided by the total number of intervals, and multiplying by 100. The average percent accuracy for each video was calculated by adding up the number of intervals for each target behavior in which the observer's data were 100% in agreement with the true value observers' data, divided by the total number of intervals for each behavior, and multiplied by 100. Using this data, the researcher assessed how the number of behaviors and the frequency of behaviors effected the accuracy of measurement.

Independent Variables

The independent variables were the number of behaviors being measured and the frequency of occurrence of each separate behavior occurred. To ensure that the effects of the number of behaviors and the effects of the frequency of behaviors were captured separately, there were a total of six different conditions (see Table 2). In order to ensure potential inaccurate recordings were not caused due to high-rate behavioral bursts (as previously described by Rolider and colleagues (2012)), the target behaviors were not set to occur in bursts. Target behaviors were set to occur in clear and independent instances and were not characterized by irregular inner-response times.

The number of behaviors chosen to be measured was based off an average number from previous literature. The average for the lower number of behaviors was found by adding all of the lowest number of behaviors in each study and divided it by the number of total number of studies (five). The lowest numbers of behaviors included: (1) one behavior (Murphy & Harrop, 1994), (2) one behavior (Farmer & Nelson-Gray, 1990), (3) two behaviors (La France et al., 2007), (4) three behaviors (Dorsey et al., 1986), and (5) four behaviors (Mash & McElwee, 1974). The average number of behaviors was 2.2 behaviors. The average for the highest number of behaviors to be measured was found by adding the highest number of behaviors recorded in each study and divided by the total number of studies (five). The highest number of behaviors included: (1) three behaviors (Murphy & Harrop, 1994), (2) three behaviors (Farmer & Nelson-Gray (1990) (3) eight behaviors (La France et al., 2007), (4) eight behaviors (Mash & McElwee (1974), and (5) nine behaviors (Dorsey et al., 1986). The average number of the highest number of behaviors was 6.2 behaviors. Although the average number for the highest number of behaviors recorded was approximately six, the current study chose to investigate nine behaviors, in addition to six behaviors, in order to identify potential changes in accuracy across a higher number of behaviors being recorded.

Similar to Dorsey and colleagues (1986) and Farmer and Nelson-Gray (1990), the frequency of behavior was also investigated in the current study to investigate if it was the number of behaviors being observed or the frequency in which they occurred that was affecting the accuracy of data collection. The low frequency of behaviors was calculated by assessing the condition with the lowest frequency and lowest number of behaviors for each study. The lowest number of behaviors and lowest frequency investigated were three behaviors with three repetitions each, with a total of nine behaviors (Dorsey et al., 1986) and one behavior with four

repetitions each, with a total of four behaviors (Farmer & Nelson-Gray, 1990). The total occurrences of behavior for the condition were then divided by the duration of the average video to get the average number of behaviors that occurred per minute. The average duration of videos for each study were 15-minutes (Dorsey et al., 1986) and 10.7-minutes (Farmer & Nelson-Gray, 1990). The average rate of behavior per minute for the lowest rates of behavior were .6 behaviors/minute in the study conducted by Dorsey and colleagues (1986) and .4 behaviors/minute in the study conducted by Farmer and Nelson-Gray (1990). To identify the average rate of behavior per minute between the two studies, the sum of the two averages were then divided by the number of studies (two) equivalating to .5 behaviors/minute. This rate of behavior was then multiplied by the number of minutes used in the current study (10-minutes) to identify the total instances of behavior during that video (5 behaviors). The total instances of behaviors were then divided by the lowest number of behaviors in the current study (three) to find the number of times each behavior would occur (1.7 occurrences). The current study rounded up and set the low frequency of behaviors to two instances of each behavior. D-9, F3

When determining the high frequency of behavior occurrences for the current study, similar calculations were used as those that were used to assess the low frequency of behaviors. The highest number of behaviors and highest frequency investigated were nine behaviors with thirty-three repetitions each, with a total of 297 behaviors (Dorsey et al., 1986) and three behavior with twenty repetitions each, with a total of sixty behaviors (Farmer & Nelson-Gray, 1990). The total occurrences of behavior for the condition were then divided by the duration of the average video to get the average number of behaviors that occurred per minute. The average rate of behavior per minute for the highest rates of behavior were 19.8 behaviors/minute in the study conducted by Dorsey and colleagues (1986) and 5.7 behaviors/minute in the study

conducted by Farmer and Nelson-Gray (1990). To identify the average rate of behavior per minute between the two studies, the sum of the two averages were then divided by the number of studies (two) equivalating to 12.8 behaviors/minute. This rate of behavior was then multiplied by the number of minutes used in the current study (10-minutes) to identify the total instances of behavior during that video (128 behaviors). The total instances of behaviors were then divided by the highest number of behaviors in the current study (nine) to find the number of times each behavior would occur (14.2 occurrences). However, due to time constraints to test additional conditions with changes to number and frequency of behaviors, a total of five occurrences of each target behavior, per video, was chosen as the highest frequency occurrence of behaviors. This made the highest total occurrence of behaviors 45 occurrences/10-minute video.

Materials

Materials used in the current study included a total of sixty-four, 2-5s video clips used to train the participants, and six, 10-minute videos used in the experimental conditions. Nine of the short video clips (one for each behavior) were used to model identification of the target behaviors. Twenty-three of the short video clips were used for the participants to practice accurate identification and recording of target behaviors. These videos included two clips for each of the nine behaviors and five for specific non-examples of behavior occurrences. These five non-occurrence clips included; 1) the child confederate stepping/stomping on the behavior analyst confederate's foot, 2) the child confederate engaging in a high-fives with the behavior analyst confederate, 3) the child confederate engaging in clapping during a reinforcement break, 4) the child confederate responding to a gross motor imitation instruction for clap hands, and 5) the child confederate laying on their side during a reinforcement break. The other thirty-two were used to test the participants on their ability to identify the target behaviors. The videos used

for testing included three clips of each of the nine behaviors and the five non-occurrences of behavior as described previously.

One male volunteer assisted in the making of the videos. He played the role of the child that engaged in problem behaviors and the researcher played the role of a behavior analyst conducting the sessions and delivering demands. The confederate was trained to engage in specified target behaviors through modeling, practice, and immediate feedback from the researcher. The training was based off of recommendations, by DiGennaro Reed, Blackman, Erath, Brand, and Novak (2018), to use Behavioral Skills Training (BST). The room in which videotaping occurred had a table, three chairs, work materials, and some toys (such as Legos, coloring book, a snake, and a frog).

When playing the role of the behavior analyst, the researcher engaged in typical clinical session activities such as implementing 3-4 skill acquisition demands (i.e., gross motor imitation, echoics, object imitation, instructional demands, etc.) and providing reinforcement breaks. In the event that the child confederate engaged in target behaviors, the behavior analyst confederate did not provide additional attention to the child confederate and continued with demands if they were in place during the occurrence of behavior.

For the making of the 10-minute condition videos, the confederate (playing the role of the child) wore a Bluetooth ear piece that played an audio recording indicating when to engage in a specific behavior. During each video session, at the proper time, the confederate engaged in behaviors based off of the audio recording that correlated with a specific behavior schedule for each 10-minute video. During times in which the target behaviors were not being engaged in, the child confederate complied with all demands given by the confederate behavior analyst and only

engaged in appropriate non-target behaviors, such as clapping hands when instructed to clap hands and appropriately playing with items during breaks.

For behavior schedules specific to each 10-minute video, each of the nine behaviors were assigned a number. The numbers were then randomly assigned to each of the 10-minute videos. Each video was split into 60, 10-s intervals. Each of the behaviors from the order assignments were then randomly assigned an interval number to occur. An audio mp3. file was then created in accordance for each of the videos' scripts. The researcher counted down from the beginning of a session and used a timer to indicate, on the audio recording, when each behavior was to occur according to the assigned interval. After filming, the videos were edited to dub over a short beep sound every 1 min to assist as a prompt to indicate to the participants when to move to the next interval for recording purposes. This was done to try to minimize data collection errors caused by recording target behaviors in the wrong interval.

To ensure that the predetermined instances of behavior were clear and precise and that no other times during the videos could be considered instances of behavior, three experienced observers collected true value data on all of the videos. The true value observers experienced the same training on data collection and behavior identification that the participants received (as described below). The true value observers were required to meet an 85% passing criteria or better on a post-training test, to display their ability to identify each target behavior.

When identifying and recording the occurrence of each target behavior for the six, 10-minute videos, true value observers were required to record the exact minute and second of which each behavior occurred, in addition to the number of occurrences of each behavior. They were allowed to pause, rewind, and fast-forward the videos in order to get the most precise time-stamp for each occurrence of target behavior. Their data were then assessed to identify

agreements and disagreements among the observers' data and the behavior occurrence schedule. If there was disagreement, the true value observers watched the videos again to assess whether or not the behavior met behavioral definitions or if the videos needed to be filmed again to meet the criteria of target behavior occurrences and non-occurrences. This process was repeated until there was 100% agreement among the observers that behavior was occurring on the predetermined schedule.

Experimental Design

For the current study, a parametric analysis was conducted using a multi-element design (Diller, Barry, & Gelino, 2016) to compare each participants' accuracy of data collection when measuring different numbers of behaviors (three, six, and nine behaviors) and to compare each participants' accuracy of data collection when measuring different frequencies of behaviors (two behavior occurrences per target behavior and nine behavior occurrences per target behavior). Maintenance measurements were also taken after three weeks without training or practice of behavior measurement. Each of the conditions was assigned in random order for each participant during the initial and maintenance assessments to avoid any carryover effects from previous condition exposure.

Procedures

Training

The participants and true value observers received individual training (at different times) on identification of the target behaviors and how to accurately use the data sheet for recording procedures. Training for the participants and true value observers, like the confederate training, involved modeling, practice, and immediate feedback from the researcher (DiGennaro Reed et al., 2018).

Training procedures were the same for the true value observers and participants, with the exception of data recording. Participants were instructed to collect behavior data using a frequency tally of each individual occurrence of target behavior under the corresponding clip or interval for which it occurred. True value observers were instructed to collect behavior data using a time-stamp of each individual occurrence of target behavior under the corresponding interval. To ensure the true value observers were able to record the most precise time stamp, they were allowed the ability to rewind and pause the experimental conditions, unlike the participants.

The participants and true value observers were trained for approximately 30-minutes on the definitions. During training, participants and true value observers discussed examples and non-examples of the target behaviors with the researcher and asked any questions in relation to the target behaviors and data collection. The researcher then modeled identifying and recording behaviors for a total of fourteen of the 2-5s video clips. These videos included one instance of each behavior and five non-occurrences of behavior. The participants and true value observers then were able to practice the identification and recording of behaviors on the twenty-three training clips in random order. The researcher delivered feedback in statements specific to the accuracy or inaccuracy of data collection. Correct responses received the verbal feedback, “Yes, that is correct”, and incorrect responses received the verbal feedback, “No, that is incorrect”, along with a short response for why the answer provided was incorrect. The specific feedback controlled for the potential influence of feedback expectancies discussed in previous literature (Kazdin, 1977; Lipinski & Nelson, 1974; O’Leary et al., 1975).

The participants and true value observers were then offered a 15-minute break before they were tested for their ability to identify and record data by watching the 32, 2-5s video clips previously described. Each participant was required to meet greater than or equal to the 85%

accuracy criteria for identification of occurrences or non-occurrences of problem behaviors out of the total number of video clips. This criterion had to be met before being able to move on to the experimental conditions. If they did not meet this criterion on the first attempt, then the training procedure was implemented again and the video clips were re-evaluated. If participants failed to meet the 85% criteria on the second attempt, they were then released from the study. However, all participants and true value observers were able to meet passing criteria on the first attempt (see Table 3). Participant 1 scored 100%, Participant 2 scored 93.8%, and Participant 3 scored 96.9% on the post-training test. The true value observers each scored 100% on the post-training test.

Experimental Manipulation

After completing the training and meeting the testing criteria, the participants were asked to watch a series of six videos and collect data on the frequency of target behaviors. Experimental videos corresponding with each condition were displayed in random order for each participant. The participants were not capable, nor were they given the option to rewind or watch the videos more than once.

A new data collection sheet was given for each video, corresponding to the specific behaviors that were supposed to be recorded for each condition. Each video also had a beep at the beginning of the video to indicate the beginning of the first 1-minute interval, and every 1-minute following that corresponded with the 1-minute intervals on the data collection sheet to indicate to move to the next interval to record frequency data.

Between each video, the participants were offered a 10-minute break to eat, drink, go to the bathroom, and/or look over the behavior definitions. However, during the initial assessment, all participants opted out of taking all but one of their offered breaks.

At the end of the initial assessment, participants were sent their $10.00 Amazon e-gift card via email. On average the entire training and data collection process took approximately 2.4 hr.

Maintenance Assessment

The participants were asked to meet again, via Zoom, after approximately three weeks following the original viewing of the videos to record behavior data on additional videos. The participants were told that the videos would be similar to the videos previously recorded. However, they were not told that the videos would be the exact same videos as previously scored. Training and testing procedures were not conducted during the maintenance assessment. Upon arrival, the participants only recorded data on the six, 10 min. videos displayed in random order. Participants were offered a 10 min. break after each video. However, all of the participants opted out of all breaks during the maintenance assessments.

At the end of the maintenance assessment, participants were notified that they would receive an email in regards to a disclosure statement and their results after all data were analyzed. They were also told where they could find the researcher's contact information on their copy of the consent form and were sent their $15.00 Amazon e-gift card via email. On average, the maintenance assessments took approximately 1.4 hr to complete.

Accuracy Assessment

Accuracy assessments were conducted to calculate the percent of accuracy for 83.3% of the conditions. Two experienced observers compared each participants' data to the true value data and independently calculated exact-count data for each participant using the methods just described. To assess percent of accuracy, each participants' data were compared to the true value observers' data using exact count-per-interval measurements for each behavior across each

condition. Exact count-per-interval measurements were calculated for each target behavior by calculating the number of intervals of which the observer's data were 100% in agreement with the true value observers' data, divided by the total number of intervals, and multiplying by 100. The average percent accuracy for each video was calculated by adding up the number of intervals for each target behavior in which the observer's data were 100% in agreement with the true value observers' data, divided by the total number of intervals for each behavior, and multiplied by 100. The agreement between observers when scoring each participants' data were on average: 100% agreement for Participant 1, 99.8% agreement for Participant 2, and 100% agreement for Participant 3. Participant 2's data did not meet an average of 100% agreement between the experienced observers due to two errors: 1) one observer marked one interval with agreement when there was not agreement, and 2) one observer marked all interval agreements correctly, but incorrectly calculated an interval in agreement with true value data when it was not.

Procedural Fidelity

To ensure that the researcher implemented the training and session protocols consistently across participants (i.e., with fidelity), two experienced observers were trained on the protocols and training manual. The observers were given Integrity Check Lists to score the researcher's implementation of training and session protocols (see Figure 1). A minimum of one session per participant was observed to ensure consistent implementation. Out of all six sessions (an initial and maintenance assessments for each participant) conducted, integrity data were collected on 83.3% of the sessions.

For the initial assessment, the researcher was given one point for doing each of the following: describing the session, discussing the session rules (applicable to the initial and maintenance

assessments), asking if the participant had any questions, viewing the signed consent form, training each of the nine target behaviors and their non-examples (one point was scored for each target behavior and their non-examples for a total of nine points), going over the practice clips and giving specific feedback as previously described, describing how to use data collection sheets, delivering the post-training test and assessing the results prior to moving on, taking a picture of each data sheet (each sheet was worth 1-point, totaling to 6-points), offering the breaks at the scheduled times throughout the initial assessment for a total of six breaks (each break was worth 1-point), having the participant seal the initial assessment data in an envelope on camera, setting up dates and times for the following initial assessment, and discussing/confirm receiving initial assessment compensation. Table 4 displays integrity scores for the initial and maintenance assessments across all three participants. During the initial assessment, procedures were implemented with 93.6% integrity for Participant 1, 96.7% integrity for Participant 2, and 100% integrity for Participant 3.

For maintenance assessments, the researcher was scored one point for doing the following: delivering a brief reminder to tally frequency of occurrences and zero out data if they did not occur, offering a 10 min. break between each video for a total of five scheduled breaks, sealing the data sheets in an envelope on camera, stating how the debriefing statement will be delivered, stating how to contact the researcher with questions or requests for results, and discussing/confirm receiving session compensation. Procedures were implemented with 100% integrity for Participants 1 and 2. Integrity data were not collected during the maintenance assessment for Participant 3.

Results

Number and Frequency of Behaviors

Data for Participant 1 are displayed in Figures 2 and 3. Although both graphs display the same data, Figure 2 shows conditions by the number of behaviors that occur, whereas Figure 3 displays by the frequency in which each behavior occurred.

During the initial assessment, Participant 1 collected data on three behaviors with a frequency of two and five occurrences per behavior with 100% accuracy, six behaviors with a frequency of two occurrences per behavior with 75% accuracy, six behaviors with a frequency of five occurrences per behavior with 98.3% accuracy, nine behaviors with a frequency of two occurrences per behavior with 100% accuracy, and nine behaviors with a frequency of five occurrences per behavior with 97.8% accuracy (see Figure 2). During the maintenance assessment for this participant, accuracy scores remained above 90% except for when that participant observed six behaviors occurring five times; in that case, accuracy fell to 71.7% (see Figure 2).

Figure 3 displays the same data, but focusing on the level of accuracy influenced by changes in frequency per behavioral category. When measuring conditions with behaviors occurring two times each, Participant 1 measured three behaviors with 100% accuracy, six behaviors with 75% accuracy, and nine behaviors with 100% accuracy. When measuring conditions with behaviors occurring five times each, this participant measured three behaviors with 100% accuracy, six behaviors with 98.3% accuracy, and nine behaviors with 97.8% accuracy (see Figure 3). During the maintenance assessment for this participant, accuracy scores remained above 90% when measuring conditions with behaviors occurring two times each. When measuring conditions with behaviors occurring five times each, accuracy scores remained

above 90% except for when that participant observed six behaviors occurring five times; in that case, accuracy fell to 71.7% (see Figure 3).

Data for Participant 2 are displayed in Figures 4 and 5. Although both graphs display the same data, Figure 4 compares conditions by the number of behaviors that occur, whereas Figure 5 compares conditions by the frequency in which each behavior occurred.

During the initial assessment, Participant 2 collected data on three behaviors with a frequency of two and five occurrences per behavior with 100% accuracy, six behaviors with a frequency of two occurrences per behavior with 78.3% accuracy, six behaviors with a frequency of five occurrences per behavior with 100% accuracy, nine behaviors with a frequency of two occurrences per behavior with 95.6% accuracy, and nine behaviors with a frequency of five occurrences per behavior with 60% accuracy (see Figure 4). During the maintenance assessment for this participant, accuracy scores remained above 90% except for when that participant observed six and nine behaviors with frequencies of five occurrences per behavior; in that case, accuracy fell to 75% accuracy for six behaviors and 83.3% accuracy for nine behavior (see Figure 4).

Figure 5 displays the same data, but focusing on the level of accuracy influenced by changes in frequency per behavioral category. When measuring conditions with behaviors occurring two times each, Participant 2 measured three behaviors with 100% accuracy, six behaviors with 78.3% accuracy, and nine behaviors with 95.6% accuracy. When measuring conditions with behaviors occurring five times each, this participant measured three and six behaviors with 100% accuracy and nine behaviors with 60% accuracy (see Figure 5). During the maintenance assessment for this participant, accuracy scores remained above 90% when measuring conditions with behaviors occurring two times each. When measuring conditions with

behaviors occurring five times each, this participant measured three behaviors with 96.7% accuracy, six behaviors with 75% accuracy, and nine behaviors with 83.3% accuracy (see Figure 5).

Data for Participant 3 are displayed in Figures 6 and 7. Although both graphs display the same data, Figure 6 compares conditions by the number of behaviors that occur, whereas Figure 7 compares conditions by the frequency in which each behavior occurred.

During the initial assessment, Participant 3 collected data on three behaviors with a frequency of two occurrences per behavior with 100% accuracy, three behaviors with a frequency of five occurrences per behavior with 93.33% accuracy, six behaviors with a frequency of two occurrences per behavior with 93.3% accuracy, six behaviors with a frequency of five occurrences per behavior with 100% accuracy, nine behaviors with a frequency of two occurrences per behavior with 98.9% accuracy, and nine behaviors with a frequency of five occurrences per behavior with 97.8% accuracy (see Figure 6). During the maintenance assessment for this participant, accuracy scores remained above 95% except for when that participant observed six behaviors occurring five times; in that case, accuracy fell to 93.3% (see Figure 6).

Figure 7 displays the same data, but focusing on the level of accuracy influenced by changes in frequency per behavioral category. When measuring conditions with behaviors occurring two times each, Participant 3 measured three behaviors with 100% accuracy, six behaviors with 93.3% accuracy, and nine behaviors with 98.9% accuracy. When measuring conditions with behaviors occurring five times each, this participant measured three behaviors with 93.3% accuracy, six behaviors with 100% accuracy, and nine behaviors with 97.8% accuracy (see Figure 7). During the maintenance assessment for this participant, accuracy scores

remained above 95% when measuring conditions with behaviors occurring two times each. When measuring conditions with behaviors occurring five times each, accuracy scores remained above 95% except for when that participant observed three behaviors occurring five times; in that case, accuracy fell to 93.3% (see Figure 7).

Order of Condition Presentation

To assess if the results were caused by the order presentation of condition videos, the accuracy percentages for Participant 1 were graphed and visually analyzed (see Figure 8). During the initial assessment, Participant 1 measured behaviors with lower accuracy in the first condition presented (75%), but scored close to 100% across all remaining conditions. During the maintenance assessment, the results seemed reversed. This participant showed high (90% or greater) levels of accuracy across the first five observation conditions, then scored 70% on the last condition (six behaviors each with a frequency of five).

Figure 9 displays accuracy percentages for Participant 2 to assess if the results were caused by the order presentation of condition videos. During the initial assessment, Participant 2 measured behaviors with lower accuracy in the first condition presented (78.3%), scored close to 100% on the following three conditions, scored 60% on the fifth condition (nine behaviors each with a frequency of five), before scoring 100% on the last condition. During the maintenance assessment, this participant measured behaviors with greater variability in the first three conditions presented scoring: 75% accuracy for six behaviors each with a frequency of two, 96.7% accuracy for nine behaviors each with a frequency of five, and 96.7% accuracy for nine behaviors each with a frequency of five. The participant then scored above 90% across all remaining conditions.

Figure 10 displays accuracy percentages for Participant 3 to assess if the results were caused by the order presentation of condition videos. This participant showed high (90% or greater) levels of accuracy across all conditions during the initial and maintenance assessments. During the initial assessment, Participant 3 measured behaviors with lower accuracy on the first condition presented (93.3%), scored close to 100% for the following four conditions, then scored 93.33% on the last condition (six behaviors each with a frequency of two). During the maintenance assessment, this participant scored all conditions with accuracy above 95% except when observing the second condition presented (three behaviors occurring five times); in that case, accuracy fell to 93.3% (see Figure 10).

Discussion

The purpose of this study was to extend research on the variables effecting accuracy of data collection by gaining a better understanding on how the number of behaviors being observed and their frequencies affect the accuracy of data collection. Two undergraduate students with no prior formal training on data collection and one graduate-level student with three years of prior training and experience were trained to identify and collect data on nine target behaviors. Participants collected behavior data on six, 10 min videos of which tested for accuracy effects on specific condition variables such as the number of behaviors (three, six, and nine behaviors) and their frequencies (two and five occurrences). Accuracy of data collection was variable across all participants when assessing the number of behaviors, however, all three participants displayed higher accuracy when scoring behaviors occurring two times each in comparison to so lower accuracy when measuring behaviors occurring five times each. The two main findings indicated by the results were: 1) the number of behaviors may not have an influence on the accuracy of

data collection, and 2) the frequency of behavioral occurrences may have an influence on the accuracy of data collection.

The results from the current study corresponded with the results from Murphy and Harrop (1994) indicating that the number of behaviors did not affect accuracy when collecting data on one, two, and three behaviors. Murphy and Harrop tested how the number of behaviors effected accuracy when using partial-interval recording and momentary-time sampling. The current study provided additional literature to support evidence suggesting the number of behaviors when measuring three, six, and nine behaviors does not have an effect on the accuracy of data collection when using continuous frequency count methods of measurement.

The results from the current study that suggested the number of behaviors had no effect on the accuracy of data collection also corresponded with the results found by Farmer and Nelson-Gray (1990). Farmer and Nelson-Gray suggested that the number of behaviors when measuring one, two, and three behaviors did not have an effect on the accuracy of data collection. Although results indicating the number effects on accuracy were corresponding with the current study, the current study's findings in regards to the frequency effects on the accuracy of data collection were contraindicated with the findings from Farmer and Nelson-Gray. When assessing behavior frequencies (four, twelve, and twenty occurrences) and their effects on accuracy, the results from Farmer and Nelson-Gray suggested that the higher accuracy occurred with more behavior occurrences.

These findings are also contraindicated with the findings from Dorsey and colleagues (1986) who found that accuracy scores were higher when a lower number of behaviors (3) were being measured compared to lower accuracy scores when a higher number of behaviors were being measured (9). However, the results regarding frequency were corresponding with the

results found by Dorsey and colleagues, who found that accuracy scores were higher when behaviors occurred with lower frequency, than when behaviors occurred with higher frequencies.

Number of Behaviors

These results suggest that the number of behaviors did not have a notable effect on the accuracy of data collection when testing three, six, and nine behaviors. Participant 1 consistently measured three and nine behaviors with high accuracy and slight variability and six behaviors with the most variability and the lowest accuracy during both the initial and maintenance assessments. Participant 2 consistently measured three behaviors with the highest accuracy during both the initial and maintenance assessments, nine behaviors with the lowest accuracy during the initial assessment, and six behaviors with the lowest accuracy during the maintenance assessment. Although Participant 2 measured conditions with nine behaviors with lower accuracy and more variability during the initial assessment, these results were not replicated during maintenance assessment. Across both the initial and maintenance assessments, Participant 3 recorded all conditions with high accuracy and little variability. Participant 3 scored conditions with nine behaviors with the most accuracy and least variability, and conditions with three behaviors with the lowest accuracy and most variability.

Frequency of Behaviors

Overall, the current study suggests that the frequency of behavior occurrences may have had an effect on data collection accuracy. Two out of the three participants measured conditions with five occurrences per behavior with lower accuracy in comparison to those conditions with two occurrences per behavior, across both the initial and maintenance assessments. Participant 1 was the only participant that did not demonstrate low accuracy when measuring conditions with a frequency of five occurrences per behaviors for both their initial and maintenance assessments.

Participant 1 obtained her lowest accuracy score in the condition with six behaviors occurring two times each, but recorded the remaining conditions with two occurrences per behavior with higher accuracy than conditions with five occurrences per behavior. Lower accurate measurement during the condition with six behaviors occurring two times each could have been due technological issues and glitches, issues with the data sheet, inability to hear the audio indicator for moving on to the next interval, etc. Conditions in which six behaviors were measured with a frequency of two occurrences per behavior were consistently lower in accuracy and more variable for Participant 1 and Participant 2 during the initial assessment. Conditions with six behaviors with five occurrences per behavior were consistently lower in accuracy and more variable for Participant 1 and Participant 2 during the maintenance assessment. Although Participant 3 measured behaviors under these conditions with high accuracy, the condition to measure six behaviors occurring two times each was one condition that Participant 3 measured with the lowest accuracy at 93.3%. The consistent measurements that were lower in accuracy during these conditions may support evidence suggesting that there was a potential, unidentified issue with the videos. If multiple videos were created for each condition, this could have assisted in identifying if low accuracy measurements were due to condition variables or due to issues with the video. However, these results could have also be a result of Participant 3's prior experience with data collection and training history in comparison to Participant 1 and Participant 2 whom did not have any prior formal training.

Order of Condition Presentation

As previously stated, previous research conducted parametric analyses using group designs (i.e., Murphy & Harrop, 1994; La France et al., 2007) or a combination of within-subject and between-subject designs (Dorsey et al., 1986; Farmer & Nelson-Gray, 1990). Additional

research was conducted to identify ways in which a parametric analysis could be done best using a single-subject design.

The current study used a multi-element design in an attempt to identify variables causally related to variability in accuracy scores. The choice was based on previous research. For example, Brand, Henley, DiGennaro, Gray, and Crabbs (2019) conducted research to identify published studies that used parametric analyses for manipulations of treatment integrity. To be included in the articles of assessment, Brand and colleagues required that the articles had to have been published in one of 10 journals, treatment integrity had to have been evaluated as an independent variable (i.e., evaluating integrity percentage, percentage of errors, or the manipulation of percentage of trials with errors), and studies including other variables, unrelated to treatment integrity were excluded. They found 19 studies that fit their criteria of which 39.1% used a multielement design, 36.8% used a ABAB design (or similar variation), 21.1% used a combined multiple baseline-multi-element design, and 5.3% used a multiple baseline, group, or other multiple sequence design.

There have also been other studies that have successfully conducted parametric analyses using alternating treatment or multi-element designs. Kranak, Alber-Morgan, and Sawyer (2017) used an alternating treatments/multi-element design to conduct a parametric analysis on how specific praise rates effect on-task behavior for elementary students. Jenkins and DiGennaro Reed (2016) conducted a parametric analysis using a multi-element embedded with a multiple baseline design to analyze how the number of rehearsal opportunities effected the efficacy of training package that taught how to correctly implement a functional analysis. Ueyama (2017) used a multi-element design to compare how exact interobserver agreement was affected by partial interval and momentary time sampling procedures.

One main feature of the multi-element design is to have the conditions alternating to establish experimental control and eliminate any effects that may be caused by condition order (Diller et al., 2016). The current study wanted to ensure that the order of conditions that was randomly selected for this experimental design was not responsible for participant results. The current study displayed a consistency of lower accuracy measurements for conditions measuring six behaviors, despite their placements in presentation order across all participants, indicated that accuracy percentage for these conditions may have been related to condition variables and not the order of presentation. Participants 1 and 3 displayed some variability in accuracy across all conditions during both the initial and maintenance assessments. During the initial assessment, Participant 2 also engaged in variable accuracy across the presentation of conditions. This assisted in supporting evidence to suggest that the presentation of conditions did not have an effect on the accuracy of measurement. However, during the maintenance assessment, there was variability in accuracy across the first three conditions presented to Participant 2. For the last three conditions presented, Participant 2 measured behavior with high (above 90%) near level accuracy. This could suggest that the order of presentation may have had an effect on accuracy during the maintenance assessment. This increase in accuracy could have been due to repeat exposure to the conditions, however, due to limited data and evidence, further investigation would be recommended. Overall, the general order in which conditions were presented did not seem to have an effect on accurate data collection for any of the participants within this study. These findings are in agreement with the results of Farmer and Nelson-Gray (1990).

When analyzing how the order effected accurate responding for specific variables, there should be some considerations in regards to the number of times each condition was replicated and the number of opportunities that are available to measure behaviors under each condition.

The limited amount of data were insufficient at indicating if the number of behaviors or the frequency of variables were affected by the order of presented conditions during the initial and maintenance assessments. This was a limitation to this study. Future research should use additional videos for each condition and provide multiple opportunities to measure behaviors under each condition. This could potentially provide additional information in regards to order of presentation and additional evidence on how the number of behaviors and frequency of variables affects the accuracy of data collection.

Limitations and Future Research

There are some potential limitations to this study detracting from declaring a clear relationship between accuracy and the two variables (behavior and frequencies) investigated. One potential limitation was the use of auditory and visual definitional requirements that were used in the target definitions. For example, “head banging” had an auditory requirement hearing an audible sound caused by head banging. Another example involved “pushing in” which the individual being pushed had to move a minimum of 6 in from their original placement. This required the participants to subjectively identify if they could hear an audible sound for definitions with audible components, and it required participants to have to subjectively measure the distance of an individual moving on the videos. This could have had an effect on accuracy and support the findings described by Smith and colleagues (2013), who found participants were able to score with higher accuracy on definitions that were more clearly operationally defined than not.

Another limitation was not all behaviors were tested across all conditions. Kicking, throwing, and hitting self were only measured in conditions three through condition six, with larger numbers of behaviors being measured. Grabbing, pushing, and dropping were only

measured in conditions five and six. These behaviors were never given the opportunity to be measured at conditions with lower number and frequencies of behaviors. Future research should ensure that all target behaviors are being tested across all conditions. This could also assist in identifying if a particular behavior (or the frequency of occurrence) exerts influence on accuracy assessment.

Participants did not repeat other videos for the conditions in which they were tested on. When returning for the maintenance assessment, the videos that they watched were the same to identify differences in how they were measuring behavior. As previously stated, future literature should ensure that multiple videos are made for each condition to potentially increase reliability and identify if accuracy scores are being replicated across the different videos within the condition.

One of the biggest limitations to the current study was implementing study procedures over video web cam. During the training portions of the initial assessment there were frequent computer glitches in which the audio and video would not line up during the training condition. For example, participants would occasionally report they could not see the video moving but they could hear the audio and vice versa. The only time that this was reported, outside of the training portion of the initial assessment, was once on the first video presented during the maintenance assessment for Participant 2. Although this was the only time it was reported during the 10 min videos, it may have happened to other participants and they may have failed to report it. This could have affected measurement accuracy by creating a video/audio delay that could have misconstrued the occurrence of an auditory that was required for some of the definitions therefore changing an instance of behavior occurrence to no longer fit criteria for the definition. Future research should ensure that technological malfunctions are accounted for or conduct

sessions in which the experimenter or an assistant is on sight when the participants are testing to ensure that they can fully assist participants through any technology malfunctions.

Another potential limitation was the complexity of the data collection sheets. At times when participants did not correctly mark an interval with behavior occurrence, it was sometimes noticed that the box for the interval next to the one that should have been marked was incorrectly marked as well. Although it was not assessed, it is possible that the layout, or insufficient observer training was responsible for these incidences. Literature has suggested that data collection sheets and/or lack of sufficient training may affect accuracy of data collection (Kazdin, 1977; Madsen et al., 2016; Vollmer et al., 2008). Future research can account for this variable by systematically testing different formats for data collection sheets, and testing if there was an effect of looking back and forth between the data sheet and the computer screen.

Another potential limitation was that participants were compensated for participation, not for accurate measurement of behavior. The participants could have recorded data as instructed or they could have failed to comply with instructions due to lack of motivating operations and consequences for correct data collection. There were not any motivating operations set-in place to establish motivation for accurate behavior recording. This could have potentially affected how well participants attended or maintained the skills taught in the training. Future research should investigate how compensation for participation compared to compensation for accurate responding effects accuracy of measurement for participation and to what extent it will have the effects.

The literature and results of the current study suggest that future researchers and practitioners should consider how the number of behaviors and their frequencies affect the accuracy of data collection when selecting target behaviors for assessment. Behaviors that occur

at with higher frequencies may be measured using alternative measurement systems such as partial interval methods, however, these methods may not be suitable for dangerous or high-risk behaviors. In those instances, frequency counts may be more appropriate to ensure that every behavioral occurrence is captured and reducing the number of behaviors tracked may assist with improving the accuracy of data collection.

Although indications, based off of the results of previous literature and the current study, regarding the number of behaviors and their influence on accuracy have not been consistent, researchers and practitioners should cautiously consider reducing the number of behaviors they track. If larger number of behaviors must be tracked, then consideration should be taken for circumstances in which that may influence the accuracy of data collection, and contingency plans should be prepared to reduce those influences in those circumstances. For example, a researcher may temporarily stop collecting data on the occurrences of a behavior that has minimal risks to an individual or others in their environment when a target behavior with higher risks occurs in high frequencies or at high rates. These suggestions should be followed up with more thorough investigation, however, they may temporarily serve as a solution to mitigate some measurement error. Measurement is a cornerstone of behavior analysis and good science. Research findings, clinical decisions, and educational planning are all based upon the assumption of valid measurement. Abundant literature has identified numerous variables that impact the accuracy of data collection. This study was an attempt to further explore the influence of two particular variables on the assessment of accuracy.

The literature and the current results will hopefully inform professionals in behavior analytic fields on how their behaviors and other environmental effects can affects can influence the accuracy of data collection, so that they can ensure to control for these effects and improve

the validity and reliability of their measurement systems, no matter what the particular target behaviors are.

References

Brand, D., Henley, A.J., DiGennaro Reed, F.D., Gray, E., & Crabbs, B. (2019). A review of published studies involving parametric manipulations of treatment integrity. *Journal of Behavioral Education* 28, 1–26 (2019). doi: 10.1007/s10864-018-09311-8

Bosch, S., & Fuqua, R. W. (2001). Behavioral cusps: A model for selecting target behaviors. *Journal of Applied Behavior Analysis, 34*(1), 123-125. doi: 10.1901/jaba.2001.34-123

Cooper, J. O., Heron, T. E., & Heward, W. L. (2007a). Improving and assessing the quality of behavioral measurement. (Eds.), *Applied Behavior Analysis* (pp. 102-123). Upper Saddle River, NJ: Pearson Education.

Cooper, J. O., Heron, T. E., & Heward, W. L. (2007b). Selecting, defining, and measuring behavior. (Eds.), *Applied Behavior Analysis* (pp. 48-71). Upper Saddle River, NJ: Pearson Education.

Cummings, A. R., & Carr, J. E. (2009). Evaluating progress in behavioral programs for children with autism spectrum disorders via continuous and discontinuous measurement. *Journal of applied behavior analysis*, *42*(1), 57–71. doi: 10.1901/jaba.2009.42-57

Dallery, J., Defulio, A., & Meredith, S. E. (2015). Contingency management to promote drug abstinence. Chapter 16. In H. S. Roane, J. E. Ringdahl, & T. S. Falcomata (Eds.), Clinical and organizational applications of applied behavior analysis (pp. 395-424). San Diego, CA: Elsevier, Inc.

Dempsey, C. M., Iwata, B. A., Fritz, J. N., & Rolider, N. U. (2012). Observer training revisited: a comparison of in vivo and video instruction. *Journal of applied behavior analysis*, *45*(4), 827–832. doi:10.1901/jaba.2012.45-827

DiGennaro Reed, F. D., Blackman, A. L., Erath, T. G., Brand, D., & Novak, M. D. (2018). Guidelines for Using Behavioral Skills Training to Provide Teacher Support. *TEACHING Exceptional Children*, *50*(6), 373–380. doi: 10.1177/0040059918777241

Diller, J. W., Barry, R. J., & Gelino, B. W. (2016). Visual analysis of data in a multielement design. *Journal of Applied Behavior Analysis, 49*(4), 980-985. doi: 10.1002/jaba.325

Dixon M. R. (2003). Creating a portable data-collection system with Microsoft Embedded Visual Tools for the Pocket PC. *Journal of Applied Behavior Analysis*, *36*(2), 271-84. doi: 10.1901/jaba.2003.36-271

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.

Farkas, G. M., & Tharp, R. G. (1980). Observation procedure, observer gender, and behavior valence as determinants of sampling error in a behavior assessment analogue. *Journal of Applied Behavior Analysis, 13*(3), 529-536. doi: 10.1901/jaba.1980.13-529

Farmer, R., & Nelson-Gray, R. (1990). The accuracy of counting versus estimating event frequency in behavioral assessment: The effects of behavior frequency, number of behaviors monitored, and time delay. *Behavioral Assessment, 12*(4), 425-442.

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

Giunta-Fede, T., Reeve, S. A., DeBar, R. M., Vladescu, J. C., & Reeve, K. F. (2016). Comparing continuous and discontinuous data collection during discrete trial teaching of tacting by children with autism. *Behavioral Interventions, 31*(4), 311-331. doi: 10.1002/bin.1446

Haaf, R. A., Brewster, M., de Saint Victor, Cynthia M., & Smith, P. H. (1989). Observer accuracy and observer agreement in the measurement of visual fixation with fixed-trial procedures. *Infant Behavior & Development, 12*(2), 211-220. Doi: 10.1016/0163-6383(89)90007-6

Hausman, N. L., & Kahng, S. (2015). Treatment of pediatric obesity: an opportunity for behavior analysts. Chapter 13. In H. S. Roane, J. E. Ringdahl, & T. S. Falcomata (Eds.), Clinical and organizational applications of applied behavior analysis (pp. 303-320). San Diego, CA: Elsevier, Inc.

Jang, J., Dixon, D. R., Tarbox, J., & Granpeesheh, D. (2011). Symptom severity and challenging behavior in children with ASD. *Research in Autism Spectrum Disorders, 5*(3), 1028-1032. doi: 10.1016/j.rasd.2010.11.008

Jenkins, S.R. & DiGennaro Reed, F.D. (2016). A parametric analysis of rehearsal opportunities on procedural integrity, *Journal of Organizational Behavior Management*, 36:4, 255-281, doi: 10.1080/01608061.2016.1236057

Jackson, J., & Dixon, M. R. (2007). A mobile computing solution for collecting functional analysis data on a Pocket PC. *Journal of Applied Behavior Analysis*, *40*(2), 359-84. doi: 10.1901/jaba.2007.46-06

Jerome, J., Kaplan, H., & Sturmey, P. (2014). The effects of in-service training alone and in-service training with feedback on data collection accuracy for direct-care staff working with individuals with intellectual disabilities. *Research in Developmental Disabilities, 35*(2), 529-536. doi: 10.1016/j.ridd.2013.11.009

Johnston, J. M. & Pennypacker, H. S. (1980) Stability and accuracy of measurement. (pp. 189-200). In J. M. Johnston & H. S. Pennypacker. *Strategies and Tactics of Human Behavioral Research.*

Kazdin, A. E. (1977). Artifact, bias, and complexity of assessment: The ABCs of reliability. *Journal of Applied Behavior Analysis, 10*(1), 141-150. doi: 10.1901/jaba.1977.10-141

Kostewicz, D. E., King, S. A., Datchuk, S. M., Brennan, K. M., & Casey, S. D. (2016). Data collection and measurement assessment in behavioral research: 1958–2013. *Behavior Analysis: Research and Practice, 16*(1), 19-33. doi: /10.1037/bar0000031

Kranak, M.P., Alber-Morgan, S.R., & Sawyer, M.R. (2017). A parametric analysis of specific praise rates on the on-task behavior of elementary students with autism. *Education and Training in Autism and Developmental Disabilities*, 52(4), 453-464. doi: 10.2307/26420417

La France, B., H., Heisel, A. D., & Beatty, M. J. (2007). A test of the cognitive load hypothesis: Investigating the impact of number of nonverbal cues coded and length of coding session on observer accuracy. *Communication Reports, 20*(1), 11.

LeBlanc, L. A., Lund, C., Kooken, C., Lund, J. B., & Fisher, W. W. (2019). Procedures and accuracy of discontinuous measurement of problem behavior in common practice of applied behavior analysis. *Behavior Analysis in Practice,* doi: 10.1007/s40617-019-00361-6

LeBlanc, L. A., Raetz, P. B., Sellers, T. P., & Carr, J. E. (2016). A proposed model for selecting measurement procedures for the assessment and treatment of problem behavior. *Behavior Analysis in Practice, 9*(1), 77-83. doi: 10.1007/s40617-015-0063-2

Lipinski, D., & Nelson, R. (1974). Problems in the use of naturalistic observation as a means of behavioral assessment. *Behavior Therapy, 5*(3), 341-351. doi: 10.1016/S0005-7894(74)80003-1

Machado, M. A., Luczynski, K. C., & Hood, S. A. (2019). Evaluation of the accuracy, reliability, efficiency, and acceptability of fast forwarding to score problem behavior. *Journal of Applied Behavior Analysis*, 52(1), 315–334. doi: 10.1002/jaba.510

Madsen, E. K., Peck, J. A., & Valdovinos, M. G. (2016). A review of research on direct-care staff data collection regarding the severity and function of challenging behavior in individuals with intellectual and developmental disabilities. *Journal of Intellectual Disabilities, 20*(3), 296-306. doi: 10.1177/1744629515612328

Martens, B. K., Daly, E. J., & Ardoin, S. P. (2015). Applications of applied behavior analysis to school-based instructional intervention. Chapter 6. In H. S. Roane, J. E. Ringdahl, & T. S. Falcomata (Eds.), Clinical and organizational applications of applied behavior analysis (pp. 95-124). San Diego, CA: Elsevier, Inc.

Mash, E. J., & McElwee, J. D. (1974). Situational effects on observer accuracy: Behavioral predictability, prior experience, and complexity of coding categories. *Child Development, 45*(2), 367-377. doi: 10.2307/1127957

Miltenberger, R. G. & Weil, T. M. (2012). Observation and measurement in behavior analysis. In G. J. Madden (Ed.) APA handbook of behavior analysis, Volume 1, American Psychological Association: Washington, DC.

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. doi: 10.1111/j.2044-8295.1994.tb02517.x

Normand, M. P., Dallery, J., & Ong, T. (2015). Applied behavior analysis for health and fitness. Chapter 22. In H. S. Roane, J. E. Ringdahl, & T. S. Falcomata (Eds.), Clinical and organizational applications of applied behavior analysis (pp. 555-582). San Diego, CA: Elsevier, Inc.

O'Leary, K. D., Kent, R. N., & Kanowitz, J. (1975). Shaping data collection congruent with experimental hypotheses. *Journal of Applied Behavior Analysis, 8*(1), 43-51. doi: 10.1901/jaba.1975.8-43

Rapp, J. T., Carroll, R. A., Stangeland, L., Swanson, G., & Higgins, W. J. (2011). A comparison of reliability measures for continuous and discontinuous recording methods: Inflated agreement scores with partial interval recording and momentary time sampling for duration events. *Behavior Modification, 35*(4), 389-402. doi: 10.1177/0145445511405512

Reis, M. H., Wine, B., & Brutzman, B. (2013). Enhancing the accuracy of low-frequency behavior data collection by direct-care staff. *Behavioral Interventions, 28*(4), 344-352. doi: 10.1002/bin.1371

Repp, A. C., Nieminen, G. S., Olinger, E., & Brusca, R. (1988). Direct observation: Factors affecting the accuracy of observers. *Exceptional Children, 55*(1), 29-36. doi: 10.1177/001440298805500103

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*(4), 501-508. doi: 10.1901/jaba.1976.9-501

Rolider, N. U., Iwata, B. A., Bullock, C. E. (2012). Influences of response rate and distribution on the calculation of interobserver reliability scores. *Journal of Applied Behavior Analysis, 45*(4), 753-5762. doi: 10.1901/jaba.2012.45-753

Sigurdson, S. O., & McGee, H. M. (2015). Organizational behavior management: systems analysis. Chapter 25. In H. S. Roane, J. E. Ringdahl, & T. S. Falcomata (Eds.), Clinical and organizational applications of applied behavior analysis (pp. 583-604). San Diego, CA: Elsevier, Inc.

Smith, G. D., Lambert, J. V., & Moore, Z. (2013). Behavior description effect on accuracy and reliability. *Journal of General Psychology, 140*(4), 269-281. doi: 10.1080/00221309.2013.818525

Tarbox, J., Wilke, A. E., Findel-Pyles, R., Bergstrom, R. M., & Granpeesheh, D. (2010). A comparison of electronic to traditional pen-and-paper data collection in discrete trial training for children with autism. *Research in Autism Spectrum Disorders, 4*(1), 65-75. doi: 10.1016/j.rasd.2009.07.008

Ueyama, S. R. (2017). *Evaluating teacher implementation of discontinuous data collection in the classroom* (Doctoral Dissertation). Retrieved from ProQuest Dissertations & Theses Global. (Order No. 10757834).

Van Houten, R., Axelrod, S., Bailey, J. S., Favell, J. E., Foxx, R. M., Iwata, B. A., & Lovaas, O. I. (1988). The right to effective behavioral treatment. *Journal of Applied behavior Analysis*, 21(4), 381–384. doi: 10.1901/jaba.1988.21-381

Vollmer, T. R., Sloman, K. N., & St Peter Pipkin, C. (2008). Practical implications of data reliability and treatment integrity monitoring. *Behavior analysis in practice*, *1*(2), 4–11. doi: 10.1007/BF03391722

Wu, S. M., Whiteside, U., & Neighbors, C. (2007). Differences in inter-rater reliability and accuracy for a treatment adherence scale. *Cognitive Behaviour Therapy, 36*(4), 230-239. doi: 10.1080/16506070701584367

Figure 1

Integrity Score Sheet

Participant: _____ Initial Assessment

Total	Criteria
___ / 1	Open by describing the session
___ / 1	Discuss session rules
___ / 1	Ask if they have questions
___ / 1	View signed consent form
	Train on the following definitions and show video examples and non-examples if applicable:
	__/1 Kicking __/1 Head Banging
	__/1 Hitting Others __/1 Dropping
	__/1 Grabbing __/1 Swiping
	__/1 Pushing __/1 Throwing
___ / 9	__/1 Hitting Self
	Go over the practice clips answers by doing the following (1 point total):
	Give general feedback on accuracy of participant answers stating "Yes, that is correct," or "No, that is Incorrect" (.5 total)
___ / 1	Discuss incorrect answers with the participant (.5 total)
___ / 1	Describe how to use the data collection sheets
___ / 1	Offer a 15-minute break prior to the post-training test
___ / 1	Play the post-training test and assess the data prior to moving on
___ / 6	Take pictures of each data sheet after each video (6 total)
___ / 5	Offer 10-minute breaks in between each of the 10-minute videos (5 total)
___ / 1	Did the experimenter have the participant seal the envelope on camera?
___ / 1	Set potential times and dates for session 2
___ / 1	Discuss compensation/Send compensation and have participant confirm receiving compensation
Total	__ / 31
Integrity Score	____%

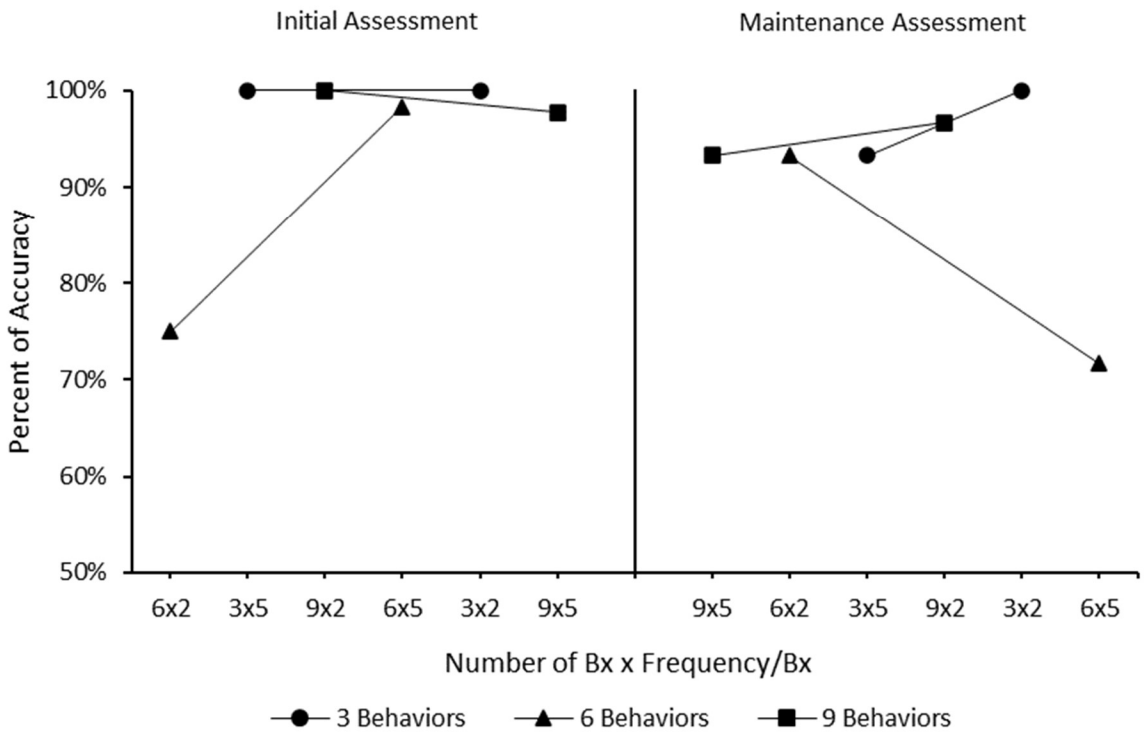
Maintenance Assessment

Total	Criteria
___ / 1	Deliver brief reminder to tally frequency of occurrences and zero out data if it did not occur
___ / 5	Offer 10-minute breaks in between each of the 10-minute videos (5 total)
___ / 1	Seal the envelope on camera
___ / 1	State how debriefing statement will be delivered
___ / 1	State how to contact the experimenter or supervisor with any questions or requesting results
___ / 1	Discuss compensation/Send session compensation and have participant confirm receiving compensation
Total	__ / 31
Integrity Score	____%

Note. This is an example of the integrity sheet that the participants filled out. It displays a breakdown of points for the integrity score sheets for the initial and maintenance assessments.

Figure 2

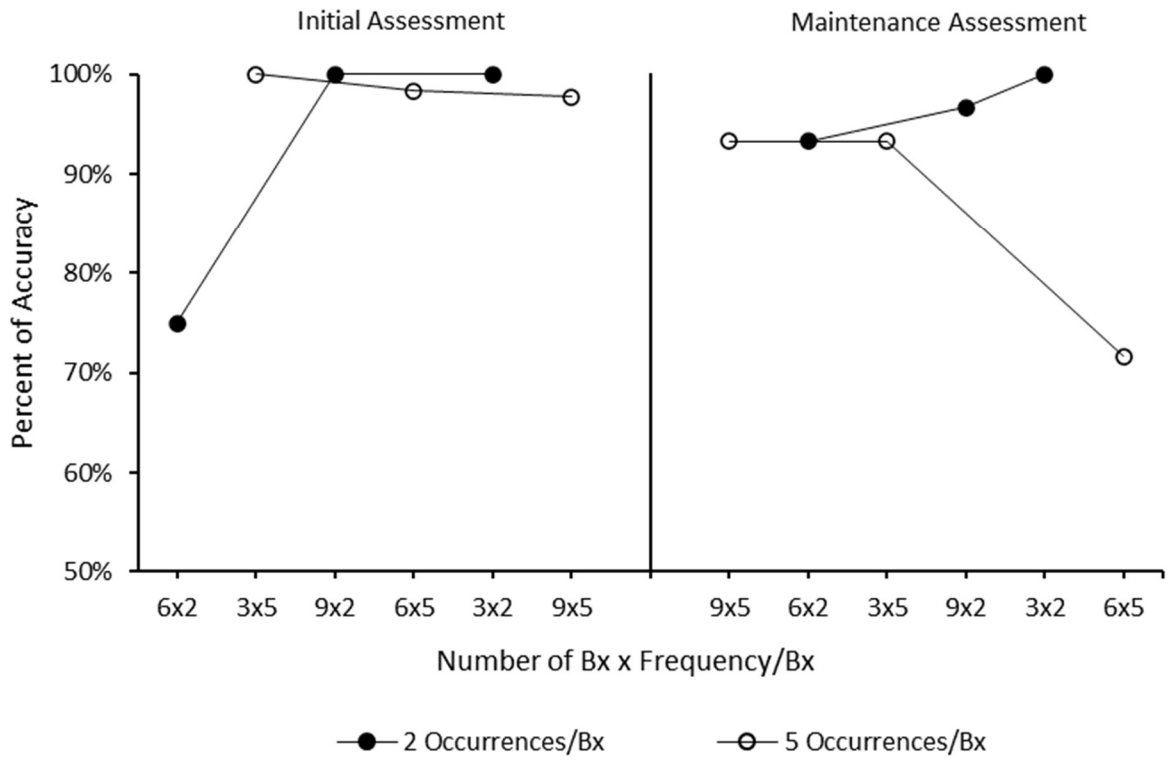
Participant 1 Number of Behavior Analysis



Note. This graph displays Participant 1's accuracy scores in the order of which the conditions were presented and how the number of behaviors effected accuracy data for the initial and maintenance assessments. The x-axis displays the number of behaviors and their frequencies of occurrences for each condition (number of behaviors x frequency of occurrences). The y-axis displays the percentage of accuracy.

Figure 3

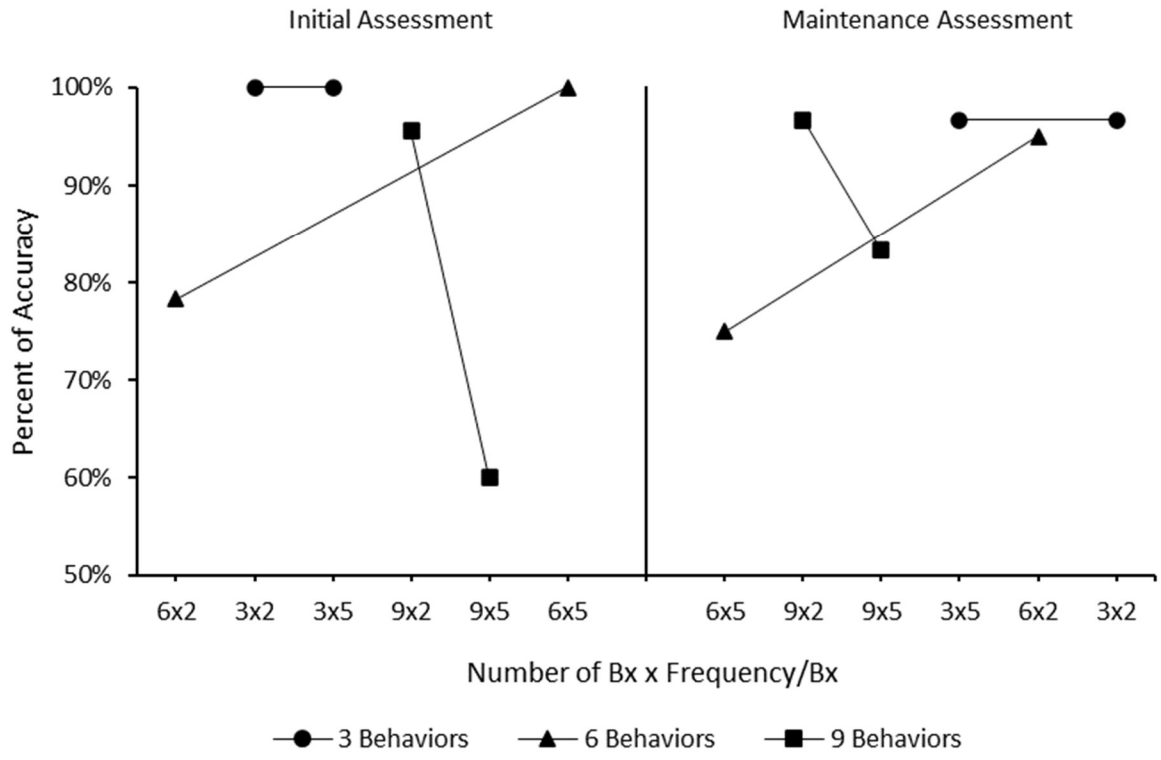
Participant 1 Frequency Analysis



Note. This graph displays Participant 1's accuracy scores in the order of which the conditions were presented and how the frequency of behaviors effected accuracy data for the initial and maintenance assessments. The x-axis displays the number of behaviors and their frequencies of occurrences for each condition (number of behaviors x frequency of occurrences). The y-axis displays the percentage of accuracy.

Figure 4

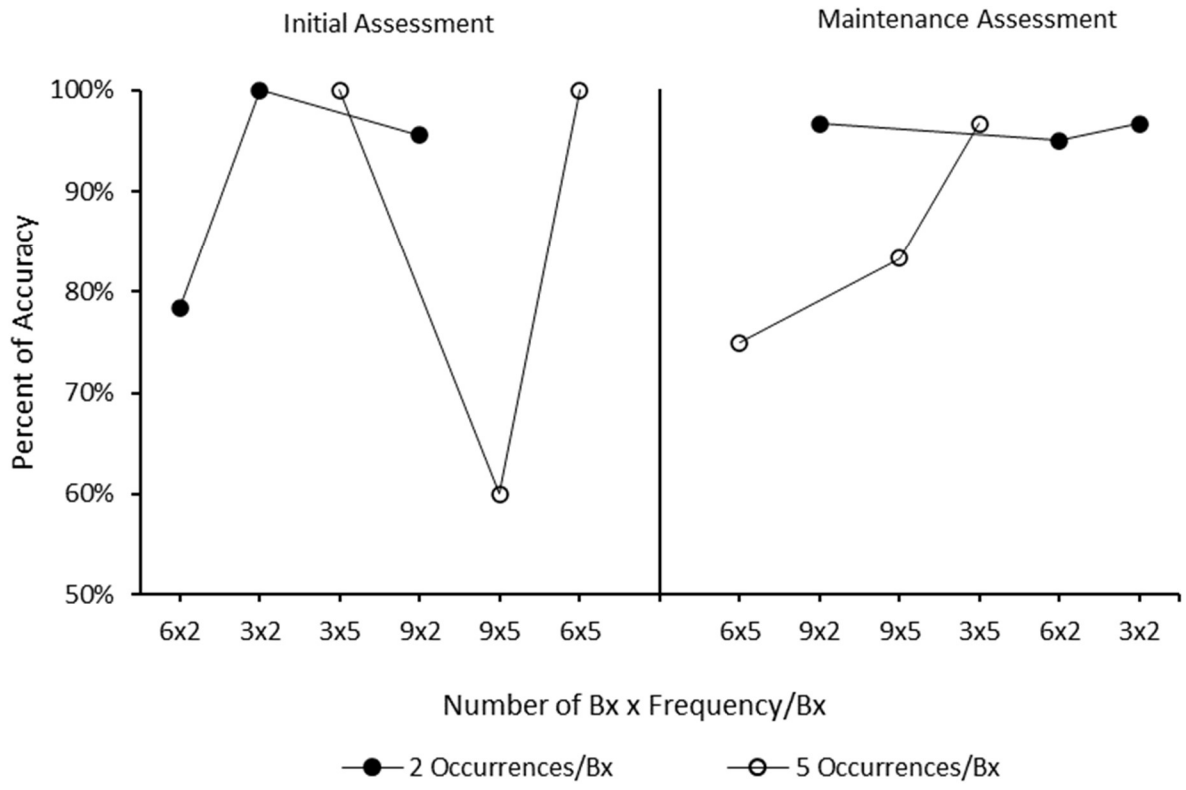
Participant 2 Number of Behavior Analysis



Note. This graph displays Participant 2's accuracy scores in the order of which the conditions were presented and how the number of behaviors effected accuracy data for the initial and maintenance assessments. The x-axis displays the number of behaviors and their frequencies of occurrences for each condition (number of behaviors x frequency of occurrences). The y-axis displays the percentage of accuracy.

Figure 5

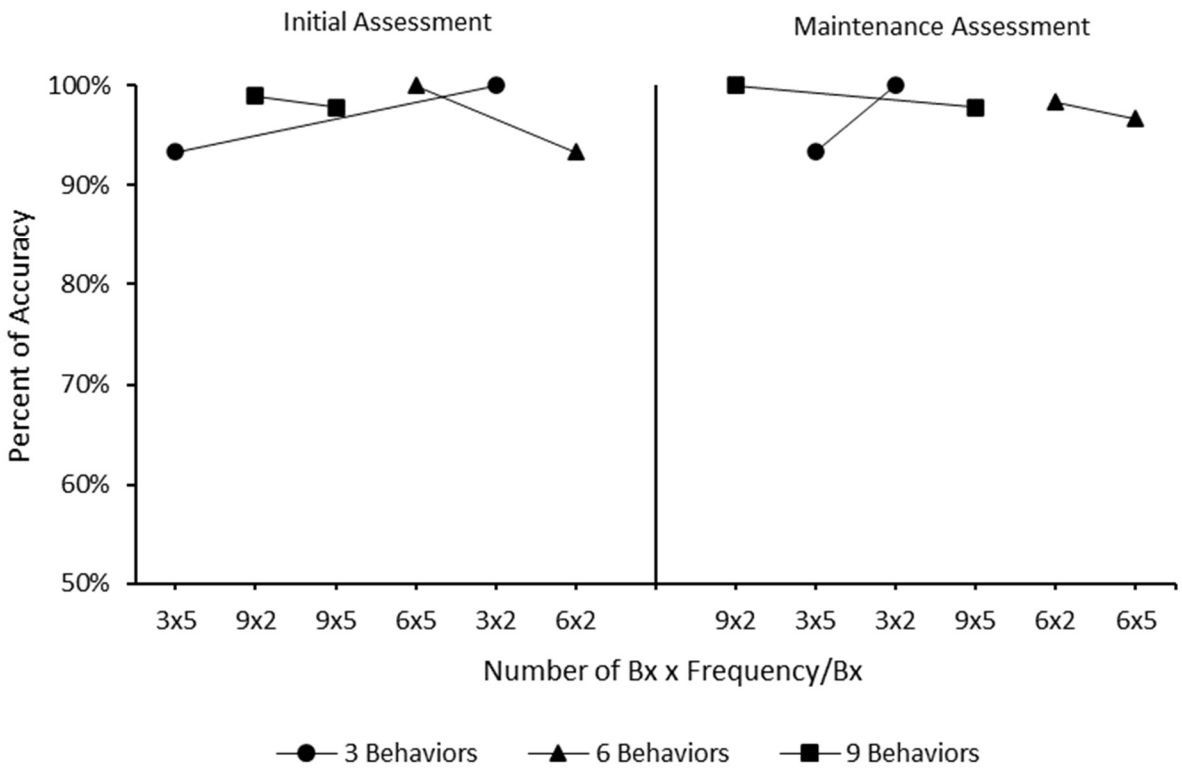
Participant 2 Frequency Analysis



Note. This graph displays Participant 2's accuracy scores in the order of which the conditions were presented and how the frequency of behaviors effected accuracy data for the initial and maintenance assessments. The x-axis displays the number of behaviors and their frequencies of occurrences for each condition (number of behaviors x frequency of occurrences). The y-axis displays the percentage of accuracy.

Figure 6

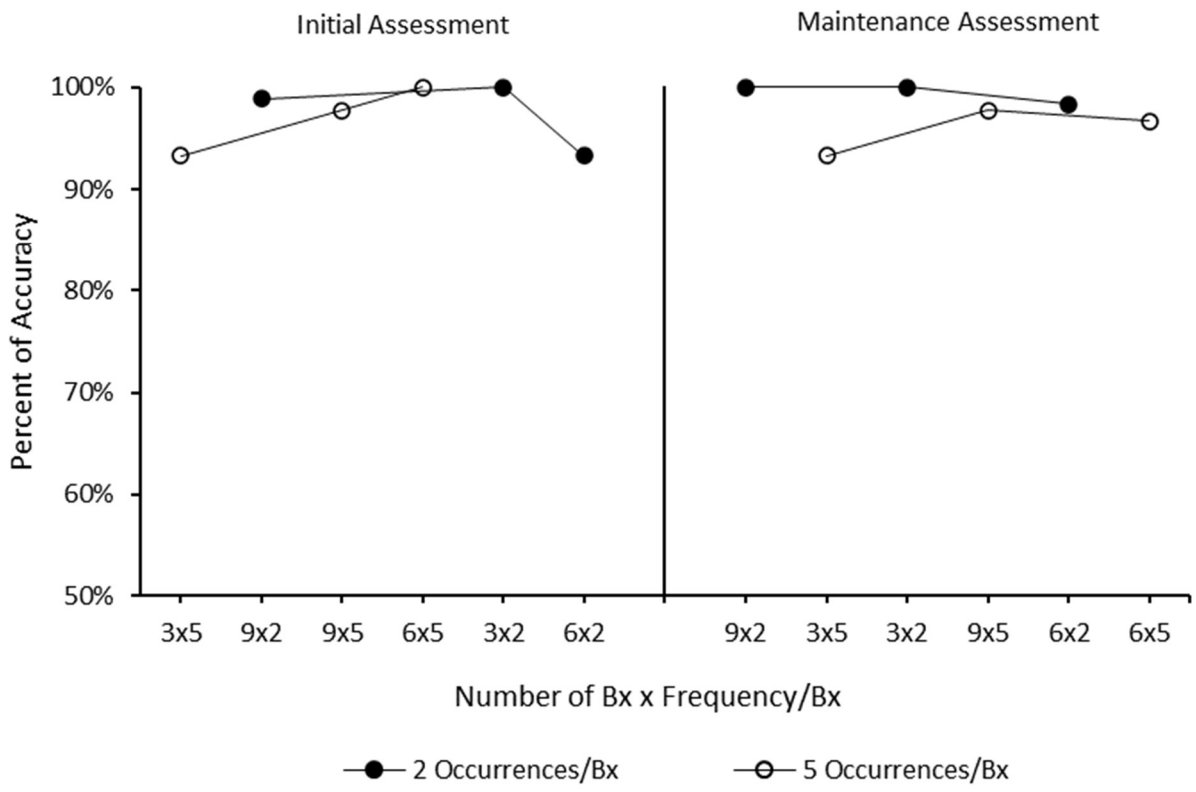
Participant 3 Number of Behavior Analysis



Note. This graph displays Participant 3's accuracy scores in the order of which the conditions were presented and how the number of behaviors effected accuracy data for the initial and maintenance assessments. The x-axis displays the number of behaviors and their frequencies of occurrences for each condition (number of behaviors x frequency of occurrences). The y-axis displays the percentage of accuracy.

Figure 7

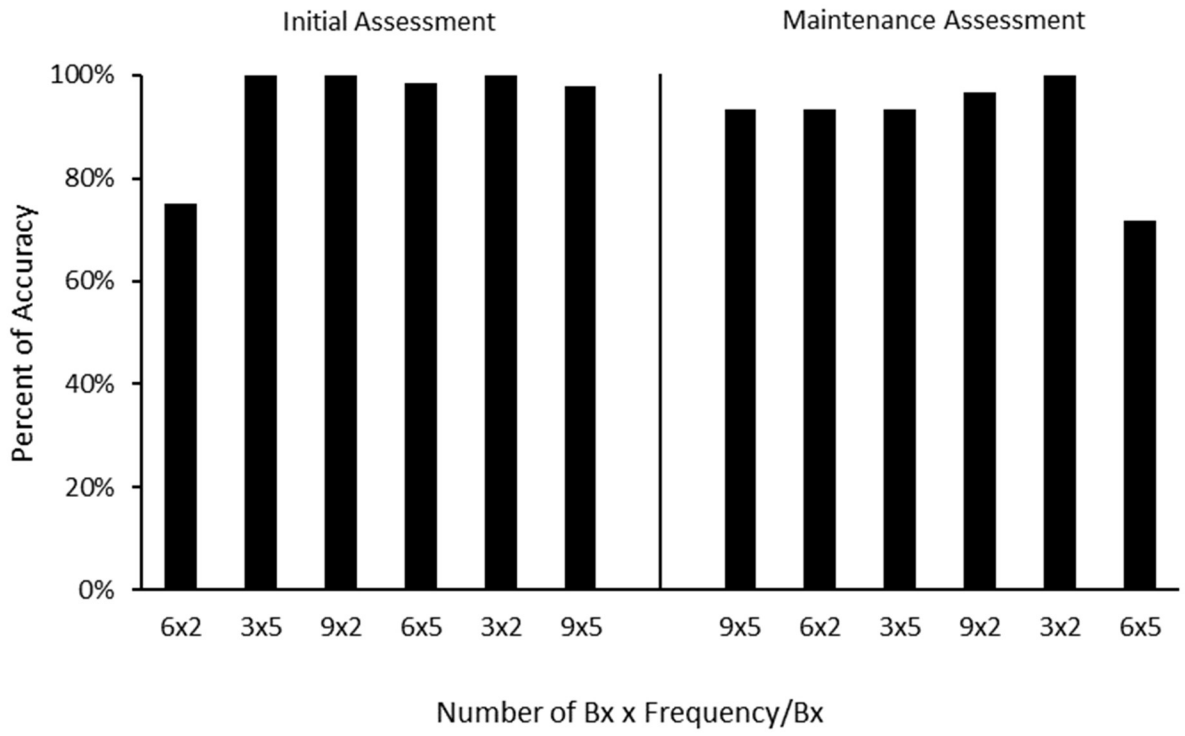
Participant 3 Frequency Analysis



Note. This graph displays Participant 3's accuracy scores in the order of which the conditions were presented and how the frequency of behaviors effected accuracy data for the initial and maintenance assessments. The x-axis displays the number of behaviors and their frequencies of occurrences for each condition (number of behaviors x frequency of occurrences). The y-axis displays the percentage of accuracy.

Figure 8

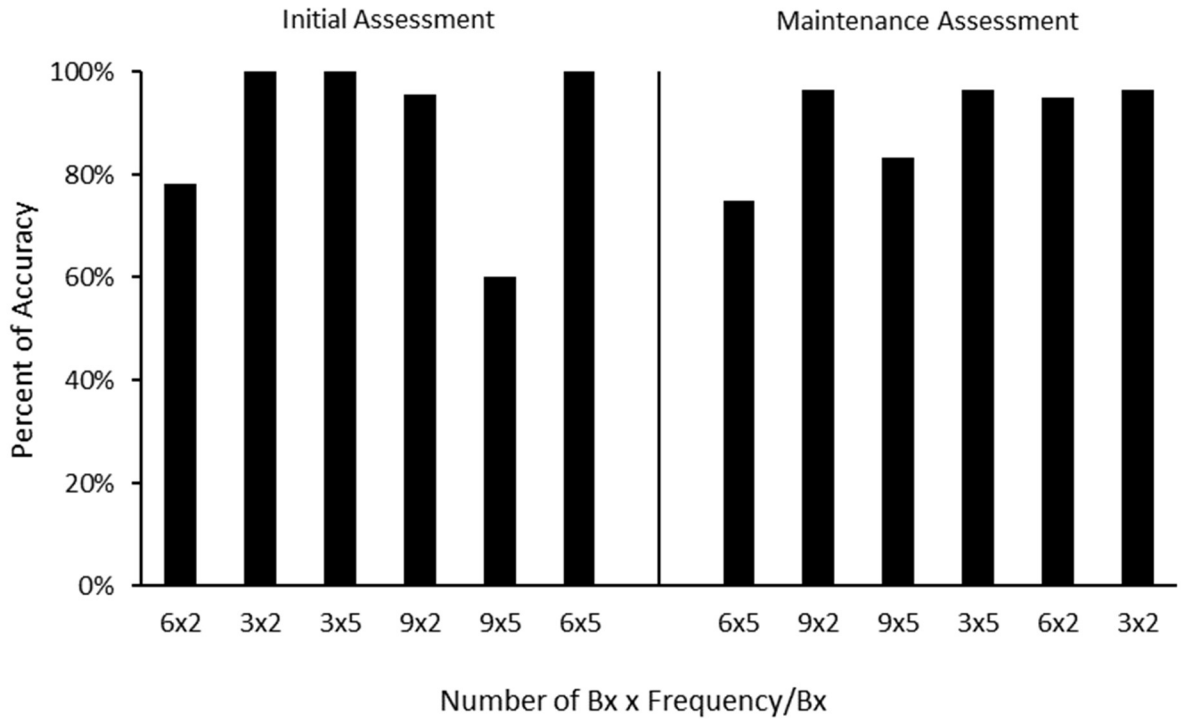
Participant 1 General Order of Condition Presentation



Note. This graph displays Participant 1's accuracy scores in the general order of which the conditions were presented for the initial and maintenance assesssments. The x-axis displays the number of behaviors and their frequencies of occurrences for each condition (number of behaviors x frequency of occurrences). The y-axis displays the percentage of accuracy.

Figure 9

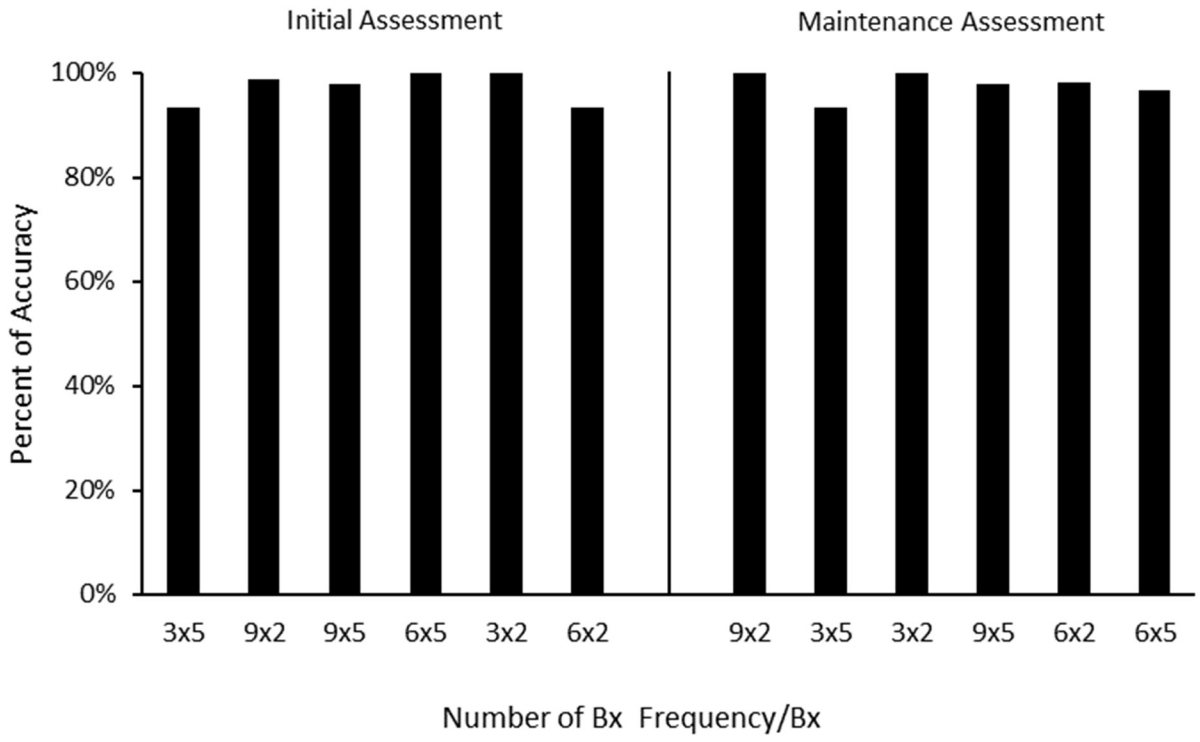
Participant 2 General Order of Condition Presentation



Note. This graph displays Participant 2's accuracy scores in the general order of which the conditions were presented for the initial and maintenance assessments. The x-axis displays the number of behaviors and their frequencies of occurrences for each condition (number of behaviors x frequency of occurrences). The y-axis displays the percentage of accuracy.

Figure 10

Participant 3 General Order of Condition Presentation



Note. This graph displays Participant 3's accuracy scores in the general order of which the conditions were presented for the initial and maintenance assessments. The x-axis displays the number of behaviors and their frequencies of occurrences for each condition (number of behaviors x frequency of occurrences). The y-axis displays the percentage of accuracy.

Table 1

Target Behavior Definitions

Behavior	Definition
Kicking	Any instance in which any part of the individual's foot made contact with any part of another individuals body. The offset of this behavior was as soon as the individual's foot no longer was touching the other individual. Non-occurrences of this behavior included when the individual stepped/stomped in a downward motion on to any part of another individual's body.
Hitting Others	Any time the individual made physical contact with another individual, using an open or closed fist, with enough force to make an audible sound. Onset of the behavior was as soon as physical contact with the other individual was made and offset was as soon as physical contact with the body part used to make contact was no longer contacting the other individual. Each contact made by each of the individual's hands were considered to be separate instance of hitting behavior. Non-occurrences included if the individual made contact with another individual during appropriate activities such as high-fives or fist bumps.
Grabbing	Any instance that the individual wraps their fingers and/or hands around any part of another individual (including body parts, hair, and/or clothing) and pulls on them with enough force to make the other individual, their clothing, and/or their hair move 1 inch or more. Onset is when the individual first makes contact and pulls. Offset is the release of the other individual, their clothing, and/or their hair, indicated by a lack of contact between the individual and another individual's body part, hair, and/or clothing.
Pushing	Any instance in which the individual uses one or two open hands to push against another individual with enough force that the other individual moves a minimum of 6-inches from their original placement.
Hitting Self	Any instance in which the individual uses an open or closed hand to make contact with their own body with enough force to make an audible sound. Onset is as soon as contact is made. Offset is as soon as their hand is no longer making contact with their body. Non-examples of this behavior include if the individual is clapping or if they are imitating another individual.
Head Banging	Any instance in which the individual moved their head in a forward or backward motion and made contact with any surface (floor, wall, table) or object they were not holding with enough force to make an audible sound. Onset of this behavior was as soon as their head made physical contact with the surface or object. Offset was as soon as their head was no longer physically in contact with that surface or object.
Dropping	Any instance in which the individual went from a standing position or sitting in a chair to laying on the ground either on their back, side, or abdomen. Onset of this behavior was as soon as their back, side, or abdomen made contact with the floor. Offset was when the individual was sitting in an up-right position with their back or abdomen no longer touching the floor.
Swiping Items	Any instance the individual used their arm or hand to move one or more item(s) on a flat surface (including the floor, a table, and/or a shelf) with enough force to move the item(s) a minimum of 2-inches from its original positioning. Onset was when the individual made contact with the item(s) and the item(s) moved more than 2-inches. Offset was as soon as the individual was no longer making physical contact with the item(s).
Throwing Items	Any instance in which one or more item(s), held in the individual's hand, left one or both of the individual's hand(s) and was in the air before falling to a surface. Onset and off set were as soon as the item(s) left the individual's hand(s) and was air bound. Non-examples included: setting an item(s) on a surface where the item(s) made contact with the surface before their hand was removed from the item(s) and anytime the individual was engaging in appropriate functional play during a break.

Note. This table displays the nine target behaviors measured by the participants, and the target behavior definitions.

Table 2

Condition Properties

Condition	Behaviors	Frequency	Rate (Behaviors/Minute)
1	Swiping Items	x2	
	Head Banging	x2	
	Hitting Others	x2	
Total	3	x6	0.6
2	Swiping Items	x5	
	Head Banging	x5	
	Hitting Others	x5	
Total	3	x15	1.5
3	Swiping Items	x2	
	Head Banging	x2	
	Hitting Others	x2	
	Hitting Self	x2	
	Kicking	x2	
	Throwing Items	x2	
Total	6	x12	1.2
4	Swiping Items	x5	
	Head Banging	x5	
	Hitting Others	x5	
	Hitting Self	x5	
	Kicking	x5	
	Throwing Items	x5	
Total	6	x30	3
5	Swiping Items	x2	
	Head Banging	x2	
	Hitting Others	x2	
	Hitting Self	x2	
	Kicking	x2	
	Throwing Items	x2	
	Dropping	x2	
	Grabbing	x2	
	Kicking	x2	
Total	9	x18	1.8
6	Swiping Items	x5	
	Head Banging	x5	
	Hitting Others	x5	
	Hitting Self	x5	
	Kicking	x5	
	Throwing Items	x5	
	Dropping	x5	
	Grabbing	x5	
	Kicking	x5	
Total	9	x45	4.5

Note. This table displays each condition's properties including the target behaviors, the number of behaviors, the frequency of each target behavior, and the rate of which the behaviors occur during the condition videos.

Table 3

Post-Training Test Results

Post-Training Test Clip	True-Value Observer 1	True-Value Observer 2	True-Value Observer 3	Participant 1	Participant 2	Participant 3
1. No Behavior	1	1	1	1	1	1
2. Throw Item(s)	1	1	1	1	1	1
3. Pushing	1	1	1	1	1	1
4. Hitting Other	1	1	1	1	1	1
5. Kicking	1	1	1	1	1	1
6. No Behavior	1	1	1	1	1	1
7. Grabbing	1	1	1	1	1	1
8. Dropping	1	1	1	1	1	1
9. Hitting Other	1	1	1	1	1	1
10. Pushing	1	1	1	1	1	1
11. Swiping Item(s)	1	1	1	1	0	1
12. No Behavior	1	1	1	1	1	1
13. Hitting Self	1	1	1	1	1	0
14. Head Banging	1	1	1	1	1	1
15. Throwing Item(s)	1	1	1	1	0	1
16. No Behavior	1	1	1	1	1	1
17. Dropping	1	1	1	1	1	1
18. Kicking	1	1	1	1	1	1
19. Hitting Self	1	1	1	1	1	1
20. Hittting Other	1	1	1	1	1	1
21. Swiping Item(s)	1	1	1	1	1	1
22. Head Banging	1	1	1	1	1	1
23. Dropping	1	1	1	1	1	1
24. Grabbing	1	1	1	1	1	1
25. Head Banging	1	1	1	1	1	1
26. No Behavior	1	1	1	1	1	1
27. Swiping Item(s)	1	1	1	1	1	1
28. Pushing	1	1	1	1	1	1
29. Throwing Item(s)	1	1	1	1	1	1
30. Hitting Self	1	1	1	1	1	1
31. Grabbing	1	1	1	1	1	1
32. Kicking	1	1	1	1	1	1
Total	32 / 32	32 / 32	32 / 32	32 / 32	30 / 32	31 / 32
Total Percent Correct	100%	100%	100%	100%	93.75%	96.88%

Note. This table displays the post-training scores for each true value observer and for each participant. To the left is each training clip in the order of which it was presented and the behavioral occurrence displayed in the video. Each Participant received a one for correct identification and a 0 for incorrect identification.

Table 4

Procedural Integrity Scores Across Participants

Initial Assessment Integrity Criteria				Participant 1	Participant 2	Participant 3
Open by describing the session				1 / 1	1 / 1	1 / 1
Discuss session rules				1 / 1	1 / 1	1 / 1
Ask if they have questions				1 / 1	1 / 1	1 / 1
View signed consent form				1 / 1	1 / 1	1 / 1
Train on the following definitions and show video examples and non-examples if applicable:						
__/1	Kicking	__/1	Head Banging			
__/1	Hitting Others	__/1	Dropping			
__/1	Grabbing	__/1	Swiping			
__/1	Pushing	__/1	Throwing			
__/1	Hitting Self			9 / 9	9 / 9	9 / 9
Go over the practice clips answers by doing the following (1 point total):						
Give general feedback on accuracy of participant answers stating "Yes, that is correct," or "No, that is Incorrect" (.5 total)						
Discuss incorrect answers with the participant (.5 total)				1 / 1	1 / 1	1 / 1
Describe how to use the data collection sheets				1 / 1	1 / 1	1 / 1
Offer a 15-minute break prior to the post-training test				1 / 1	1 / 1	1 / 1
Play the post-training test and assess the data prior to moving on				1 / 1	1 / 1	1 / 1
Take pictures of each data sheet after each video (6 total)				6 / 6	6 / 6	6 / 6
Offer 10-minute breaks in between each of the 10-minute videos (5 total)				3 / 5	4 / 5	5 / 5
Did the experimenter have the participant seal the envelope on camera?				1 / 1	N/A	1 / 1
Set potential times and dates for session 2				1 / 1	1 / 1	1 / 1
Discuss compensation/Send compensation and have participant confirm receiving compensation				1 / 1	1 / 1	1 / 1
			Total	29 / 31	29 / 30	31 / 31
			Integrity Score	93.55%	96.67%	100.00%
Maintenance Assessment Integrity Criteria						
Deliver brief reminder to tally frequency of occurrences and zero out data if it did not occur				1 / 1	1 / 1	N/A
Offer 10-minute breaks in between each of the 10-minute videos (5 total)				5 / 5	5 / 5	N/A
Seal the envelope on camera				1 / 1	N/A	N/A
State how debriefing statement will be delivered				1 / 1	1 / 1	N/A
State how to contact the experimenter or supervisor with any questions or requesting results				1 / 1	1 / 1	N/A
Discuss compensation/Send session compensation and have participant confirm receiving compensation				1 / 1	1 / 1	N/A
			Total	10 / 10	9 / 9	N/A
			Integrity Score	100.00%	100.00%	N/A

Note. This table displays the integrity scores for each of the participants' initial and maintenance assessments. To the left is the criteria that was required by the experimenter to meet when running sessions. To the right is the integrity scores scored by the observers observing procedural implementation.