2023

Impacts of Error Rate and Therapist Appearance on the Accuracy of Fidelity Data Collection

Marisela A. Aguilar
maa00019@mix.wvu.edu

Impacts of Error Rate and Therapist Appearance on the Accuracy of Fidelity Data Collection

Marisela Aguilar

Master's Thesis Defense submitted

to the Eberly College of Arts and Sciences

at West Virginia University

in partial fulfillment of the requirements for the degree of

Master of Science in

Psychology

Claire St. Peter, Ph.D., Chair

Kathryn Kestner, Ph.D.

Shari Steinman, Ph. D.

Department of Psychology

Morgantown, West Virginia

2023

Keywords: procedural fidelity, data collection, data accuracy, interobserver agreement

ABSTRACT

Impacts of Error Rate and Therapist Appearance on the Accuracy of Fidelity Data Collection

Marisela Aguilar

Procedural fidelity is the extent to which a procedure is implemented as designed. Analyzing procedural-fidelity data can improve treatment outcomes. Fidelity data are generally collected by a supervisor or trained data collector using a checklist that operationalizes each component of the procedure and accounts for errors in implementation of the components. However, little is known about variables that may affect the accuracy of supervisor-collected data generally, and even less is known about variables that may affect the accuracy of procedural-fidelity data. Therefore, the current studies explored the extent to which Board Certified Behavior Analysts (BCBAs) accurately detected programmed fidelity errors when using a tally checklist (Experiment 1) or rating scale (Experiment 2) for a resetting differential reinforcement of other behavior procedure (DRO). Nine participants were exposed to four conditions in which they watched videos of a resetting DRO with two therapists of different races/ethnicities with varied programmed errors (i.e., 80% and 40% fidelity). Participants were generally accurate regardless of the programmed level of fidelity but were slightly less accurate for the low (40%) fidelity condition with one therapist and when using a rating scale.

Table of Contents

Impacts of Error Rate and Therapist Appearance on the Accuracy of Fidelity Data Collection

Procedural fidelity is the extent to which a procedure is implemented as designed (also termed *treatment integrity, adherence,* or *procedural integrity*; DiGennaro et al., 2011). Procedural fidelity is typically reported as a percentage and can be collected and calculated in a variety of ways (Bergmann et al., 2023). One way to measure procedural fidelity is to calculate the number of steps implemented correctly and divide that value by the number of steps implemented correctly plus the number of steps implemented incorrectly (DiGennaro et al., 2005; DiGennaro et al., 2007). When a procedure is implemented exactly as designed, procedural fidelity should be 100%. Any value below 100% would mean that the implementer engaged in errors in implementation. The closer the value is to zero, the more errors there were in implementation.

Fidelity scores of trained implementers have been reported to span the entire possible range from 0% – 100%, with mean scores being as low as 3% (Cook et al., 2015; Foreman et al., 2021). In many cases, poor procedural fidelity reduces the efficacy of procedures, including increasing rates of challenging behavior (e.g., Arkoosh et al., 2007; Vollmer et al., 1999), and slowing skill acquisition (e.g., Carroll et al., 2013; Pence & St. Peter, 2015). Implementing procedures with high fidelity can reduce the risk of negative outcomes for clients. Additionally, the reporting and monitoring of procedural fidelity is increasing, suggesting that fidelity is an area of interest with opportunities for further development in experimentation (Falakfarsa et al., 2021).

As research interest increases for the monitoring and reporting of procedural fidelity, perceptions of its importance in practice might be increasing as well. Fallon et al. (2020)

surveyed 314 Board Certified Behavior Analysts (BCBAs) and Board Certified Behavior Analysts – Doctoral (BCBA-Ds) who provided in-home services to clients. Over 99% of participants agreed or strongly agreed that procedural fidelity is a key component to intervention success, and 97% of participants reported receiving some training in fidelity and its measurement. However, only 77% agreed or strongly agreed that their training was sufficient. Most participants (82%) reported that this training was primarily offered during graduate school. Thus, most BCBAs receive some form of training on how to measure fidelity but may not feel fluent or confident with these skills.

The Behavior Analyst Certification Board (BACB; 2021) tasks supervisors with determining the competency of their supervisees. One way to measure competency is to assess a supervisee's fidelity of implementation. According to the BACB ethics code, it is the responsibility of the supervisor to provide adequate training and supervision to their trainees (BACB, 2020). A supervisor may be a BCBA or a BCBA-D. Both certifications are at the graduate level, with the BCBA-D requiring a doctoral degree. There are no additional practice privileges for a BCBA-D beyond the BCBA certification. Individuals with a BCBA or BCBA-D must maintain their proficiency and work within their scope of competence. However, it may be difficult to assess how accurate a BCBA is in their collection of procedural fidelity data because, typically these data are collected during direct observation, and measuring interobserver agreement for procedural fidelity data in research is markedly low (Collier-Meek et al., 2018; Essig et al., 2023).

There are many ways that supervisors can measure fidelity, including time sampling, event recording, or checklists (Collier-Meek et al., 2021). Perhaps most often, checklists during direct observation are used to measure fidelity (Barnett et al., 2014). Procedural-fidelity

checklists are created by breaking down the procedure into discrete operationalized components (Codding et al., 2005). Procedural-fidelity checklists generally allow for recording of a component to be scored or rated as occurring (or not occurring) as written or designed (Noell et al., 2005). Bergmann et al. (2023) evaluated multiple methods for collecting procedural fidelity data for a discrete-trial procedure. Rating scales by component were used in which each component was rated on either a 3-point or 5-point Likert scale where each rating corresponded to a percentage. Aguilar et al. (2023) used a tally checklist in which each time a component was implemented it was tallied as correct or incorrect for a free-operant procedure. Checklists can look structurally different from one another but serve the same function, to assess procedural implementation. The formulation of these checklists allows for individualized performance feedback for an implementer and can identify strengths in implementation and areas for improvement (Codding et al., 2008).

Checklists can capture multiple types of fidelity errors. Two potentially important types of fidelity errors are omission errors and commission errors. Omission errors consist of failing to implement a step in the procedure as described (St. Peter Pipkin et al., 2010). For example, an omission error might consist of a therapist failing to deliver an earned reinforcer. In contrast, commission errors consist of the addition of steps or implementation of a step in the procedure in a way that is not described (St. Peter Pipkin et al.). For example, a commission error might consist of a therapist delivering a reinforcer early before a timer elapses. Because omission and commission errors may produce differential effects on intervention outcomes (e.g., Foreman et al., 2023; St. Peter Pipkin et al.), collecting accurate data on each of these error types may be important for supervisors.

Recall that it may be difficult for supervisors to assess the accuracy of their procedural fidelity data collection. One way that accuracy might be established is through calculating interobserver agreement (IOA) for fidelity data collection. IOA is the extent to which the data of two independent observers agree when recording the same response(s) using the same measurement system. There are standards for the frequency of measurement and level of agreement of IOA (i.e., 33% of sessions and at least 80% agreement; Cooper et al., 2019). Although there are standards for IOA, there are no standards for the reporting of procedural fidelity. In a recent review of two major behavior-analytic journals (*Behavior Analysis in Practice and Journal of Applied Behavior Analysis*) from 2017 – 2021 IOA of procedural fidelity was reported for 17.7% of studies for both journals (Essig et al., 2023). In contrast, IOA for the dependent variable(s) from 2017 – 2021 was 94.25% of studies for both journals. Because there is limited reporting on the IOA of procedural fidelity, little is known about how researchers collect IOA for fidelity data, or what factors might contribute to high or low levels of IOA.

Learning to collect valid and reliable data is an important aspect of behavior-analytic training. Yet, 76% of BCBAs and BCBA-Ds reported doubting the accuracy of their data, and 72% reported doubting the reliability of their data (Morris et al., 2022). Inaccurate or unreliable data may lead to negative client outcomes, particularly if "data-based" decisions are based on inaccurate data. Clients' right to effective behavioral treatment may be violated if decisions are being made using inaccurate data (Van Houten et al., 1988). There are several known factors that affect the accuracy of data collection, but some of the results are mixed.

Rate of behavior has been reported to affect the accuracy of data collection of target behavior (Dorsey et al., 1986; Kapust & Nelson, 1984; Smith et al., 1981; Smith & Sheaffer, 1984; Van Acker et. al., 1991). However, it is unclear the extent to which rate of behavior affects

accuracy of data collection. For example, Kapust and Nelson found that participants were more accurate in their data collection for low rates of behavior as opposed to high rates of behavior. In direct contrast with this finding, Smith et al. and Van Acker et al. found that participants were more accurate when scoring high rates of behavior than when scoring low rates of behavior.

Additionally, the extent to which behavior occurs in a predictable sequence may affect accuracy. Van Acker et al. (1991) provided participants with video tapes in which some tapes displayed high behavioral predictability while others displayed low predictability. Predictability was defined as the likelihood that the target response would occur following some other target event. Accuracy scores were higher when behavior occurred in a predictable sequence in comparison to a less predictable sequence. In contrast, Mash and McElwee (1974) found no differences between accuracy when observers scored predictable versus unpredictable sequences of behavior.

Another factor that might influence the accuracy of procedural-fidelity data collection is the complexity of the checklist itself. Mash and McElwee (1974) found that observers had higher accuracy when recording data from an audio tape with a four-category measurement system in comparison to an eight-category measurement system. Collier-Meek et al. (2018) identified that on average, fidelity checklists have 13 steps, with the number of steps ranging from 3 to 99. Although studies have not yet been conducted to evaluate checklist complexity, it seems plausible that more complex checklists may decrease the accuracy of the data.

In addition to factors such as checklist complexity and error rates, therapist appearance or mannerisms may affect the accuracy of fidelity data. Fidelity measures reflect an implementer's competency with implementation of a procedure. The physical appearance and mannerisms of an individual such as the clothes that they wear, their posture, and their nonverbal immediacy (i.e.,

vocal variability, gesturing, smiling) have been reported to influence ratings of the individual's competency (Glascock & Ruggiero, 2006; Gurney et al., 2017; Morris et al., 1996). An individual's perceived racial identity may also affect performance ratings. White supervisors rate the performance of Black employees significantly lower than that of White employees when the only difference between their performance is the color of their skin (Stauffer & Buckley, 2005). It seems plausible that these factors may also affect the way in which supervisors rate the performance of their trainees in behavior analysis.

Despite the importance of accurate measurement of fidelity for therapist training and subsequent client outcomes, little attention has been paid to variables that influence the accuracy of fidelity data. However, rate of behavior affects accuracy of data collection for target behavior (Dorsey et al., 1986; Kapust and Nelson, 1984; Smith et al., 1981; Van Acker et. al., 1991). Additionally, the appearance and mannerisms of individuals have affected ratings of that person's competence (Glascock & Ruggiero, 2006; Gurney et al., 2017). Although it seems logical that similar variables may influence the accuracy of fidelity data, we could find no studies to date that have directly evaluated their effects.

Moreover, research on data accuracy suggests that when checks for accuracy are performed by someone who is known, accuracy is improved (Romanczyk et al., 1973; Weinrott and Jones, 1984). Individuals in the field of behavior analysis report that they trust that a procedure has been implemented with high fidelity if it was implemented by a BCBA or from a well-known research lab/clinical site (St. Peter et al., 2023). If individuals trust that a BCBA implements a behavior-change procedure accurately, they may also trust that individual to collect fidelity data accurately. Yet, there are no studies that investigate or demonstrate that BCBAs collect fidelity-data accurately for a free-operant procedure. Therefore, the purpose of

Experiment 1 and Experiment 2 was to assess possible influences of error rate and therapist appearance on the accuracy with which BCBAs collect fidelity data for a free-operant procedure as well as if the method of procedural fidelity data collection (i.e., tally checklist or rating scale) affects accuracy.

General Method

Participants and Setting

Eight BCBAs and one BCBA-D participated (M = 32.4 years, range = 26 – 39 years). There were six participants in Experiment 1 and three in Experiment 2. Participants were recruited via email. Participation was limited to those within a 100-mile radius of a specific geographic area. Due to the small number of BCBAs-D's in and around this area, reporting demographic information at the individual level may make the participants identifiable. Therefore, participants are described in the aggregate. Participation was not limited by age, gender, ethnicity, or years of experience, but individuals whose primary affiliation was the West Virginia University Psychology Department were excluded. Seven participants identified as female and White. One participant identified as male and White. Eight participants held the supervising credential and were able to supervise individuals for their BCBA hours. To become a supervising BCBA at the time of the study, individuals were required to hold their BCBA certification for at least one year and complete an 8-hour supervision training based on the Supervisor Training Curriculum Outline by the BACB. Participants received their training to become BCBAs at multiple different programs across the United States.

Participants were compensated $8 for each checklist they completed plus $0.625 per mile traveled (range = $161.25 – $203.25). Sessions lasted on average 214.4 min (range = 172 min – 259 min). In addition to monetary compensation, all participants earned a continuing education

unit (CEU) in the domain of supervision for their participation in this study (valued at $30). At the time of the study, BCBAs were required to obtain continuing education units to "ensure that certificants continue to engage in activities that expand their behavior-analytic skills beyond the requirements for initial certification and help them stay up to date on developments in the profession." (BACB, n.d.-a, p. 2). Specifically, BCBAs with a supervision credential are required to obtain three CEUs on supervision in every recertification cycle. Participation in this study met the goal of expanding their skill sets as they practiced collecting fidelity data on a specific procedure and checklist to which they had not been previously exposed. Additionally, participants were provided feedback on the accuracy of their data collection during debriefing.

The primary employment of six participants were as BCBAs. The remaining three participants were educators. All participants reported that procedural fidelity was important, and all rated themselves as familiar or very familiar on a Likert scale of familiarity with a resetting DRO. All participants worked with children. Six participants also worked with adolescents, and three also worked with both adolescents and adults. Two participants reported that when they saw their supervisee(s), they collected fidelity data for 10% to 30% of sessions, four reported 40% to 60% of sessions, and two reported 70% to 90% of sessions. All participants reported that they collected fidelity data from direct observation, one reported that they used permanent product (e.g., tokens delivered to student in correspondence with work completed by student; Sheridan et al., 2009) in addition to direction observation, and one participant reported using self-report and interviews in addition to the previously mentioned methods for evaluating procedural fidelity. Three participants had been supervising BCBAs for less than 3 years, two participants had been supervising for 3 to 5 years, two participants had been supervising for 6 to 11 years, and one participant had been supervising for 11 to 15 years.

Participants completed the study in person in a 4.1 m by 3 m laboratory room with a table and a chair. Participants watched videos of a behavior intervention procedure (described in more detail below) on a 46.99 cm by 29.21 cm computer screen. Participants were provided pens and pencils, an eraser, a stopwatch, and a bell. All participants were allowed to access their personal belongings, including watches and phones, during the experiment.

Materials

Participants collected data using a tally checklist (Experiment 1) or rating scale (Experiment 2) for a resetting differential reinforcement of other behavior procedure (DRO; see Appendix A, B) created from a corresponding behavior intervention plan (BIP; see Appendix C). Each measurement system had 14 components.

Participants collected fidelity data from videos. Participants watched a total of 20 videos, 5 videos were from each condition (conditions described in more detail below). The content of each video showed the resetting DRO procedure implemented with varying programmed errors (described in more detail below). Each video was 5 min in duration and featured one of two therapists and mock learner. Therapist 1 was a 23-year-old White woman and Therapist 2 was a 20-year-old Black woman. The mock learner was a 20-year-old White woman and was the same individual across all videos.

We systematically manipulated both global procedural fidelity and therapists across the videos to produce four conditions: videos with 80% fidelity with Therapist 1, 40% fidelity with Therapist 1, 80% fidelity with Therapist 2, and 40% fidelity with Therapist 2. Global procedural fidelity was calculated by dividing the total number of steps completed correctly by the total number of steps completed correctly and incorrectly and multiplying this value by 100 to yield a percentage. Errors took the form of omitted procedural steps (*omission* errors) and steps that

were implemented at the wrong time or incorrectly (*commission* errors). Before filming, the therapist was instructed to engage in errors that were written on a notecard and placed out of view from the camera. The therapist timed the programmed errors at their discretion. The mock learner was instructed to engage in two instances of targeted challenging behavior (see definitions in Appendix C) at their own discretion, and an unspecified number of instances of non-target challenging behavior (e.g., tapping pencil, putting head down, sighing) in each video.

Regardless of procedural fidelity, videos included an average of 37 events that participants should have scored for procedural fidelity (range, 30 – 43). Thus, there were multiple opportunities to implement some of the components. For example, delivering a token and behavior-specific praise typically occurred multiple times within a single video. Once the videos were created, an experimenter used the procedural-fidelity checklist to verify that the programmed fidelity level in the video matched the "true value" of the procedural fidelity for that video. Any videos for which the true value was not within 5% of the programmed fidelity (i.e., 35 – 45% for a 40% video) were re-recorded until there was a set of 20 videos (five in each condition) for which the programmed and true values aligned. Videos that were programmed to have 80% fidelity with Therapist 1 had an average true value of 80% (range, 79% – 81%). Videos that were programmed to have 40% fidelity with Therapist 1 had an average true value of 38% (range, 36% - 40%). Videos that were programmed to have 80% fidelity with Therapist 2 had an average true value of 81% (range, 78% – 83%). Videos that were programmed to have 40% fidelity with Therapist 2 had an average true value of 41% (range, 40% – 44%).

Experimental Design

Each participant was exposed to four conditions in a multielement design. The order of conditions was randomized without replacement for each participant, using a random sequence

generator. The four conditions were as follows: high fidelity (Therapist 1), high fidelity (Therapist 2), low fidelity (Therapist 1), and low fidelity (Therapist 2).

Procedure

At that start of the session, participants were consented. During the consent process, the experimenter informed the participants that the purpose of the study was to evaluate measures of procedural fidelity, that their participation was voluntary, and that study findings would not be disclosed to, or affect their standing with, the BACB. Once consenting was finished, the experimenter provided a brief overview of the BIP and procedural-fidelity checklist. The experimenter read each step of the BIP aloud, described how to use the procedural-fidelity checklist, and answered participant questions using only the information from the BIP or consent form.

Immediately following the brief overview, the experimenter reviewed two videos of the BIP implemented with 100% fidelity (one video of each therapist). The order of the 100% videos was randomized. While the experimenter reviewed videos with participants, she collected data alongside the participant using the same measurement system as the participant. When the first video was reviewed, the experimenter provided examples and nonexamples of correct implementation of the procedure and demonstrated how to use the measurement system. For example, the experimenter described correct implementation of the following procedural component, “Starts/restarts 30-second timer (within 5s) at start of work block or after targeted challenging behavior.” Following correct implementation of the step, the experimenter paused the video and explained why the step was implemented correctly (e.g., the therapist accurately restarted the timer following challenging behavior within 5s of the challenging behavior). The experimenter provided at least one example of an omission and commission error if applicable

(e.g., therapist does not restart the DRO timer following an instance of targeted challenging behavior, therapist restarts the DRO timer following an instance of non-targeted challenging behavior, respectively). While reviewing the first video with participants, the experimenter paused, rewound, and fast-forwarded the video to describe correct and incorrect implementation of each step of the procedure the first time it occurred. The second video was shown in real time to mimic in-vivo data collection as well as to demonstrate data-collection procedures for the rest of the experiment. Participants were allowed to ask questions during review of videos with 100% implementation. Their questions were answered using only information from the BIP, checklist, and consent form.

After reviewing the 100% videos, the experimenter loaded the first video for the participant, provided them with a blank data sheet, instructed the participant that the experimenter would start the video from the next room, and left the room. The mouse and keyboard were removed from the computer to ensure that the participant could not pause, rewind, or fast-forward the videos. The experimenter started the videos from the observation room using a wireless keyboard.

The video displayed a title that was randomly assigned using a letter and number generator (e.g., B23)[1] for 5 s, then counted down from three and started. Each video played for 5 min, after which a black screen appeared for 2 min. Participants had 2 min to review their checklist and finish data collection after the video was over. When 2 min had elapsed, the experimenter entered the room, retrieved the participant's completed checklist, presented them with a new one, and loaded their next video. Participants were given a bell and the experimenter informed them that, if they did not need the entire 2 min to review the checklist, they could ring

[1] The purpose of the random label was to ensure that participants could not guess the fidelity level of each video and to provide a means for determining that the experimenter was displaying the correct video to the participant.

the bell and the experimenter would enter early to retrieve their checklist, provide them with a new one, and load their next video. Participants were offered a break after every fourth video and were able to request additional breaks at any time during the experiment. This process repeated for all 20 videos. Participant questions were not answered during the data-collection portion of the experiment.

After the 20th video, the experimenter informed the participant that the experiment was over and asked the participant to complete the demographic survey (Appendix D). The experimenter asked if the participant would like to receive a CEU for their participation. All participants received a CEU. While the participant completed the demographic survey, the experimenter prepared the participant's graphs. The experimenter provided feedback on the accuracy of the participant's procedural fidelity data collection by showing them graphs of their difference scores and accuracy coefficients.[2] During debriefing, the experimenter informed the participants that a secondary purpose of the study was to determine if there was bias in the participant's data collection. Participants also completed a brief interview with the experimenter about aspects of the data collection that they found easy or challenging. The experimenter provided each participant with a list of resources on procedural fidelity and cultural competency in behavior analysis. The resources included articles, podcasts, and training modules. Participants were paid in the form of a visa gift card either immediately after the experiment or by mail within one week of their participation. Participants were emailed their CEU certificate within one week of their participation.

[2] These were like the graphs in Figures 1,2, 6, and 7 but displayed only graphs for that participant.

Experimenter Fidelity

Experimenter fidelity was collected using a checklist by an undergraduate research assistant. Experimenter fidelity was calculated by adding the total number of steps completed correctly by the total number of steps correctly and incorrectly. Steps of the experimental protocol included components such as showing the correct video to the participant, offering a break after every fourth video, and answering questions using only information from the BIP, scripts, checklist, or consent form (see Appendix E). Experimenter fidelity was assessed for 89% of sessions and was 100% for all sessions.

Experiment 1

Procedural-Fidelity Measure

The procedural-fidelity checklist for Experiment 1 required participants to tally each instance of implementation for each component as correct, incorrect, or not applicable (see Appendix A). Participants were also instructed to tally data on challenging behavior to ensure that the therapist was implementing steps related to challenging behavior correctly, and additionally were asked to rate their confidence about the accuracy of their procedural-fidelity data collection on a 5-point Likert scale, for which a rating of a 1 indicated that they were not very confident and a 5 indicated that they were very confident.

Data Analysis

The primary dependent variable was accuracy of fidelity data. Fidelity was calculated by dividing the total number of tallies of correct implementation over the total number of tallies of correct and incorrect implementation and multiplying this value by 100. Accuracy was evaluated in two ways. The first way was using difference scores. Difference scores were calculated by comparing the “true fidelity” (i.e., experimenter-collected fidelity) to the participant’s obtained

fidelity. The formula for this calculation was *participant obtained fidelity – true fidelity*. The further the difference score was from zero, the lower the accuracy. Values above zero indicate that the participant over-scored the therapists implementation. Values below zero indicate that the participant under-scored the therapists implementation. The difference score has limitations. The most striking limitation is that the possible range of obtained difference scores differs across the two conditions. For the high-fidelity condition, difference scores could range from -80 to 20. For the low-fidelity condition, difference scores could range from -40 to 60. Additionally, using difference scores as a measure of accuracy is similar to using total agreement as a measure of IOA (Bijou et al. 1968). A difference score of zero may be yielded, but this does not indicate that the participant and experimenter collected identical data. However, difference scores provide a picture that may be more akin to what might be used by supervisors in clinical practice to provide feedback to trainees (Mudford et al., 2009). Additionally, difference scores allow identification of underscoring and overscoring, which is particularly important for the detection of systematic bias in scoring for one therapist.

The second way that accuracy was assessed was using accuracy coefficients. Recall that participants collected fidelity data by tallying correct implementation and errors for each component of the procedure in different cells of the data sheet. These tallies constitute the number of instances of correct and incorrect implementation for that component. Cell-by-cell accuracy was calculated by comparing the "true fidelity" to the participant's obtained fidelity. For each cell of the checklist, the smaller number was divided by the larger number between the "true fidelity" and obtained fidelity and averaging the quotients across the cells. This measure is like interobserver agreement (IOA) within blocks (Mudford et al., 2009). Accuracy coefficients

detect components in which disagreements occurred. This measure provided more information about the components, or types of errors, for which accuracy was low.

A secondary dependent variable was accuracy of component fidelity. Average accuracy for each component was calculated to determine if there were certain components that had high or low accuracy values. Cell-by-cell accuracy was calculated by comparing the "true fidelity" to the participant's obtained fidelity. For each cell of the checklist, the smaller number was divided by the larger number between the "true fidelity" and obtained fidelity and averaging the quotients across the two cells for each component (accuracy for correct and incorrect implementation). Because there was more observed differentiation between levels of fidelity rather than therapist appearance, component accuracy was separated by level of fidelity.

Another secondary dependent variable was the latency to finish data collection for each participant. Latency was calculated by setting a timer when the video was over (i.e., black screen appeared) and stopping the timer either when the participant rang the bell, or the experimenter entered the room. This was collected by a secondary data collector for each video for all but one participant, for whom the first six latencies were not recorded. To obtain the latencies after the experiment was over, the experimenter listened to the audio file and calculated the seconds between the end of the video and the time it took the participant to ring the bell or when the experimenter entered the room.

Experiment 1 Results and Discussion

Recall that Therapist 1 was the White therapist and Therapist 2 was the Black therapist. Figure 1 displays the difference scores for each participant. There were no clear patterns in difference scores across conditions for any participants. Both over scoring and under scoring occurred across videos for all participants.

Figure 2 displays the accuracy coefficients for each participant. All participants had generally high accuracy coefficients that were undifferentiated across conditions. For each participant, there are several instances of agreement above the standard in behavior analysis of 80% (Cooper et al., 2019). The lowest agreement score was observed for Elsa[3] at 73% in the Therapist 1 high-fidelity condition (Figure 2 video 3). This was an interesting finding as Elsa had the most graduate education in behavior analysis compared to the other participants. However, an accuracy coefficient of 73% is only 7% lower than the 80% standard. Table 1 displays the average accuracy for each participant for each condition. All participants except for Peter, had the lowest average accuracy value for the Therapist 1 low-fidelity condition. Taken together, these data suggest that the BCBAs in this experiment were generally accurate in their data collection for a resetting DRO procedure but were slightly less accurate when there were more errors for Therapist 1.

It was difficult to ascertain why videos of Therapist 1 implementing with low fidelity had the lowest accuracy coefficients, considering that each therapist engaged in roughly the same number of errors and the same types of errors. Recall that the therapists were instructed to engage in errors at their own discretion. Therefore, it was possible that Therapist 1 might have engaged in errors that were clustered together whereas Therapist 2 might have engaged in errors that were spaced apart. When behavior occurs in bursts, IOA can be reduced (Rolider et al., 2012). It may also be the case that errors that occur in bursts may reduce the accuracy of procedural-fidelity data collection. To examine this discrepancy, interresponse times between errors for Therapist 1 and Therapist 2 were calculated. Interresponse times were calculated by pausing the videos when an error occurred and recording the time stamp for each error. The time

[3] Participant names are pseudonyms.

from the subsequent error was subtracted from the previous error to determine the time between errors. This was then graphed in seconds (Figure 3). Therapist 1 was more likely to engage in two simultaneous errors (an interresponse time of 0s) than was Therapist 2. Future research could systematically manipulate the likelihood of simultaneous errors to determine effects on data collection accuracy.

Participant accuracy of component fidelity for all 14 components was averaged across all 40% and 80% videos (Figure 4). This analysis revealed that participants collected data less accurately for Components 2 and 14 across both levels of fidelity. Component 2 had an average accuracy value of 80.6% across all participants for high-fidelity videos and 77.2% for low-fidelity videos. Component 2 was starting/restarting the DRO timer. Component 14 had an average accuracy value of 84.2% across all participants for high-fidelity videos and 76.3% for low-fidelity videos. Component 14 was reprompting and removing the iPad when play time has ended. Component 2 occurred multiple times throughout all videos and Component 14 occurred only once or twice in each video. Component 14 was not observed during training with 100% videos, but a description of accurate and inaccurate implementation of this step was provided. This may have resulted in their lower fidelity scores for this component. In the lower-fidelity conditions, participants also collected data less accurately on Components 3 (does not comment about targeted challenging behavior; average accuracy 84.3%), 4 (Records occurrence of targeted behavior on data sheet within 5s of targeted behavior; average accuracy 82.8%), 5 (places token on the board; average accuracy 87.5%, and 6 (delivers behavior specific praise; average accuracy 85.5%). These are all components that occurred multiple times in each video, meaning that there were multiple opportunities to score implementation as correct or incorrect.

Latency to finish data collection is displayed in the first panel of Figure 5. All participants have generally similar latencies except for Kelly. On most occasions participants did not need the entire 2 min duration allotted to them to finish data collection and rang the bell to signal that they were finished after each video.

Comparing the difference scores (Figure 1) and accuracy coefficients (Figure 2) demonstrates the differences between the two measurements. For example, Kelly had a 0% difference score for Video 15 (Therapist 1 high procedural-fidelity) but had an 89% accuracy coefficient. Different conclusions may be drawn about accuracy when only evaluating difference scores than when observing accuracy coefficients. This supports findings from Repp et al. (1976): the method of agreement that an experimenter uses may impact the agreement percentages yielded. Difference scores in this study were a more lenient method for calculating agreement than were accuracy coefficients. When evaluating difference scores, one might draw the conclusion that Kelly did not require additional training or feedback to improve her fidelity data collection. However, when observing accuracy coefficients, Kelly demonstrates some small room for improvement.

Recall that a secondary purpose of the study was to determine if BCBAs were biased in their data collection. No systematic bias between therapists was detected in any of the participants' data collection. There may be several reasons for this finding. Participants were instructed to collect data using an operationalized checklist in which the measurement system was objective rather than subjective. It may be that collection of procedural-fidelity data using an operationalized checklist eliminates bias. For example, Horn and Haynes (1981) demonstrated that training observers to use an objective coding method that focused their attention on overt, operationally defined responses may reduce bias. Therefore, the purpose of Experiment 2 was to

determine if more subjective measurements of fidelity, such as the use of a rating scale, would reduce the accuracy of fidelity data collection.

Experiment 2

Procedural-Fidelity Measure

The procedural-fidelity measure for Experiment 2 required participants to rate the accuracy of implementation of each component (see Appendix B). Participants were instructed to rate implementation on a 5-point Likert scale. The rating scale was anchored to percentages adapted from Suhrheinrich et al. (2019) and Bergmann et al. (2023). For example, a rating of zero meant that the component of the procedure was not applicable or did not occur in the video, the rest of the ratings spanned from 1% – 100% with unequal bin sizes (see Table 2). To obtain the "true ratings" the experimenter converted true component fidelity calculated from the checklist used for Experiment 1 (Appendix A) to the ratings described above. This was done by dividing the number of times each component was implemented correctly by the number of times each component was implemented correctly and incorrectly and multiplying the value by 100 to yield a percentage, then converting the percentage to a corresponding rating. For example, if Component 2 was implemented with 50% fidelity, it was assigned a rating of 3. This was done for all components for all videos to compare participants' obtained ratings to the experimenter's "true ratings."

Participants were instructed to tally data on challenging behavior to ensure that the therapist was implementing steps related to challenging behavior correctly, as well as to provide an overall rating of the therapist's implementation of the entire procedure. Additionally, participants were asked to rate their confidence of procedural-fidelity data collection on a 5-point

Likert scale in which a rating of a 1 indicated that they were not very confident and a 5 indicated that they were very confident.

During debriefing, participants in Experiment 2 were shown the measure they used and the measure from Experiment 1. Participants were asked which checklist they would prefer to use in their own practice.

Data Analysis

The primary dependent variable was accuracy of the fidelity data. Global fidelity was calculated by averaging the ratings across each component to yield a mean rating for each video. Accuracy was evaluated in two ways. The first way was using difference scores. Difference scores were evaluated by comparing the "true rating" (i.e., experimenter rating) to the participant's obtained rating. The formula for this calculation was *participant obtained rating – true rating*. The further the difference score was from zero, the lower the accuracy. Values above zero indicated that the participant over-scored the therapist's implementation, values below zero indicated that the participant under-scored the therapist's implementation. The possible range of obtained difference scores was -3 to 2 for low-fidelity videos and 4 to -2 for high-fidelity videos. Like Experiment 1, evaluation of difference scores has limitations in that the true rating and experimenter rating could average to be the same rating (i.e., difference score of zero), however, individual component ratings might differ.

The second way that accuracy was assessed was using accuracy coefficients. Recall that participants collected fidelity data by rating each component of the procedure in different cells of the data sheet. Cell-by-cell accuracy was calculated by dividing the smaller rating by the larger rating for each cell of the procedural fidelity checklist and averaging the quotients across the cells. This measure is like IOA within blocks (Mudford et al., 2009). A benefit to using this

measure is that it is sensitive to components in which disagreements were detected. This measure provided more information about the components for which accuracy was low.

Accuracy of component fidelity was calculated by dividing the smaller rating by the larger rating between the "true rating" and the participant obtained rating for each component. Like in Experiment 1, accuracy values for each participant were averaged for each component across the 40% and 80% videos.

Latency to complete data collection was evaluated in the same way as Experiment 1. A secondary data collector collected latency data for Ellowyn and Nicole. The experimenter collected all latency data for Lola from the audio recording. Average latency to finish data collection was calculated for each participant by adding all the latencies for each video and dividing the value by the number of videos (20).

Experiment 2 Results and Discussion

Recall that the three BCBAs enrolled in Experiment 2 differed from the six enrolled in Experiment 1. Figure 6 displays the difference scores for each participant. All participants overscored implementation more often than they underscored: they gave the therapist more credit for their implementation than they deserved. This is consistent with findings that the use of rating scales produces higher fidelity values than other methods of fidelity-data collection (Bergmann et al., 2023). Not only did rating scales yield higher fidelity scores, but they may also have reduced the extent to which an observer may detect errors in implementation.

Figure 7 displays the accuracy coefficients for each participant. All participants obtained lower accuracy coefficients for videos of Therapist 1 with low fidelity on average (Table 1). Recall that this finding is like that of the findings of Experiment 1. The videos were the same across Experiment 1 and 2. It is still unknown why this condition resulted in lower accuracy

coefficients on average. Additionally, each participant obtained their lowest accuracy coefficient in the low fidelity condition at 75%, 80%, and 0% for Ellowyn, Nicole, and Lola, respectively.

An analysis of component fidelity revealed that Ellowyn and Nicole were generally accurate in their component fidelity data collection for high-fidelity videos (Figure 8). Lola had reduced accuracy for Components 7 (labels break period; average accuracy 52%), and 11 (makes at least 3 positive statements about behavior during play period [score 1 per play]; average accuracy 51.5%). All participants were less accurate in their data collection for low-fidelity videos for Components 3 (does not comment about targeted challenging behavior; average accuracy 72.7%) and 14 (re-prompts and gently removes iPad after 5s if Emma does not put it away; average accuracy 55.7%).

Latency to finish data collection is displayed in the second panel of Figure 5. All participants have generally similar latencies. In Experiment 2 (rating measure) latencies to finish data collection are longer than latencies to finish data collection for 5 of 6 participants in Experiment 1 (tally measure). An unpaired t-test was conducted to compare latencies for Experiment 2 and Experiment 1 (Figure 9). There was a significant difference in the latencies for the rating measure ($M = 88.3$, $SD = 22.2$) and tally measure ($M = 31.5$, $SD = 31.2$); $t(7) = 2.8$, $p = 0.027$. Readers should interpret this finding with caution. The data may not meet all the assumptions for an unpaired t-test. The sample sizes for both experiments are too small to determine if the data are normally distributed.

This experiment was a systematic replication of Bergmann et al. (2023) using a free-operant procedure as opposed to a discrete-trial procedure. Similar to the findings of Bergmann et al. the rating scale inflated fidelity compared to the calculations from Experiment 1. For example, when using the rating scale, a score of 3 could mean that the therapist implemented the

procedure with 50% or 79% fidelity. A participant could score both a high-fidelity and low-fidelity video as a 3 and could be accurate (high-fidelity video) or slightly inaccurate (low-fidelity video).

Lola's data were more differentiated than were Ellowyn's and Nicole's data. This may be because she had not been a BCBA for as long as the other participants and had not gone through the training to become a supervising BCBA. Lola's data are limited because she often did not record data for a specific component. She either ran out of time or reported not remembering how well the therapist implemented that component. The missing data resulted in multiple agreements of zero between the participant obtained ratings and the true ratings.

All participants sometimes tallied instances of correct and incorrect implementation in the comments section to assist them in their ratings. Without the prompt for a more objective measure of data collection (like in Experiment 1), participants used objective measures to help their subjective ratings of implementation. In contrast, Bergmann et al. (2023) did not permit participants to collect additional data while watching videos. To keep the instructions the same across Experiments 1 and 2 in the current studies, the experimenter modeled collecting tally data for one component (Component 11); makes at least 3 positive statements about behavior during play period (score 1 per play). The experimenter modeled keeping track of these statements by writing tallies for each statement on the procedural-fidelity measure for all participants in Experiment 1 and 2. If collecting additional data had not been permitted, participants may have been less accurate.

Recall that all participants were asked which data sheet they would prefer to use in their own practice. All participants preferred the data sheet from Experiment 1. Although rating scales may take less time to complete than do checklists in which tallies need to be made for each

instance of implementation, making tallies was preferred by the BCBAs in this study. All participants in Experiment 2 only experienced the rating scale; therefore, their stated preference should be interpreted with caution. It may be pertinent to directly assess preferences for methods of procedural-fidelity data collection to ensure that the data are accurate and frequently collected. Individuals are more likely to implement a behavior-change procedure that is preferred in comparison to a non-preferred behavior-change procedure, the same may be true for preferred data collection methods (Johnson et al., 2014). A BCBA may reduce the frequency with which they collect fidelity data if the data sheet is not preferred or is difficult to use.

General Discussion

The purpose of the current experiments was to determine the accuracy with which BCBAs collected fidelity data and if factors such as therapist appearance, error rate, and method of data collection influenced accuracy. In the environmental arrangement of the study, increased error rate slightly reduced the accuracy with which BCBAs collected fidelity data. The participating BCBAs were slightly less accurate in their collection of procedural fidelity data when using a rating scale than when using a tally checklist.

Recall that the accuracy of procedural-fidelity data collection for BCBAs for a free-operant procedure was previously unknown. In both experiments, most accuracy coefficients met or exceeded the 80% agreement standard in behavior analysis (Cooper et al., 2019). We may have created environmental conditions that supported the collection of accurate data. For example, being trained by a high-status experimenter (i.e., faculty member) has yielded lower agreement scores than a lower-status experimenter (i.e., graduate student; Taplin & Reid, 1973). In the current study, a graduate student trained the participants on data collection. When checks for accuracy are known, performed by a known person, and emphasizing accuracy over IOA also

improves accuracy of data collection (Boykin & Nelson, 1981; Reid, 1970; Romanczyk et al., 1973; Taplin & Reid, 1973; Weinrott & Jones, 1984). In the present study, participants were informed that the purpose was to assess the *accuracy* with which BCBAs collect fidelity data, that their accuracy would be assessed for *each* checklist, and that the check for accuracy was performed by the experimenter, with whom the participants interacted throughout the experiment. These variables may have supported participants in their collection of accurate data. Future studies may manipulate the status of the experimenter to determine if the accuracy of fidelity data is affected. Additionally, a comparison of overt and covert checks for accuracy of procedural-fidelity data collection could be explored.

To my knowledge, this is the first study that assesses the extent to which BCBAs can collect accurate fidelity data from a complex free-operant procedure. A previous study (Aguilar et al., 2023) recruited a community sample to collect procedural fidelity data from video models with varying levels of fidelity (i.e., 40%, 80%, 100%) of a DRO procedure, including videos used in the current study. Aguilar et al. found that participants collected fidelity data with significantly reduced accuracy when the programmed level of fidelity was lower (i.e., 40%) compared to when it was higher (i.e., 80% and 100%). The present study recruited participants with graduate training in behavior analysis, and a non-parametric Friedman test of differences among repeated measures was conducted and rendered no significant differences in the accuracy of their data collection across either programmed level of fidelity.

Bergmann et al. (2023) compared the duration of time each participant spent collecting fidelity data using different methods of data collection. Participants spent the least amount of time using the 5-pt Likert by component method (18.1 min). In the present study, participants had longer latencies to finishing data collection when using the rating scale than when using the

tally checklist (Figures 5, 9). Latencies to finish data collection were significantly different between these two methods. Additionally, Bergmann et al. compared the extent to which each method of data collection was sensitive to errors in implementation. It was found that the Likert by component methods were likely to result in missed errors. We also found that Likert scales may result in missed errors. However, our findings in comparison with Bergmann et al. should be evaluated with caution. Bergmann et al. used a discrete-trial procedure, and we used a free-operant procedure. Future research might determine if the type of behavior-change procedure used affects accuracy of fidelity data collection.

A limitation to the generality of the present findings is that the study was conducted in a laboratory. Although efforts were made to maintain ecological validity (e.g., participants were not allowed to pause, rewind, or fast-forward the videos), the laboratory is much different than a naturalistic context such as a school or clinic. In a classroom or clinic, there may be competing responsibilities that distract from the collection of procedural-fidelity data (e.g., ongoing classroom instruction). Additionally, the behavior of other individuals in the environment (not the subject of observation) can reduce the accuracy with which individuals collect data (Cunningham & Tharp, 1981). In the videos used in the present study, there were only two actors, and the laboratory setting was free of any distractions or additional personnel. Participants may not be as accurate in naturalistic contexts with additional distractions or responsibilities. Future studies could be conducted in naturalistic contexts or with a live role-play in the place of video model to determine if additional environmental factors reduce accuracy.

Another limitation is that the data were collected using permanent products. Because there were no time stamps for the participants' data collection, it is unknown if the participants were detecting the programmed errors when they occurred. It is also unknown if there are certain

types of errors (i.e., omission or commission) that participants were less accurate in detecting due to the method of data collection. Future research might assess if certain types of errors are more difficult to detect in implementation than others.

Participants were asked to rate their confidence for accuracy of their data collection on a 1 – 5 scale on the bottom of the data sheet after each video. Participants did not complete the confidence rating for 3% of the data sheets. This resulted in multiple instances of missing data within and across participants. Due to the missing data, we were unable to analyze any meaningful correspondence between confidence and accuracy. Future studies might address this limitation by requiring participants to complete the confidence rating before allowing them to continue to the next video.

Recall that participants in Experiment 2 were asked which data collection system they preferred. This was not possible to ask of participants in Experiment 1 as the second data collection system was not created at the time of the study. Future research might conduct a concurrent choice assessment in which BCBAs choose, and then use, their preferred method of data collection. Previous research has demonstrated preference for an intervention increases the fidelity with which teachers implement their preferred intervention in comparison to teachers that did not have the opportunity to state and use their preferred intervention (Johnson et al., 2014). Additionally, the teachers who selected a preferred intervention continued to use the intervention in their classrooms after coaching ended. Experimental assessment of preference for methods of procedural-fidelity data collection could uncover if preference affects accuracy of fidelity data which to my knowledge, has not yet been studied.

Future research might determine if rates of challenging behavior affect the accuracy with which individuals collect fidelity data. In the present study, each video only displayed two

instances of challenging behavior although rates of non-target challenging behavior (e.g., pencil tapping, sighing) varied. Kapust and Nelson (1984) found that high rates of challenging behavior reduce the accuracy with which participants collected data on the dependent variable. It may be particularly important to determine if rate of challenging behavior affects accuracy of fidelity data.

This study contributes to the literature in at least two ways. First, this study is the first of its kind in which racial bias was experimentally assessed for the practice and supervision of BCBAs. Results from this study suggest that, at least in the present environmental arrangement, BCBAs were unbiased in their data collection for a White therapist and Black therapist. Additional characteristics may be manipulated in future studies such as accent, age, and gender to determine if these variables affect bias in behavior analysis as they have been reported to increase bias in other fields (Baquiran & Nicoladis 2020; Gerull et al., 2019; Lev-Ari & Keysar, 2010; Richardson et al., 2013). Second, this study provides multiple opportunities for systematic replications to uncover variables that affect not only accuracy of data collection in general, but accuracy with which procedural fidelity data are collected more specifically.

References

Aguilar, M., Cooper, A. R., & St. Peter, C. C. (2023). Frequency of implementation errors negatively affects accuracy of fidelity data. *Education and Treatment of Children*. https://doi.org/10.1007/s43494-023-00097-7

Arkoosh, M. K., Derby, K. M., Wacker, D. P., Berg, W., McLaughlin, T. F., & Barretto, A. (2007). A descriptive evaluation of long-term treatment integrity. *Behavior Modification*, *31*(6), 880–895. https://doi.org/10.1177/0145445507302254

Baquiran, C., & Nicoladis, E. (2020). A doctor's foreign accent affects perceptions of competence. *Health communication*, *35*(6), 726–730. https://doi.org/10.1080/10410236.2019.1584779

Behavior Analyst Certification Board. (2020). *Ethics code for behavior analysts*. Littleton, CO: Author.

Behavior Analyst Certification Board. (n.d.-a). *Authorized continuing education providers.* https://www.bacb.com/authorized-continuing-education-providers/

Barnett, D., Hawkins, R., McCoy, D., Wahl, E., Shier, A., Denune, H., & Kimener, L. (2014). Methods used to document procedural fidelity in school-based intervention research. *Journal of Behavioral Education*, *23*(1), 89–107. https://doi.org/10.1007/s10864-013-9188-y

Behavior Analyst Certification Board (2021). Professional and ethical compliance code for behavior analysts. Littleton, CO.

Bijou, S. W., Peterson, R. F., & Ault, M. H. (1968). A method to integrate descriptive and experimental field studies at the level of data and empirical concepts. *Journal of Applied Behavior Analysis*, *1*(2), 175–191. https://doi.org/10.1901/jaba.1968.1-175

Carroll, R. A., Kodak, T., & Fisher, W. W. (2013). An evaluation of programmed treatment-integrity errors during discrete-trial instruction. *Journal of Applied Behavior Analysis*, 46, 379–394. https://doi.org/10.1002/jaba.49

Codding, R. S., Feinberg, A. B., Dunn, E. K., & Pace, G. M. (2005). Effects of immediate performance feedback on implementation of behavior support plans. *Journal of Applied Behavior Analysi*s, 38, 205–219. https://doi.org/10.1901/jaba.2005.98-04

Codding, R. S., Livanis, A., Pace, G. M., & Vaca, L. (2008). Using performance feedback to improve treatment integrity of classwide behavior plans: An investigation of observer reactivity. *Journal of Applied Behavior Analysis*, *41*(3), 417–422. https://doi.org/10.1901/jaba.2008.41-417

Collier-Meek, M. A., Fallon, L. M., & Gould, K. (2018). How are treatment integrity data assessed? Reviewing the performance feedback literature. *School Psychology Quarterly, 33*(4), 517–526. https://doi.org/10.1037/spq0000239

Collier-Meek, M. A., Sanetti, L. M. H., Gould, K., & Pereira, B., (2021). An exploratory comparison of three treatment fidelity assessment methods: time sampling, event recording, and post-observation checklist. *Journal of Educational and Psychological Consultation*, 31(3), 334-359, https://doi.org/10.1080/10474412.2020.1777874

Cook, J. E., Subramaniam, S., Brunson, L. Y., Larson, N. A., Poe, S. G., & St. Peter, C. C. (2015). Global measures of treatment integrity may mask important errors in discrete-trial training. *Behavior Analysis in Practice, 8*(1), 37–47. https://doi.org/10.1007/s40617-014-0039-7

Cooper, J. O., Heron, T. E., & Heward, W. L. (2019). Applied Behavior Analysis (3rd Edition) (p. 117). Hoboken, NJ: Pearson Education

Cunningham, T. R., & Tharp, R. G. (1981). The influence of settings on accuracy and reliability of behavioral observation. *Behavioral Assessment*, *3*(1), 67–78.

DiGennaro, F. D., Martens, B. K., & McIntyre, L. L. (2005). Increasing treatment integrity through negative reinforcement: Effects on teacher and student behavior. *School Psychology Review*, *34*(2), 220–231. https://doi.org/10.1080/02796015.2005.12086284

DiGennaro, F. D., Martens, B. K., & Kleinmann, A. E. (2007). A comparison of performance feedback procedures on teachers' treatment implementation integrity and students' inappropriate behavior in special education classrooms. *Journal of Applied Behavior Analysis*, *40*(3), 447–461. https://doi.org/10.1901/jaba.2007.40-447

DiGennaro Reed, F. D., Reed, D. D., Baez, C. N., & Maguire, H. (2011). A parametric analysis of errors of commission during discrete-trial training. *Journal of Applied Behavior Analysis*, 44(3), 611–615. https://doi.org/10.1901/jaba.2011.44-611

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.

Essig, L., Rotta, K., & Poling, A. (2023). Interobserver agreement and procedural fidelity: An odd asymmetry. *Journal of Applied Behavior Analysis*, *56*(1), 78–85. https://doi.org/10.1002/jaba.961

Falakfarsa, G., Brand, D., Jones, L., Godinez, E. S., Richardson, D. C., Hanson, R. J., Velazquez, S. D., & Wills, C. (2021). Treatment Integrity Reporting in *Behavior Analysis in Practice* 2008-2019. *Behavior Analysis in Practice*, *15*(2), 443–453. https://doi.org/10.1007/s40617-021-00573-9

Fallon, L. M., Cathcart, S. C., & Sanetti, L. M. H. (2020). Assessing parents' treatment fidelity: A survey of practitioners in home settings. *Focus on Autism and Other Developmental Disabilities*, *35*(1), 15–25. https://doi.org/10.1177/1088357619866192

Foreman, A. P., Peter, C., Mesches, G. A., Robinson, N., & Romano, L. M. (2021). Treatment integrity failures during timeout from play. *Behavior Modification*, *45*(6), 988–1010. https://doi.org/10.1177/0145445520935392

Foreman, A. P., Romano, L. M., Mesches, G. A., & St. Peter, C. C. (2023). A translational evaluation of commission fidelity errors on differential reinforcement of other behavior. *The Psychological Record*, *73*(1), 97–104. https://doi.org/10.1007/s40732-022-00528-8

Gerull, K. M., Loe, M., Seiler, K., McAllister, J., & Salles, A. (2019). Assessing gender bias in qualitative evaluations of surgical residents. *American Journal of Surgery*, *217*(2), 306–313. https://doi.org/10.1016/j.amjsurg.2018.09.029

Glascock, J., & Ruggiero, T. E. (2006). The relationship of ethnicity and sex to professor credibility at a culturally diverse university. *Communication Education, 55*(2), 197–207. https://doi.org/10.1080/03634520600566165

Gurney, D. J., Howlett, N., Pine, K., Tracey, M., & Moggridge, R. (2017). Dressing up posture: The interactive effects of posture and clothing on competency judgements. *British Journal of Psychology*, *108*(2), 436–451. https://doi.org/10.1111/bjop.12209

Hanley, G. P., Piazza, C. C., Fisher, W. W., & Maglieri, K. A. (2005). On the effectiveness of and preference for punishment and extinction components of function-based interventions. *Journal of Applied Behavior Analysis*, *38*(1), 51–65. https://doi.org/10.1901/jaba.2005.6-04

Horn, W. F., & Haynes, S. N. (1981). An investigation of sex bias in behavioral observations and ratings. *Behavioral Assessment*, 3, 173-183.

Johnson, L. D., Symons, F. J., Wehby, J. H., Moore, T. C., Maggin, D. M., & Sutherland, K. S. (2014). An analysis of preference relative to teacher implementation of intervention. *Journal of Special Education*, *48*(3), 214–224. https://doi.org/10.1177/0022466913475872

Kapust, J. A., & Nelson, R. O. (1984). Effects of the rate and spatial separation of target behaviors on observer accuracy and interobserver agreement. *Behavioral Assessment*, *6*(3), 253–262.

Lev-Ari, S., & Keysar, B. (2010). Why don't we believe non-native speakers? The influence of accent on credibility. *Journal of Experimental Social Psychology, 46*(6), 1093–1096. https://doi.org/10.1016/j.jesp.2010.05.025

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. https://doi.org/10.2307/1127957

Morris, C., Conway, A. A., Becraft, J. L., & Ferrucci, B. J. (2022). Toward an understanding of data collection integrity. *Behavior Analysis in Practice*, *15*(4), 1361–1372. https://doi.org/10.1007/s40617-022-00684-x

Morris, T. L., Gorham, J., Cohen, S. H., & Huffman, D. (1996). Fashion in the classroom: Effects of attire on student perceptions of instructors in college classes. *Communication Education, 45*(2), 135–148. https://doi.org/10.1080/03634529609379043

Mudford, O. C., Martin, N. T., Hui, J. K., & Taylor, S. A. (2009). Assessing observer accuracy in continuous recording of rate and duration: Three algorithms compared. *Journal of Applied Behavior Analysis*, *42*(3), 527–539. https://doi.org/10.1901/jaba.2009.42-527

Mudford, O. C., Taylor, S. A., & Martin, N. T. (2009). Continuous recording and interobserver agreement algorithms reported in the Journal of Applied Behavior Analysis (1995-2005). *Journal of applied behavior analysis*, *42*(1), 165–169. https://doi.org/10.1901/jaba.2009.42-165

Noell, G. H., Witt, J. C., Slider, N. J., Connell, J. E., Gatti, S. L., Williams, K. L., . . . Duhon, G. J. (2005). Treatment implementation following behavioral consultation in schools: A comparison of three follow-up strategies. *School Psychology Review*, *34*(1), 87–106.

Pence, S. T., & St Peter, C. C. (2015). Evaluation of treatment integrity errors on mand acquisition. *Journal of Applied Behavior Analysis*, *48*(3), 575–589. https://doi.org/10.1002/jaba.238

Reid, J. B. (1970). Reliability assessment of observation data: A possible methodological problem. *Child Development*, *41*(4), 1143–1150. https://doi.org/10.2307/1127341

Repp, A. C., Deitz, D. E., Boles, S. M., Deitz, S. M., & Repp, C. F. (1976). Differences among common methods for calculating interobserver agreement. *Journal of Applied Behavior Analysis*, *9*(1), 109–113. https://doi.org/10.1901/jaba.1976.9-109

Richardson, B., Webb, J., Webber, L., & Smith, K. (2013). Age discrimination in the evaluation of job applicants. *Journal of Applied Social Psychology*, *43*(1), 35–44. https://doi.org/10.1111/j.1559-1816.2012.00979.x

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), 753–762. https://doi.org/10.1901/jaba.2012.45-753

Romanczyk, R. G., Kent, R. N., Diament, C., & O'leary, K. D. (1973). Measuring the reliability of observational data: a reactive process. *Journal of Applied Behavior Analysis*, *6*(1), 175–184. https://doi.org/10.1901/jaba.1973.6-175

Sheridan, S. M., Swanger-Gagné, M., Welch, G. W., Kwon, K., & Garbacz, S. A. (2009). Fidelity measurement in consultation: Psychometric issues and preliminary examination. *School Psychology Review*, *38*(4), 476–495.

Smith, J. B., Madsen Jr., C. H., & Cipani, E. C. (1981). The effects of observational session length, method of recording, and frequency of teacher behavior on reliability and accuracy of observational data. *Behavior Therapy*, *12*(4), 565–569. https://doi.org/10.1016/S0005-7894(81)80096-2

Smith, G. A., & Sheaffer, B. (1984). Observer reactivity in monitored and unmonitored analogue conditions. *Journal of Psychoeducational Assessment*, *2*(3), 249–255. https://doi.org/10.1177/073428298400200310

Stauffer, J. M., & Buckley, M. R. (2005). The existence and nature of racial bias in supervisory ratings. *Journal of Applied Psychology*, *90*(3), 586–591. https://doi.org/10.1037/0021-9010.90.3.586

St. Peter, C. C., Brand, D., Jones, S. H., Wolgemuth, J. R., & Lipien, L. (2023). On a persisting curious double standard in behavior analysis: Behavioral scholars' perspectives on procedural fidelity. *Journal of Applied Behavior Analysis*, *56*(2), 336–351. https://doi.org/10.1002/jaba.974

St. Peter Pipkin, C., Vollmer, T. R., & Sloman, K. N. (2010). Effects of treatment integrity failures during differential reinforcement of alternative behavior: A translational model. *Journal of Applied Behavior Analysis*, *43*(1), 47–70. https://doi.org/10.1901/jaba.2010.43-47

Suhrheinrich, J., Dickson, K. S., Chan, N., Chan, J. C., Wang, T., & Stahmer, A. C. (2019). Fidelity assessment in community programs: An approach to validating simplified methodology. *Behavior Analysis in Practice*, *13*(1), 29–39. https://doi.org/10.1007/s40617-019-00337-6

Taplin, P. S., & Reid, J. B. (1973). Effects of instructional set and experimenter influence on observer reliability. *Child Development*, *44*(3), 547–554. https://doi.org/10.2307/1128011

Van Acker, R., Grant, S. H., & Getty, J. E. (1991). Observer accuracy under two different methods of data collection: The effect of behavior frequency and predictability. *Journal of Special Education Technology*, *11*(3), 155-166. https://doi.org/10.1177/016264349101100304

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. https://doi.org/10.1901/jaba.1988.21-381

Vollmer, T., Roane, H., Ringdahl, J., & Marcus, B. (1999). Evaluating treatment challenges with differential reinforcement of alternative behavior. *Journal of Applied Behavior Analysis*, *32*(1), 9–23. https://doi.org/10.1901/jaba.1999.32-9

Weinrott, M. R., & Jones, R. R. (1984). Overt versus covert assessment of observer reliability. *Child Development*, *55*(3), 1125–1137.

Table 1

Average Accuracy Coefficient for Each Condition

	Therapist 1 High Fidelity	Therapist 1 Low Fidelity	Therapist 2 High Fidelity	Therapist 2 Low Fidelity
Bianca	96	84	91	96
Quinn	97	91	93	94
Elsa	92	84	94	87
Kelly	95	87	92	92
Jamie	92	86	93	91
Peter	92	89	85	97
Ellowyn	96	88	95	90
Nicole	94	86	95	90
Lola	79	63	82	70

Note. The cells shaded in gray are participants from Experiment 2.

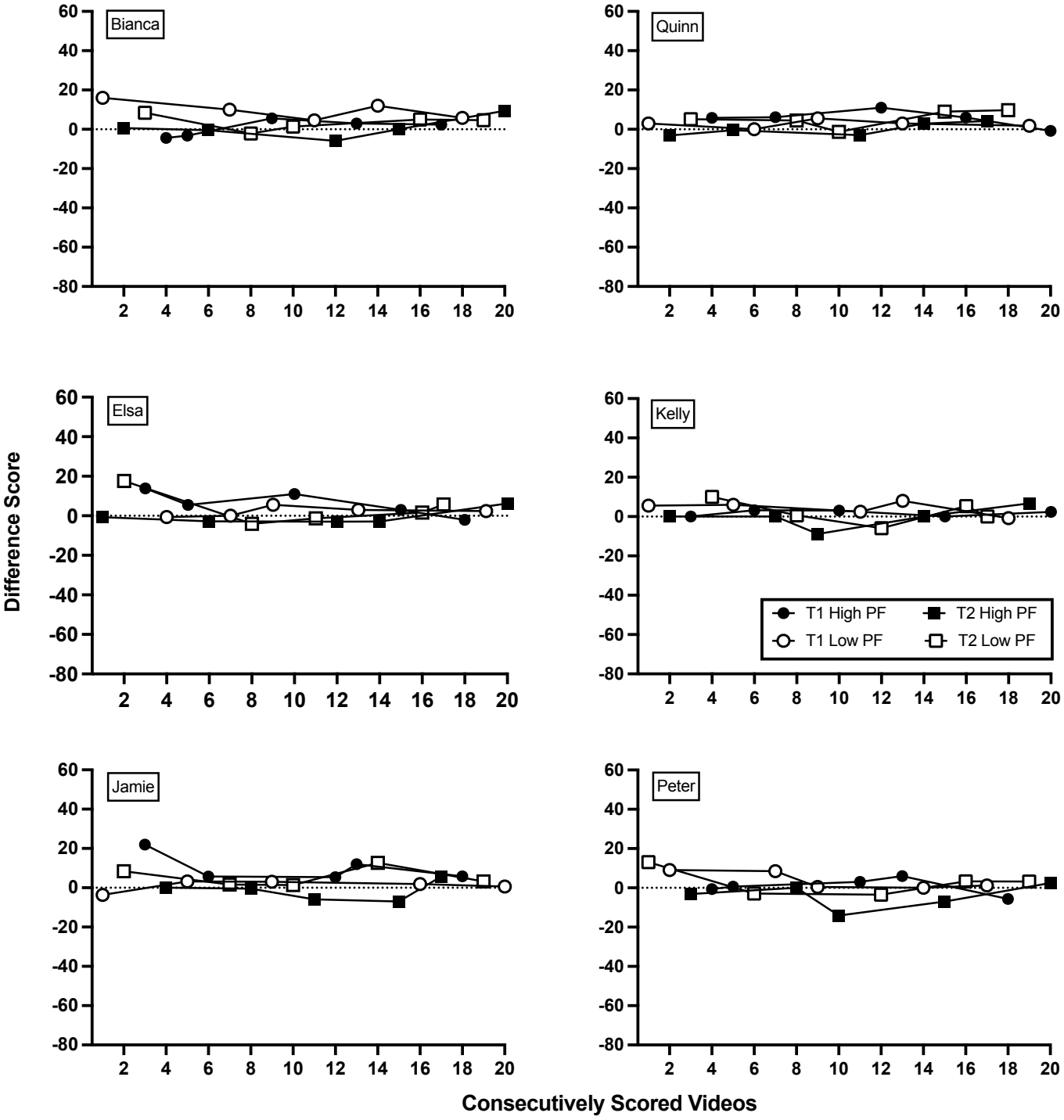
Table 2

Experiment 2 Rating Scale

Component Level	Rating
0	Not visible or not applicable
1	Never implemented appropriately (0%)
2	Implemented competently occasionally, but misses many opportunities (1%-49%)
3	Implemented competently half the time, but misses many opportunities (50%-79%)
4	Implemented competently most of the time, but misses many opportunities (80%-99%)
5	Implemented competently throughout the session (100%)

Figure 1

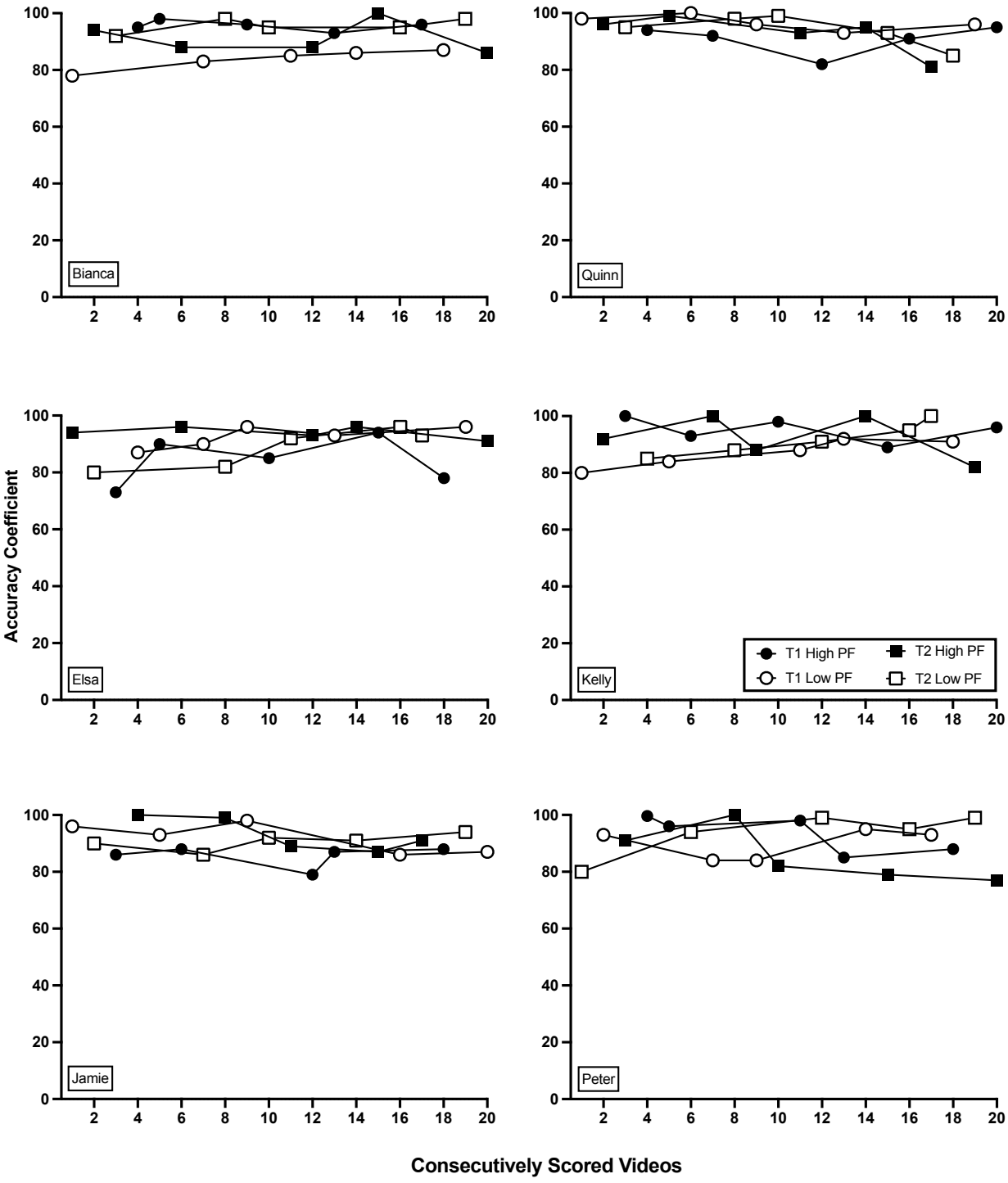
Experiment 1 Difference Scores



Note: T1 = Therapist 1; T2 = Therapist 2; High = 80%, Low = 40%, PF = procedural fidelity.

Figure 2

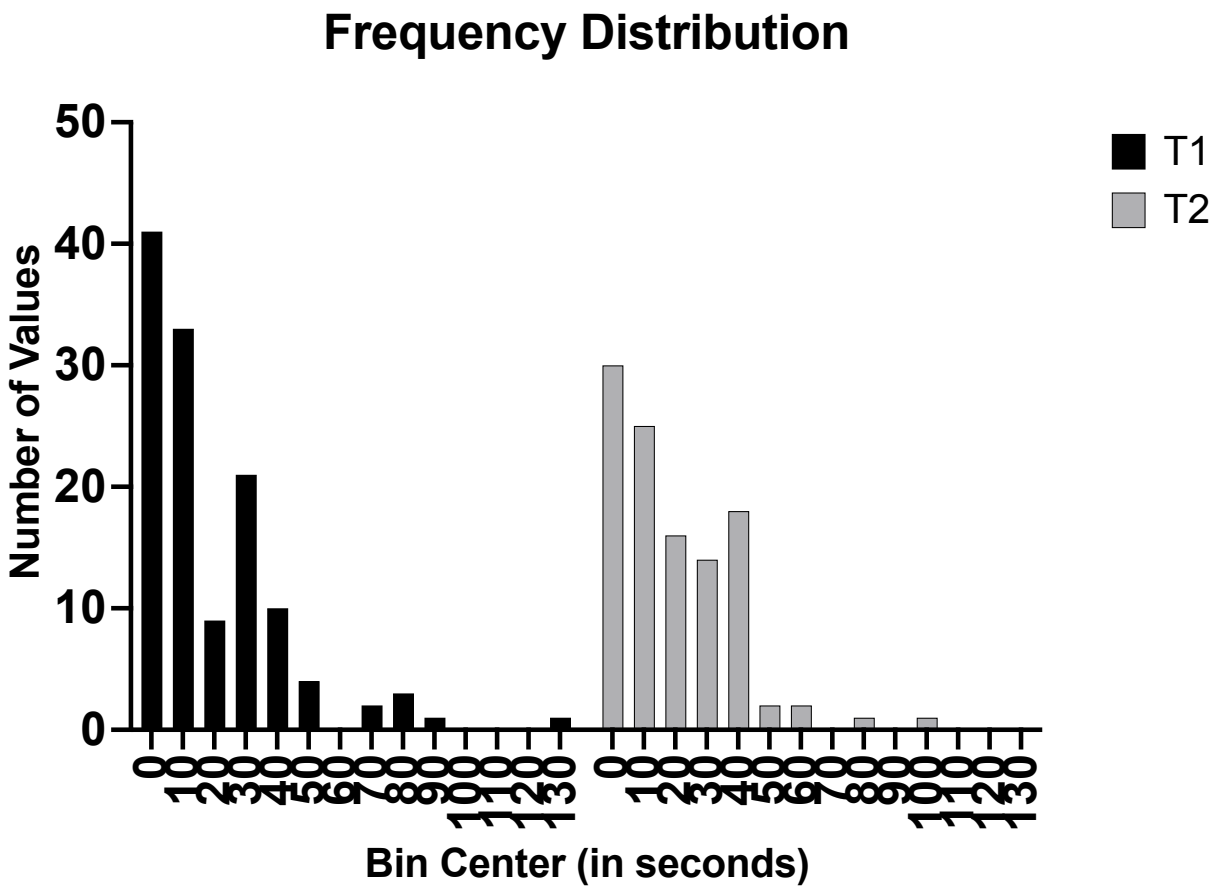
Experiment 1 Accuracy Coefficients



Note: T1 = Therapist 1; T2 = Therapist 2; High = 80%, Low = 40%, PF = procedural fidelity

Figure 3

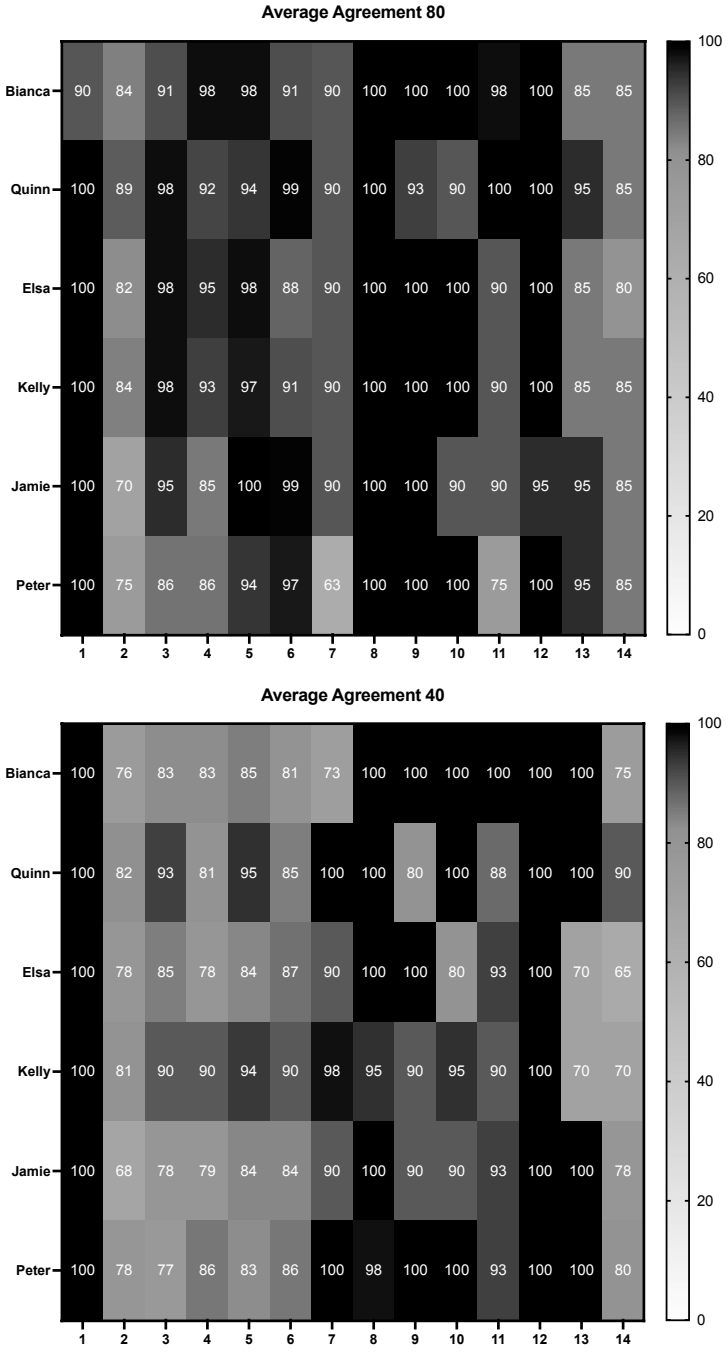
Interresponse Time (IRT) Distribution



Note. T1 = Therapist 1, T2 = Therapist 2.

Figure 4

Experiment 1 Average Agreement of All Components



Note. As the color gradient darkens to black, agreement scores are higher. As the color gradient lightens to white, agreement scores are lower. The numbers inside the boxes indicate mean agreement for that component.

Figure 5

Latency to Finish Data Collection for Each Participant

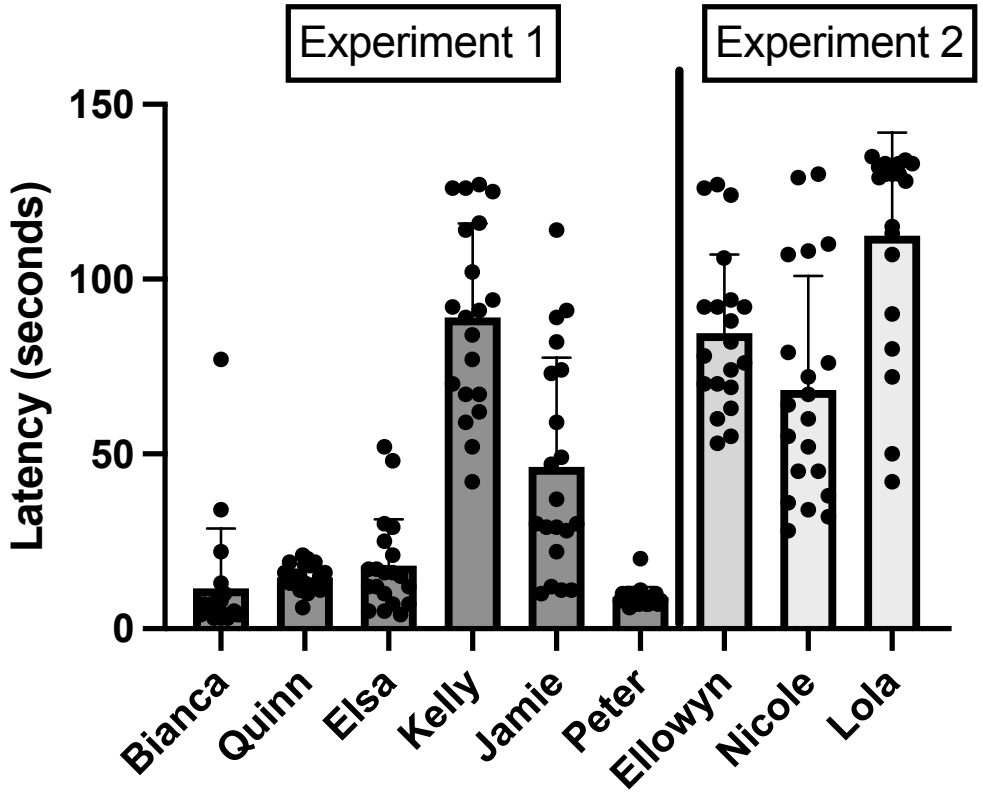
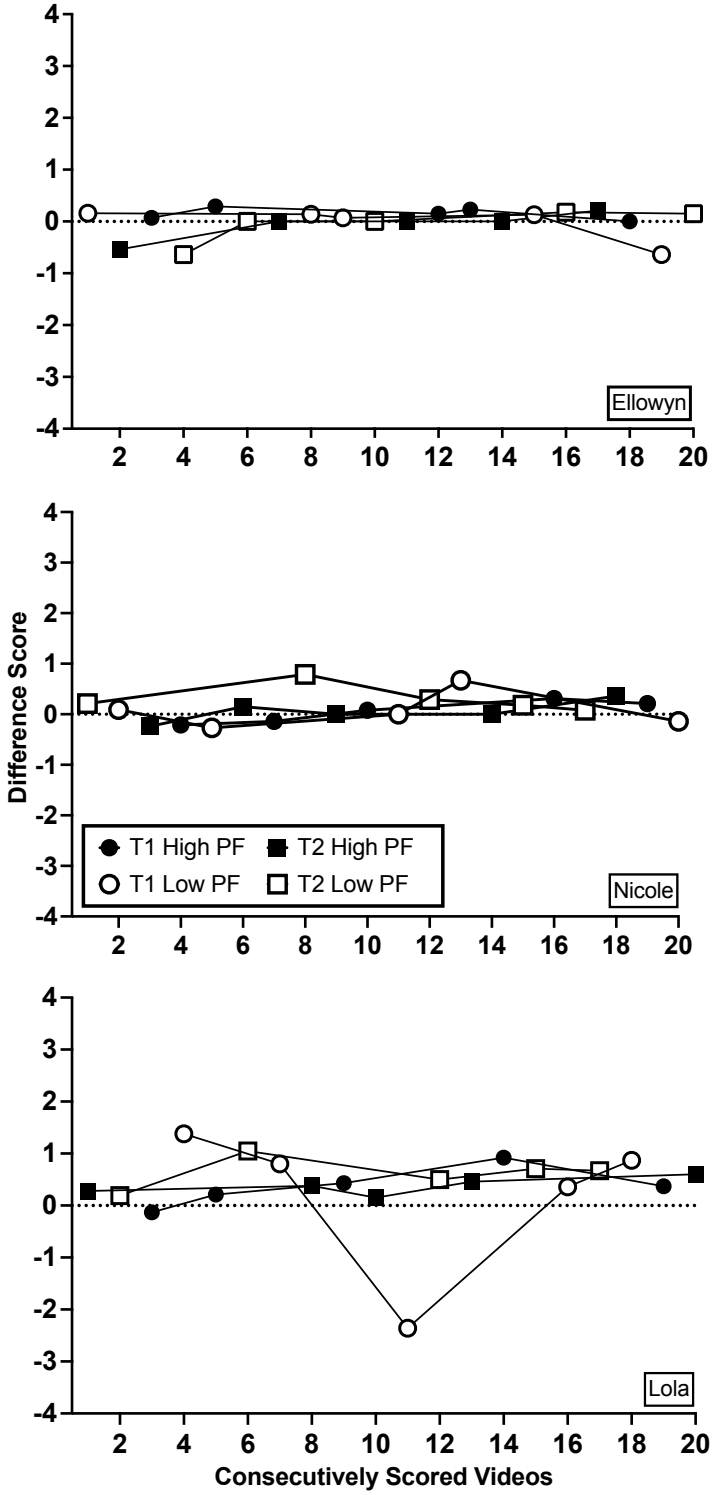


Figure 6

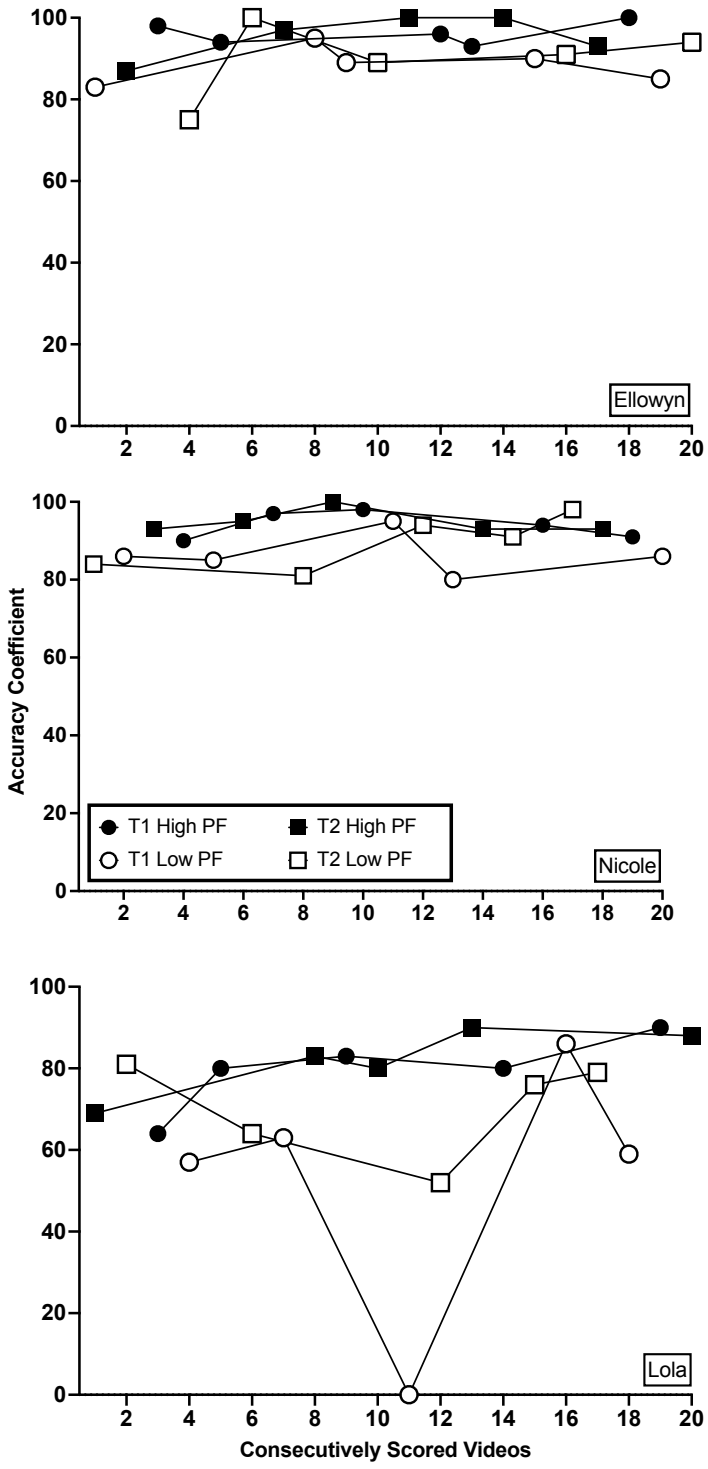
Experiment 2 Difference Scores



Note: T1 = Therapist 1; T2 = Therapist 2; High = 80%, Low = 40%, PF = procedural fidelity.

Figure 7

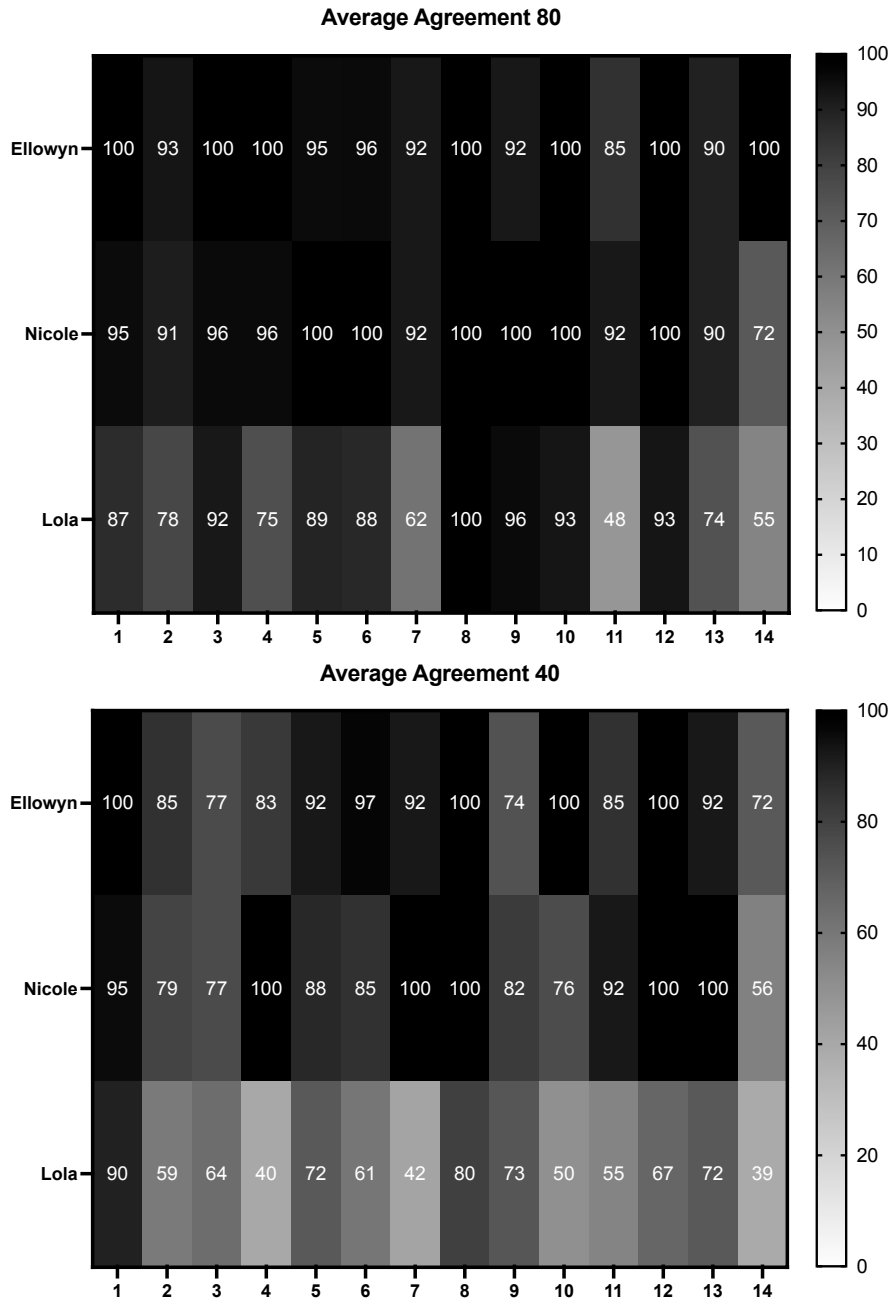
Experiment 2 Accuracy Coefficients



Note: T1 = Therapist 1; T2 = Therapist 2; High = 80%, Low = 40%, PF = procedural fidelity.

Figure 8

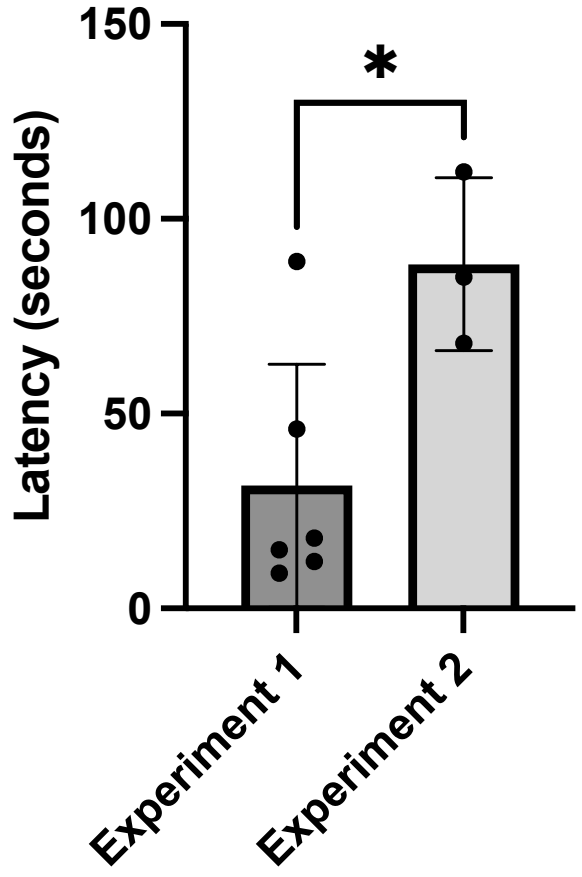
Experiment 2 Average Agreement of All Components



Note. As the color gradient darkens to black, agreement scores are higher. As the color gradient lightens to white, agreement scores are lower. The numbers inside the boxes indicate mean agreement for that component.

Figure 9

Mean Latency to Finish Data Collection by Experiment



Note. Each filled circle is the average latency to finish data collection for one participant.

Appendix

Appendix A

Experiment 1 Procedural Fidelity Checklist

Emma DRO Procedure				
Procedure Component	Correct Implementation (Tally)	Incorrect Implementation (Tally)	N/A	Comments
Says how to earn tokens				
Behavior-Change Procedure:				
Starts/restarts 30-second timer (within 5s) at start of work block or after targeted challenging behavior				
Does not comment about targeted challenging behavior				
Records occurrence of targeted behavior on data sheet within 5s of targeted behavior				
Timer elapses (within 5s):				
Places token on the board				
Delivers behavior specific praise				
Immediately after delivery of the third token (within 3s):				
Labels break period				
Asks Emma to remove and give each token				
Delivers blue iPad after removal of third token				
Starts 30-second play timer				
Makes at least 3 positive statements about behavior during play period (score 1 per play)				
End of play period (within 3s):				
Tells Emma play time is over				
Prompts Emma to put away blue iPad				
Re-prompts and gently removes iPad after 5s if Emma does not put it away				
Count of Target Behavior:				
Protest (tally):		Property destruction (tally):		

Appendix B

Experiment 2 Procedural Fidelity Checklist

Use the following rating scale to evaluate how well the therapist implemented each step of the procedure:

0	Not visible or not applicable
1	Never implemented appropriately (0%)
2	Implemented competently occasionally, but misses many opportunities (1%-49%)
3	Implemented competently half the time, but misses many opportunities (50%-79%)
4	Implemented competently most of the time, but misses many opportunities (80%-99%)
5	Implemented competently throughout the session (100%)

Procedure Component	Rating (0-5)	Comments
Says how to earn tokens		
Starts/restarts 30-second timer (within 5s) at start of work block or after targeted challenging behavior		
Does not comment about targeted challenging behavior		
Records occurrence of targeted behavior on data sheet within 5s of targeted behavior		
Places token on the board		
Delivers behavior specific praise		
Labels break period		
Asks Emma to remove and give each token		
Delivers blue iPad after removal of third token		
Starts 30-second play timer		
Makes at least 3 positive statements about behavior during play period (score 1 per play)		
Tells Emma play time is over		
Prompts Emma to put away blue iPad		
Re-prompts and gently removes iPad after 5s if Emma does not put it away		
Count of Target Behavior:		
Protest (tally):	Property Destruction (tally):	
Overall Fidelity		
1 2 3 4 5		

Appendix C

Behavior Intervention Plan

Emma Behavior Intervention Plan

Background: Emma protests and destroys property during one-on-one instruction. An FBA showed that these responses are maintained by attention and access to preferred items (like an iPad). Emma enjoys chatting with her therapist about her interests and likes to show off her knowledge.

Goal: Emma will use kind words and be respectful to classroom materials during 1:1 instruction for at least 5 min.

Operational Definitions

RESPONSE	DEFINITION	EXAMPLES	NONEXAMPLES
Protesting	Says contrary statement	"No", "I won't", "I'm not doing it" "I don't want to" "I don't like writing"	Statements about difficulty ("this is hard"), saying "I don't know", asking for help
Destroying property	Tears, swipes, shoves, or throws materials	Rips a paper, pushes pencils onto the floor, throws pencils	Crumpling paper, pounding materials on the table

Procedure

1. At the start of 1:1 instruction, tell Emma how she can earn tokens. For example, "If you use kind words and use materials correctly, you will earn a token. Once you earn 3 tokens, we can play on the iPad together."
2. Start a 30-second DRO timer.
3. If Emma protests or destroys property, immediately (within 5s):
    a. Silently reset the DRO timer to 30s.
    b. Do not comment about the targeted behavior.
    c. Record occurrence of targeted challenging behavior on the data sheet.
4. When the DRO timer elapses, immediately (within 5s):
    a. Place a token on the board.
    b. Deliver behavior-specific praise (e.g., "nice job using kind words"; "great job using materials appropriately").
    c. Restart the 30-second DRO timer.
5. When Emma earns 3 tokens, give her a 30-second break immediately after the delivery of the third token.
    a. Label the break period (e.g., "You got three tokens! Let's play on the iPad").
    b. Instruct Emma to remove and hand you each token.
    c. Give Emma the iPad when the third token is removed from the board.
    d. Start a 30-second timer.

e. Makes at least 3 positive statements about behavior during play period (e.g., “you are using the iPad so nicely, great job!” or “you are so good at that game!”). Avoid questions and reprimands.
f. At the end of the break period:
    i. Tell Emma that play time is over.
    ii. Prompt Emma to hand over the iPad.
    iii. If Emma does not independently put away the iPad after 5s, re-prompt and gently remove the iPad.
    iv. Restart the 30s DRO timer.

Appendix D

Demographic Survey

Demographics Survey

General Questions

What is your name? _____

What is your address? _____

What is your BCBA Certification number? _____

What is your date of birth? _____ What is your gender? _____

What is your race? _____ What is your ethnicity? _____

What is your personal annual income?_____

What is your social security number? (this is needed for payment) _____

What language(s) do you speak fluently (list)

_____ _____

_____ _____

_____ _____

BCBA Questions

Is your primary employment as a BCBA? _____ Yes _____ No (list job title: _____)

What is/was your primary work as a BCBA? (check all that apply)

- ☐ Behavioral treatment of autism and other developmental disabilities
- ☐ Organizational behavior management
- ☐ Brain injury rehabilitation
- ☐ Behavioral gerontology
- ☐ Clinical
- ☐ Education
- ☐ Behavioral sport psychology
- ☐ Prevention and behavioral intervention of child maltreatment
- ☐ Behavioral treatment of substance use disorders
- ☐ Environmental sustainability
- ☐ Health and fitness
- ☐ Behavioral pediatrics
- ☐ Organizational behavior management
- ☐ Supervision
- ☐ Other; please specify: _____

What populations have you worked with as a BCBA? (check all that apply)

- ☐ Intellectual and developmental disabilities
- ☐ Social/emotional disorders
- ☐ Autism spectrum disorder
- ☐ Brain injury
- ☐ Industry professionals (e.g., health care, human services, education, government, nonprofits, retail, etc.)
- ☐ Mental disorders
- ☐ Athletes/Sports teams/Coaches
- ☐ Substance use disorders
- ☐ Seniors
- ☐ Other; please specify: _____

What age groups have you worked with? (check all that apply)

- ☐ Children (age 12 and younger)
- ☐ Adolescents (13 – 17)
- ☐ Adults (18 – 65)
- ☐ Older adults (65 and older)

How familiar are you with a resetting DRO procedure? 1 2 3 4 5
(not familiar) (very familiar)

Supervision Questions

Have you supervised or trained other people in the last 12 months? (check one) ___Yes ___No

If yes, who do you supervise or train? (check all that apply)

- ☐ Registered behavior technician
- ☐ Direct care staff
- ☐ Board certified assistant behavior analyst
- ☐ Teacher
- ☐ Industry professional
- ☐ BACB supervisee
- ☐ Parent/legal guardian
- ☐ Other; please specify: _____

How long have you been supervising?

- ☐ Less than 3 years
- ☐ 3 – 5 years
- ☐ 6 – 10 years
- ☐ 11 – 15 years
- ☐ 16 – 20 years
- ☐ More than 20 years

What kinds of supervision activities do you typically do?

Procedural-Fidelity Questions (*"Procedural fidelity" refers to the extent to which procedures are implemented as described.)*

Do you think collecting fidelity data is important? ___Yes ___No

What type of training have you received on procedural-fidelity? (check all that apply)

- ☐ Graduate training
- ☐ Webinar
- ☐ Workshop
- ☐ On-the-job training
- ☐ In-service training
- ☐ Self-study
- ☐ Other; please specify: _____

When was the date of your last training?

- ☐ Within the past calendar year
- ☐ 3 years

- 5 years
- Longer than 5 years

With what frequency do you collect fidelity data?

- Less than once a year
- More than once a year; less than once a month
- More than once a month; less than once a week
- More than once a week

How often do you collect fidelity data when you see your supervisee?

- 0% of supervision sessions
- 10 – 30% of supervision sessions
- 40 – 60% of supervision sessions
- 70 – 90% of supervision sessions
- 100% of supervision sessions

When do you choose to collect fidelity data? (check all that apply)

- I don't collect fidelity data
- During training
- With novel implementors
- To check-in after a specified amount of time
    - Please specify the amount of time: _____

How do you collect fidelity data? (check all that apply)

- Direct observation
- Self-report
- Interview
- Permanent product
- Other; please specify:

  _____

Appendix E

Experimenter Fidelity Checklist

Experimenter reviews 100% video of Therapist 1
Yes ☐ No ☐
Number of questions asked (tally):

Experimenter reviews 100% video of Therapist 2
Yes ☐ No ☐
Number of questions asked (tally):

Participant questions are answered using only information from consents, scripts, or BIP documents, or by a statement like "do your best" or "I can't answer that"
Yes ☐ No ☐ N/A ☐

Participant questions about BIP or checklist are not answered during data collection
Yes ☐ No ☐ N/A ☐

Experimenter enters room 2-2.25 min from end of video or within 15s of when participant rings bell (score "no" if an error occurs at any point)
Yes ☐ No ☐

Order of videos
Video 1:_____Video 5: _____Video 9: _____Video 13: _____Video 17: _____

Video 2: _____Video 6: _____Video 10: _____Video 14: _____Video 18: _____

Video 3: _____Video 7: _____Video 11: _____Video 15: _____Video 19: _____

Video 4: _____Video 8: _____Video: 12_____Video 16: _____Video 20: _____

All videos were correct
Yes ☐ No ☐

Time Bar on Videos

Video 1: Yes ☐ No ☐ Video 5: Yes ☐ No ☐ Video 9: Yes ☐ No ☐ Video 13: Yes ☐ No ☐ Video 17: Yes ☐ No ☐

Video 2: Yes ☐ No ☐ Video 6: Yes ☐ No ☐ Video 10: Yes ☐ No ☐ Video 14: Yes ☐ No ☐ Video 18: Yes ☐ No ☐

Video 3: Yes ☐ No ☐ Video 7: Yes ☐ No ☐ Video 11: Yes ☐ No ☐ Video 15: Yes ☐ No ☐ Video 19: Yes ☐ No ☐

Video 4: Yes ☐ No ☐ Video 8: Yes ☐ No ☐ Video 12: Yes ☐ No ☐ Video 16: Yes ☐ No ☐ Video 20: Yes ☐ No ☐

All videos had no time bar

Yes ☐ No ☐

Prompts for break after every 4th video (score "no" if an error occurs at any point)
Yes ☐ No ☐

Break start time: Break end time:
Break start time: Break end time:
Break start time: Break end time:
Break start time: Break end time:

Experimenter reviews graphs with participant
Yes ☐ No ☐ N/A ☐

Experimenter gives participant list of resources
Yes ☐ No ☐ N/A ☐