4-26-2013

The Effects of Rating Accuracy on Treatment Fidelity

Joshua LaBeff
Western Michigan University, joshua.d.labeff@wmich.edu

Assessing the Relationship Between Rating Accuracy and Treatment Fidelity in the Context of Functional Analyses

Joshua D. LaBeff

Undergraduate Honors Thesis

Western Michigan University

Contents

Abstract

The current study attempts to evaluate if there is a relationship between treatment fidelity and interobserver agreement. Participant performance on rating the accuracy of a video demonstration of a functional analysis session was analyzed in comparison to their performance in implementing a mock functional analysis session. Video models (some with confederates and some with the actual participant) of functional analysis conditions were shown to and graded by participants and were immediately followed by their participation in a mock functional analysis session with a confederate child. The results of the study tend to show little to no relationship between the two variables.

Keywords*: Treatment Fidelity, Interobserver Agreement, Rating Accuracy, Functional Analysis, Video Modeling*

The Effects of Rating Accuracy on Treatment Fidelity in the Context of Functional Analyses

In general the rationale of experimentation is to inspect or examine relationships between variables. One aspect that makes experimentation particularly unique is that it attempts to make an understanding of the direct influence from one variable (independent variable) on another (dependent variable). Typically experiments only attempt to evaluate a relationship with the use of one variable, or a small quantity of variables, in order to keep results from being vague or indistinguishable. Experiments and studies seek to isolate the influence from only the independent variable so that that influences from other factors do not play a part in the results of the experiment. Therefore, the better the experiment is designed, the more likely experimenters will be able to discard alternative explanations of their results. In a perfect experiment, only one explanation of the results during experimentation would be possible, specifically that the independent variable would account for any changes seen to the independent variable(s) (Kazdin, 1982).

With any type of research approach, researchers must do their best to exclude alternative explanations to the outcome of the study in order to be confident that the independent variable was in fact the variable responsible for the research findings. The degree to which researchers assess and manage for outside or confounding variables help determine the level of confidence that one may have for the results of the study. One way researchers can help in assuring control for confounding variables is to assure that the protocols described are consistently followed with precision throughout the entire duration of the study (Gast, 2009).

Treatment Fidelity

The results and outcomes of a study can be unclear if the independent variable has been implemented inconsistently. Often, practitioners and researchers alike strive to ensure that the

independent variable is implemented exactly as planned to avoid the influence of extraneous variables that may have been accidentally administered at the time of treatment. The term treatment fidelity refers to the extent to which the independent variable has been implemented or carried out as planned (Cooper, Heron, & Heward, 2007).

Low levels of treatment fidelity may allow a plethora of additional variables to play a part when measuring change in an experiment. Oftentimes these additional variables make it near impossible to confidently interpret results. Major problems when interpreting data can occur if an intervention has been employed inconsistently, inaccurately, or at inappropriate times. This can lead researchers or practitioners to report a false positive conclusion, in which a claim is made that some type of relationship between the two variables exist when in fact there is no such relationship. There is also the possibility of a false negative, coming to the conclusion that there is no relationship between two variables when in fact there is a relationship. If a relationship or a lack there of, has been identified, one cannot be sure whether or not said relationship can be attributed to the effects of treatment or an extraneous variable that has not been accounted for due to the lack fidelity in which the intervention has been delivered (Cooper, Heron, & Heward, 2007).

There are many threats to treatment fidelity in the applied setting. One of these threats is treatment drift. Treatment drift takes place when the delivery of the independent variable changes over the duration of a study to the extent that the independent variable being applied at the end of the study is different than that of the independent variable implemented at the beginning of the study. There are many reasons treatment drift can occur over the course of a study, periodically this can be because practitioners either forget the protocol over the course of the study or become fatigued when implementing treatment. Treatment drift can occur if an

experiment has a complex independent variable. This is due to the fact that it can be difficult for individuals to apply each of the elements needed for the experiment with consistency across the entirety of the study. Treatment drift may also be caused by additional contingencies influencing the individual(s) that are responsible for the implementation of the independent variable. For instance, if the implementer(s) (e.g., teachers, parents, and school staff) find it too difficult to implement certain portions of the protocol, these aspects may get implemented incorrectly or not at all. Likewise, some components of an intervention may be more favorable or easier to implement, and these aspects may get over-implemented or consistently applied in lieu of components that are arduous (Cooper, Heron, & Heward, 2007).

If subjects in the study are being exposed to multiple treatment interventions during the time of treatment this can also hurt the study's treatment integrity. During the course of investigation only one treatment intervention should be implemented. If multiple similar interventions are being exposed to the subject(s) of the study it can become too difficult to determine what particular intervention or combination of interventions is or is not actually being measured. Researchers must always strive to only expose subjects of the study to only one treatment intervention in order to lessen the chance of there being multiple effects on the change in the subject's behavior (Gast, 2009).

Another threat to treatment integrity in a study is unclear behavioral definitions or interventions. Clearly defining an operational definition of the treatment procedures is also an essential part of experimentation. If the definition of the treatment procedures is not clear and precise the person implementing treatment may misinterpret the way treatment should be given. Other than training the individual that will be applying the treatment and determining a high level of competency in which they are able to implement treatment, having a clear operational

definitions of treatment conditions are required in order to meet the technological requirements of applied behavior analysis (Baer et al., 1968). Failure to supply a definition of the treatment protocol that is unambiguous, precise, and clear hinders application and proper execution of the treatment in the future. An experiment in which the treatment protocol has not been operationally defined also makes it difficult for other researchers to replicate the experiment and validate claims that experimenters may have made (Cooper, Heron, & Heward, 2007).

Having operationally defined treatment condition (e.g. a treatment definition that is unambiguous, clear and precise) also makes it much easier for data collectors to record data. Throughout the course of a research study, the participant's behavior is constantly being observed and measured under a minimum of two different conditions, a baseline condition and an intervention/treatment condition. To make certain that data recording and collection is less likely to be misinterpreted, independent observers frequently check the accuracy in which they are scoring fidelity. During this time secondary observers will measure treatment fidelity using the same definitions and recording procedures used by a primary observer of the study. This comparison of data is most commonly shown as a percentage (Gast, 2009).

Inter-Observer Agreement

Any time that human observers are involved in making direct observations, there is a possibility that behavior may not be recorded consistently. Recorders need to determine whether or not a target behavior has actually occurred so it will not be accidentally looked over, or documented incorrectly. It is imperative that agreement amongst observers is high when collecting data. If observer agreement is high it can help in assuring that any variation or inconsistency between observer reports remain at a minimum (Kazdin, 1982).

"Interobserver agreement (IOA) is the most commonly used indicator of measurement quality in applied behavior analysis. IOA refers to the degree to which two or more independent observers report the same observed values after measuring the same events" (Cooper, Heron, & Heward, 2007, p.113). There are a great deal of benefits and uses to an experiment from measuring IOA. Among those many benefits, calculating IOA can allow those implementing the study to account for any observer drift. By continuously comparing ratings throughout the study researchers can check to see if any of the observer's definition of the target behavior has drifted away from what it had been in the beginning portion of the study. One of the observers may be using an alternate meaning of the target behavior than the rest of the observers. Measuring IOA does not allow the researchers to determine which of the observer's definition of the target behavior has drifted, or any reasons behind the observer disagreement. However, it can help to determine when disagreements are being made, can provide insight as to when a problem in data collecting has occurred, or perhaps when retraining of data recorders would be appropriate (Cooper, Heron, & Heward, 2007).

Another reason why it is important to have agreement between independent observers would be it often times is indicative of whether or not the target behavior has been clearly defined. IOA on behavioral observations is a good way to assess whether or not the target behavior or protocol is empirically and operationally defined. Having high levels of IOA also helps the individuals who are implementing the intervention. Kazdin, (1982) states "If observers readily agree on the occurrence of the response, it may be easier for persons who eventually carry out an intervention to agree on the occurrences and to apply the intervention (e.g., reinforcing consequences) consistently" (p. 49). Having a high percentage of IOA insures that the definition of the target behavior was clear and concise, and the way in which observers were

instructed to measure the behavior was not too difficult to do. If the target behavior is not clear enough, it can easily be the case that observers will be likely to miss or mistakenly include instances of behavior.

Lastly, having a high level of IOA insures that fluctuation in the data can be attributed to changes in behavior, and not the observer that happened to be recording at the given time. Having high levels of IOA helps the trustworthiness of the study being conducted, and allows other researchers to judge the relative believability of the data presented (Cooper, Heron, & Heward, 2007). When a high percentage of IOA was attained in most (if not all) instances, a stronger assurance (or believability) of the reported data is suggested.

There are also certain standards in which observers should adhere to when acquiring observations to complete and calculate IOA. Typically, assessments of interobserver agreement should meet the criteria of using the same system of measurement, measure the same events, and observers must be independent of one another. Although these standards may seem relatively apparent, it is still crucial that data collectors do their best to adhere these three criteria.

Observers must be measuring and recording data with the same measurement system. Observers need to be required to use the same definitions of target behaviors, same observational procedures and codes, and should be using the same type of measurement tools and devices (e.g., stopwatches, counters). In addition to using equivalent measurement systems any observer's recordings being used to measure IOA, should be trained in the same manner as the primary observer, and should demonstrate a high level of proficiency when recording instances of behavior (Cooper, Heron, & Heward, 2007).

During recording sessions, observers should be recording data from the same participant(s) at the same recording interval. This is especially true if observations are happening

in real time. Data recorders sitting on opposite sides of the room may gather different results due to the different vantage points of the participant(s) during the time of the target behavior. Data intervals among recorders should start and stop at the exact time. Even in a matter of a few seconds, an observer may potentially miss or not have the opportunity to record a crucial piece of data or interaction by the participant(s). Often during this time data recording strategies are used to avoid problems of timing. For example, the use of synchronized timing devices, having one of the observers signal the start and stop of each interval, or simply predetermined starting and stopping time for the duration of the observation interval (Cooper, Heron, & Heward, 2007) may be used to help facilitate accurate recording between two or more observers.

When evaluating the percentage of IOA for a behavior that has a permanent product(s), separate observers need not measure the behavior at the same time. For example, if data recorders were to observe and report data from the same video showing the implementation of treatment, at separate times. Independent recorders would just need to be sure that the video(s) started and stopped at the same times when viewing and scoring the footage in order to make certain that they had each measured behavior from the exact points in the clip (Cooper, Heron, & Heward, 2007).

Observers recording data and measuring the target behavior of the study should also be independent of one another. If observers are able to prompt one another to score in a certain way it can affect the way they record and measure data. In order to ensure that neither of the observers can prompt one another, they should be situated in a way in which they cannot hear or see when the other observer is recording instances of the target behavior. Meaning, observers should not be seated so close to one another that one observer can influence the way in which the other observer is recording. However, this is not as much of an issue when dealing with

observers recording data from video clips of participant performance since the same video can be viewed at different times (Cooper, Heron, & Heward, 2007).

There are many ways in which to measure the percentage of IOA between independent observers. Point-by-point agreement ratio is one of the more typical for computing whether agreement has been reached between data recorders. Point-by-point agreement ratio is generally used when measuring instances of discrete opportunities (e.g., trials or intervals) in which behavior is occurring, or not occurring, when behavior is present or absent, or occurring at appropriate or inappropriate times. IOA looks at total agreement (observers agree at each interval) and does not take into account the presence or absence of behavior during the calculation. Agreements are defined as occasions in which observers recorded the same thing. As long as both recorders record the same occurrence, or non-occurrence of the target behavior an agreement has been made. Conversely an example of a disagreement would be considered as in instance in which one observer recorded an occurrence of a target behavior and the other observer did not record the same occurrence. Hence, agreements and disagreements are then compared on a point-by-point basis (Kazdin, 1982). Observer scoring can be compared directly to determine whether or not both data recorders have recorded a particular target behavior's occurrence. The way in which point-by-point agreement is computed is by dividing the total number of agreements made (from the independent observers) by the total number of agreements plus the total number of disagreements, multiplied by one hundred to obtain a percentage of agreement.

Video Modeling

Alvero and Austin (2004) evaluated the relationship of individuals observing and grading video models of others properly implementing safety office techniques (i.e., proper posture while

typing, lifting, sitting) on the individuals' ability to implement the same safety techniques without the use of feedback from the practitioners. During the course of the study a fidelity grading sheet was used to evaluate the implementation of the safety protocol by the participants. The fidelity sheets were used both to record fidelity for the participants implementation of the safety protocol as well as the participants to grade the video models of the office safety techniques.

Throughout the Avero and Austin (2004) study participants were divided into groups, one of which was given specific instructions on the correct employment of said safety office protocol. During that time the other half of the participants did not receive any type of written or verbal protocol. After both groups performed the office safety behaviors in a controlled office setting, participants then viewed and rated videos models of individuals who were implementing the same procedures. Alvero and Austin (2004) then reported that participants that completed video ratings and were provided written safety protocol, exhibited more instances of office safety behavior than those individuals whom were not provided any written instruction and had not graded the fidelity of video models performing office safety techniques. Alvero and Austin (2004) proposed that only after being given the proper protocol, observing video models of others performing the correct protocol, and rating the fidelity of the protocol (in the video models) did increases in treatment fidelity occur among participants. An important point to make is that Avero and Austin (2004) failed to report the level of accuracy in which participants graded the office safety video models. A questions that may arise from this is whether or not the accuracy with which the participants rated the video models has any relationship with the level of fidelity in which the participants implemented the protocol themselves when subsequently asked to perform these behaviors?

It can be somewhat taxing to train individuals to acquire certain skills such as safety office techniques and behaviors. Due to the fact that Alvero and Austin (2004) reported their subjects having an increase in safety office behaviors, other research has been conducted to determine whether or not other taxing skills can be acquired in a similar manner. One such taxing skill being trained in a similar way that Alvero and Austin (2004) trained safety office behaviors is the implementation of functional analysis conditions.

Functional Analysis

A functional analysis is a methodology used to aid in the identification of the contingencies that are accountable for a behavior in a given condition or situation. For any particular behavior of interest there are multiple environmental antecedents and contingencies occurring simultaneously that may influence or control the behavior. A functional analysis attempts to determine exactly which antecedents and/or contingencies are maintaining a specific behavior. This is accomplished by manipulating the environmental variables that may be related to the behavior of interest (Cooper, Heron, & Heward, 2007).

Typically there are four conditions that are tested during the course of a functional analysis. The four conditions are the unstructured play condition, attention or social disapproval condition, the demand condition, and the tangible condition. The first of the four conditions being an unstructured play condition, (play control) in which non-contingent attention is given to the patient, only after a set amount of without the occurrence of the behavior. The second of the four conditions being tested consists of an attention or social disapproval condition. In this condition, attention is given (for an allotted amount of time) to patient by the practitioner contingent on the target behavior. The third condition being the demand condition, in this condition, the patient is given a continuous amount of prompted tasks to complete throughout the

condition. The breaks from the prompted tasks then are contingent on the individual exhibiting the target behavior. The last condition being the tangible condition, during this time of the functional analysis session, a tangible reinforcer is given to the patient contingent of the target behavior.

There is new research being conducted as to how participants will fair with the absence of direct feedback from the instructor during the training and implementation of the functional analysis. In the Field (2013) study, students were exposed to and graded videos of a functional analyses condition and immediately afterwards conducted a mock functional analysis condition of their own with a confederate child. The Field (2013) research was very similar to the research of Alvero and Austin (2004) in the sense that both studies involve training in similar manor. The use of participants viewing and grading video models, and the subsequent measurement of participant fidelity pertaining to given protocol were implemented in both studies. The Field (2013) study as well as the Alvero and Austin (2004) study also reported participants treatment fidelity (when implementing protocol) increasing throughout the treatment. Both studies attributed the change in the participant's behavior to the given written protocol, and the viewing and grading of the video models. However, in both experiments the participant's accuracy in which they graded said video models was not measured.

During the first session of the Field (2013) study participants a probe phase was conducted to determine how well they could implement the functional analysis protocol without the use of any type of video model or written procedures. The purpose of this was to attempt to measure the level of competency each participant had prior to receiving the written protocol or any additional intervention (i.e., video models). Thus, baseline commenced following this probe phase, prior to each functional analysis implementation with a confederate child, participants

were given specific written protocol as to how each of the functional analysis conditions were to be implemented. Participants (prior to running mock functional analysis) also viewed one of the six video models for each the four functional analysis conditions (i.e., unstructured play, attention, demand, and tangible).

Video models consisted of a two-minute clip showing each functional analysis condition being implemented. Each condition had six separate two-minute video models. Thus, the selected video for a given functional analysis condition was selected at random.. Each of the video clips demonstrated 100% accurate implementation of the functional analysis condition according to the written protocol.

During the treatment phase of the experiment each participant would then grade one of the functional analysis conditions with a specific fidelity grading sheet that was provided for them (Appendix A). The participants were only instructed to grade one of the conditions at a time; that is they were only given one of the grading sheets in (Appendix A) and instructed to grade the corresponding condition. Determining the treatment phase that each participant was given was based upon their performance in the baseline phase. Some of the participants of the study were also instructed to grade videos of themselves rather than just video models. Self videos included a two minute clip of the participant performing a functional analysis condition. The videos of participants themselves were the same functional analysis condition in which participant was already receiving treatment (i.e., the original video model). Each of the self videos contained a minimum of at least one error as according to the functional analysis protocol. The reason that a participant was switched to grading self videos was due primarily to the participant not responding well to the particular treatment condition.

It was through this grading of the video model (or video self) functional analysis conditions by the participants that the IOA from this research was used. Immediately following the watching of all four functional analysis conditions (one of which they had graded for treatment fidelity), the participants entered a small room across the hall where they implemented the mock functional analysis session with a confederate.

This research will attempt to make an understanding of any potential relationship between scoring accuracy (as measured by the IOA between the participants and primary observer grading of functional analysis video models and/or self videos) and the fidelity with which individuals implemented the functional analysis conditions during the mock functional analysis. There are potential relationships that could exist between rating accuracy and treatment fidelity. They could have a positive relationship, in which as change on a graphical analysis would show the variables change in the same direction. The two variables could have a negative relationship in which these two variables would change in opposite directions on a graphical analysis. There is also a possibility that the two variables may have not real relationship at all.

Covariance refers to the extent to which two variables change together (Snedecor, & Cochran, 1980). If one variable changes and the second variable changes in the same direction, the variables are said to have a positive covariance. An example of positive covariance would be the change in speed of a race horse and the change in the likely hood the horse will win the race. As changes in horse's speed increases the changes in its chances of winning also increase. Likewise, as the changes in speed of the race horse decrease so do the likely hood of the horse winning the race.

When one of the variables changes in and the other variable then changes in the opposite direction the two variables are said to have a negative covariance. An example of negative

covariance would be the change in a jockey's weight and the change in the speed of a race horse. As the change in weight of a jockey on a race horse increases, the speed in which the horse runs will decrease, and as the weight in the jockey decreased the speed of the race horse would increase.

The two variables may show positive covariance in such a way that as the percentage of rating accuracy increased the fidelity with which the participant implemented the treatment protocol with fidelity also increases. The two variables may be negatively collated; as the percentage of rating accuracy increased the percentage of treatment fidelity would decrease. However, there is also the potential that there will be no relationship between the two variables; meaning that there is no clear relationship between rating accuracy and the fidelity in which the individual implements treatment. The goal of this paper is to evaluate a potential relationship between the accuracy of rating a video and one's ability to implement the functional analysis procedure. In particular, this study will examine the relationship between the rating accuracy of the participants in the Field (2013) study (pertaining to the functional analysis video models and/or self videos), and the subsequent fidelity implementation of functional analysis conditions.

Method

Participants

The participants involved were 15 graduate level students at Western Michigan University. Students were enrolled in a course to teach them techniques to analyze and assess problem behavior. The age range of the students was from 18 to 30. Student participants that took part in the Field (2013) study received class credit for their participation. Participants enrolled in this class had either a bachelor's degree or a master's degree. It was possible that

some students enrolled in the study had experience with implementation of functional analysis protocol.

Setting

The study took place on the campus of Western Michigan University. All participant sessions were graded in a research lab room; videos were watched in a different lab room across the hall from which the participant sessions were being implemented. All grading and calculating was done in an isolated lab room on a research computer.

Materials

Materials included a camera to record each of the participant's sessions, a variety of two-minute video models of each functional analysis condition; each condition had six different two-minute video models. Each video model had been preformed with 100% accuracy (according the written protocol). Functional analysis scoring sheets were used by the participants to score video models and were also used by the primary observer to score treatment fidelity (Appendix B) for participants. Each condition of the functional analysis had its own specific scoring sheet (Appendix B) that was to be used during the time of grading treatment fidelity. Self-videos were also used (for specific participants only) in order for them to view and grade for treatment fidelity. Participants self videos were graded for fidelity by the primary observer of the study in order to later compare point-by-point IOA with individuals scoring said self videos. Participants pertaining to data sets 3, 5, 6, 7, 10, 11, and 14 were the participants instructed to grade self functional analysis videos.

Dependent Variable

There was no dependent variable for this study. This is study is a data analysis of the Field (2013) study. However, the variables of interest in the study were the way in which the

specific protocol was implemented. Such variables included, only the relevant materials on the table at the starting time of the each condition, the length and timing of interventions (e.g., how long and when attention or object was given to the confederate), the way in which attention, demands and objects were provided, the amount of prompting during the session, and the length of each condition. The variables recorded were based on protocol as to when and how to react to the target behavior elicited by the confederate of the study (Appendix A).

Measurement

Each video was given its own scoring sheet, with specific check boxes to be completed based on the participant's performance (Appendix A). Each participants video's fidelity was measured by assessing how many components of the protocol were preformed correctly divide by the number of components scored in the condition. Each condition in each session had its own corresponding percentage of fidelity. The fidelity score for each participant's specific session condition was used to compare to the accuracy rating that was conducted by participants immediately prior using the video models and/or self-videos. A graphic analysis was then made to determine if any differential effect in the impact of the intervention occurred. This was done by assessing the how accurately each participant graded the video model (and/or self video) compared to fidelity in which they implemented the functional analysis protocol.

Inter-observer Agreement

The primary observer used each of the fidelity grading sheets (Appendix B) to grade each participant's treatment fidelity in all of the functional analysis conditions. A second observer scored twenty four percent of sessions for IOA and a point-by-point agreement formula was used to calculate the IOA (Kazdin, 1982). When IOA between the primary and secondary observers dropped below 80% pertaining to one of the participant's functional analysis conditions both

observers were then retrained as to how to grade the condition in question. Retraining involved both observers reviewing grading criteria and regarding the condition in question.

Participant accuracy ratings for the video models and/or self-videos were assessed by the primary observer. While grading a video model of a functional analysis condition, each participant scored the condition using one of the fidelity grading sheets (Appendix B). Point-by-point agreement (Kazdin, 1982) with the primary observer was then used to determine how accurately each of the participants had actually measured the video model or self-videos. No IOA was measured with the primary observer when measuring the rating accuracy of the participants.

Procedure

Participant's mock functional analysis videos were graded using the condition fidelity grading sheets (Appendix B). Target criteria of each condition include, the time and length in which attention was given, the way in which attention was given, usage of correct materials, the length of the condition, amount of prompting, and whether or not non-target behavior had been ignored.

Once each of the participant's videos had been scored, a percentage of accuracy was calculated. Participant's treatment fidelity was measured by the number of correct components preformed in each condition (according to the fidelity sheets) divided by the number of correct and incorrect components, this quotient was then multiplied by 100 to become a percentage. Percentages of participants treatment fidelity were then added to an excel spreadsheet.

When scoring participants rating accuracy of model and self-videos, accuracy was measured using point-by-point IOA compared to the primary observer (Kazdin, 1982). As discussed earlier all video models showed models implementing with 100% accuracy as

according to the functional analysis protocol. Therefore disagreements were measured by any time the participant had marked a component of the grading sheet as incorrect. Self-videos were first graded for fidelity by a primary observer and then IOA was measured using the participants ratings compared to the primary observers' score. The percentage of IOA was then also added to an excel spreadsheet.

Design

No research designed was implemented during the course of the study. This research aimed to show the comparison between the variables of rating accuracy and the subsequent fidelity of the implemented functional analysis protocol.

Results

Data was imported to an excel spreadsheet and was used to make a comparison line graph in the excel program comparing both data sets (i.e., accuracy and fidelity ratings) for each condition of treatment. Line graphs compare both variables (treatment fidelity and rating accuracy) for each participant pertaining to the particular condition in which they had graded a model or self video. Orienting to each of the data set graphs (Appendix B), the Y-axis is a representation of two percentages, the first being the percentage of fidelity with which the particular participant implemented the functional analysis protocol when interacting with confederate child during the particular condition of treatment. The second percentage represented by the Y-axis is the accuracy (as measured according to the point-by-point IOA formula) in which the participant were able to correctly score a functional analysis model (or self) video using a specific fidelity grading sheet for that particular condition. The X-axis for each of the graphs represents the session number in which the participant was implementing their mock functional analysis. The percentage of rating accuracy for each of the participants grading the

model (or self) videos will be represented by the triangle data series in each of the data sets. The treatment fidelity (as a percentage) in which each of the participants was able to accurately implement the functional analysis protocol will be represented in each of the data sets as the square data series.

Data set number one shows slight variability in rating accuracy during time of the demand condition treatment phase from 100% to 75% with a large increase in percentage of treatment fidelity during the time of treatment, with a slight dip and upwards trend towards the end of treatment for that condition. When treatment is administered for the attention phase both the participants rating accuracy and treatment fidelity percentage stay high and stable.

Data set number two shows slight variability in rating accuracy from 100% to 62% in the attention condition treatment phase, and a slight upward trend in treatment fidelity during that same phase. Upon switching to the control condition treatment rating accuracy stays constant at around 65% with a slight upward trend in treatment fidelity.

Data set number three shows relative stability in rating accuracy during the control condition treatment phase, with a large increase in treatment fidelity from baseline, showing stability at around 70%. Neither variable showed very much change when the participant to graded self videos.

Data set number four showed rating accuracy during the tangible treatment phase as relatively stable around 95% with treatment fidelity showing a large increase during treatment from baseline in which it remained stable around 95%. When switched to the attention treatment phase both rating accuracy and treatment fidelity stayed stable around 95%.

Data set number five shows rating accuracy remained very stable at 100% during the demand condition treatment phase whereas treatment fidelity remained relatively stable at

around 60%. Upon switching to grading self videos rating accuracy dropped to about 90% and treatment fidelity increased and stayed stable around 90%.

Data set number six shows rating accuracy remained relatively stable at around 95% and treatment fidelity staying stable after dropping to 53%. After the participant switched to grading self videos there was a drop in rating accuracy to 60% followed by a slight upward trend, and a slight upward trend in treatment fidelity.

Data set number seven showed a slight upward trend in both rating accuracy and treatment fidelity during treatment, and a slight drop followed by stability at 85% in accuracy of rating, with a somewhat dip in treatment fidelity upon grading self videos.

Data set number eight showed relative stability in rating accuracy during demand condition treatment with a slight drop off and the end of the treatment phase, and relatively stable treatment fidelity around 95% during the demand condition treatment. During the attention condition treatment there was a slight dip in rating accuracy (from around 100% to 80% back to 100%) with an upward trend in treatment fidelity.

Data set number nine showed a great deal of variability in rating accuracy during the attention condition treatment with treatment fidelity remaining stable around 95%.

Data set number ten showed slight variability in rating accuracy during the attention condition treatment with stable levels of treatment fidelity around 80% when switching to the grading of self videos there was no real apparent change in rating accuracy with a slight increase in treatment fidelity.

Date set number eleven showed a slight dip in rating accuracy during the attention condition treatment phase with a stable level of treatment fidelity at 86%. When the participant

switched to grading self videos there was an upward trend in rating accuracy with treatment fidelity stable at 100%.

Data set number twelve showed large amounts of variability in rating accuracy during the control condition treatment with an increase in treatment fidelity baseline to 100% where it remained stable.

Data set number thirteen showed a dip in rating accuracy during also during the control condition of the experiment with treatment fidelity remaining relatively stable 85% during treatment.

Data set number 14 showed relatively stable rating accuracy around 95% during the tangible condition treatment phase an increase and peak of treatment fidelity. Upon switching to grading self videos there was a drop in rating accuracy where it remained relatively stable around 45%, while showing relative variability in treatment fidelity.

Data set number fifteen showed relatively stable rating accuracy around 95% during the tangible condition phase of treatment with a slight variability in treatment fidelity from 100% to 75%

Discussion

During the course of the study 6 of the 15 data sets showed potential signs of covariance between their treatment fidelity and their rating accuracy data series (data sets 1, 3, 4, 6, 7 and 11). However, this potential relationship did not ensure that high levels of treatment fidelity resulted in high levels of rating accuracy, or that high levels of rating accuracy ensured high levels of treatment fidelity. The remaining seven participants' trends in data did not seem to vary in one or both of the variables across their performance graphs throughout the study.

The results of the study seems to support that there is no real relationship between rating accuracy when grading procedural fidelity of video models and/or self-videos, and the treatment fidelity of implementation of those same procedures in a controlled setting. This conclusion has been made based on the lack evidence that for a relationship between the two variables with in most of the data sets, and relatively no apparent evidence for any type of relationship in the other seven data sets. Although there seems to be no apparent relationship between the two variables the results of the study can still give some insight when performing similar types of training in the future. When employing comparable types training to individuals the practitioners of the study need not be as concerned with level of accuracy in which participants are scoring video models.

This outcome is beneficial to future studies similar to the Field (2013) study. It seems that just viewing and rating video models of others or of one's self to implement the functional analysis protocol is enough to increase treatment fidelity. In new and upcoming studies researchers need not be as concerned with giving feedback or the levels of accuracy in which participants rate the video models they are viewing. This can be very beneficial when training individuals, and can allow researchers to have much less involvement from the board certified behavior analyst.

Some potential limitations to this study include the fact that only the primary observer measured how accurate participants were grading the self-videos. Potentially one or more of the participants using the self-videos may have scored higher or lower when scoring the self-videos and this would not have been detected with the use of only one primary observer. Thus, an additional observer could have completed IOA with the primary observer.

Another limitation to the current study is that during the course of investigation no statistical analysis of covariance was measured. A statistical analysis may give better insight as to whether or not some type of covariance actually exists between the two variables. In future studies, researchers could reach a solution to this problem by performing an analysis of the data to determine if a statistical covariance of the two variables actually exists.

There are also some limitations to the Field (2013) study in which the data sets have been derived. Potential limitations to the Field (2013) study may include participant fatigue, and participants providing feedback to one another outside of the study. While implementing the functional analysis protocol participants of the study may have experienced fatigue when implementing the different functional analysis conditions. That is to say that perhaps the conditions that were being applied by the participants in the beginning of the session may have been implemented with greater fidelity according to the protocol that the conditions towards the end of the session. Due to the participants being tired they may have not graded videos as accurately or implemented protocol as accurately. Another potential limitation being participants may perhaps have been giving one another feedback on the proper ways in which to grade the video models and/or how to properly implement the treatment protocol that had been given to them. If participants were in fact giving one another feedback during the course of the study this may have altered their performances in either variables (treatment fidelity, or accuracy of rating).

There are a few implications can be taken from the findings of this study. In the future during similar training as the Field (2013) study, practitioners do not need to be as concerned about the level of accuracy in which participants are rating any type of video model. These findings imply that the level of treatment fidelity in which participants implement treatment seems to not be indicative of how accurate they can grade video models.

References

Alvero, A. M. & Austin, J. (2004). The effects of conducting behavioral observations on the behavior of the observer. *Journal of Applied Behavior Analysis, 37*(4), 457-648.

Baer, D.M., Wolf, M. M., & Risley, T. R. (1968). Some current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis, 1*, 91-97.

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

Field, S. P., Frieder, J. E., McGee, H. M., & Peterson, S. M. (2013). [Assessing observer effects on the fidelity of implementation of functional analysis procedures]. Unpublished raw data.

Gast, D. L. (2009). *Single Subject Research Methodology in Behavioral Sciences*. New York, NY: Routledge

Kazdin, A. E. (1982). *Single-case research designs: Methods for clinical and applied setting.* New York, NY: Oxford University Press.

Snedecor, G. W., & Cochran, W. G. (1980). *Statistical Methods*, (7th ed.). Ames, IA: Iowa State Press.

Appendix A

Procedure Fidelity Rating Sheets

Each Specific to Functional Analysis Condition

Social Disapproval
Date:
Participant:
Session Number:

Step	Correct + Incorrect -	Misc. Comments
Relevant Materials on Desk		
Provide interaction for 12-18 seconds		
State that they have work to do following interaction		
Face away from participant		
When TB occurs intervene with prompts to stop and redirection		
Intervene for 12-18 seconds		
End intervention without comment and face away		
All verbal attempts to gain attention were ignored		
All non target behavior was ignored		
End session at 5 min (+/- 5 sec)		

Academic Demand

Step	Correct + Incorrect -	Misc. Comments
Relevant Materials on Desk		
Prompting occurred at least 45 times per condition.		
If target behavior occurs provide 12-18 second break with no comment or attention and work materials are moved away from the participant		
Following break, immediately prompt work		
All verbal attempts to escape work were ignored		
All non-target behavior was ignored		
End session at 5 min (+/- 5 sec)		

Contingent Tangible

Step	Correct + Incorrect -	Misc. Comments
Relevant Materials on Desk		
Allow play with toy for 12-18 seconds		
After 12-18 seconds remove toy and place in lap		
When TB occurs intervene with access to the toy Do not say anything		
Allow access for 12-18 seconds		
After 12-18 seconds take toy without comment		
All verbal attempts to access toy were ignored		
All non-target behavior was ignored		
End session at 5 min (+/- 5 sec)		

Unstructured Play

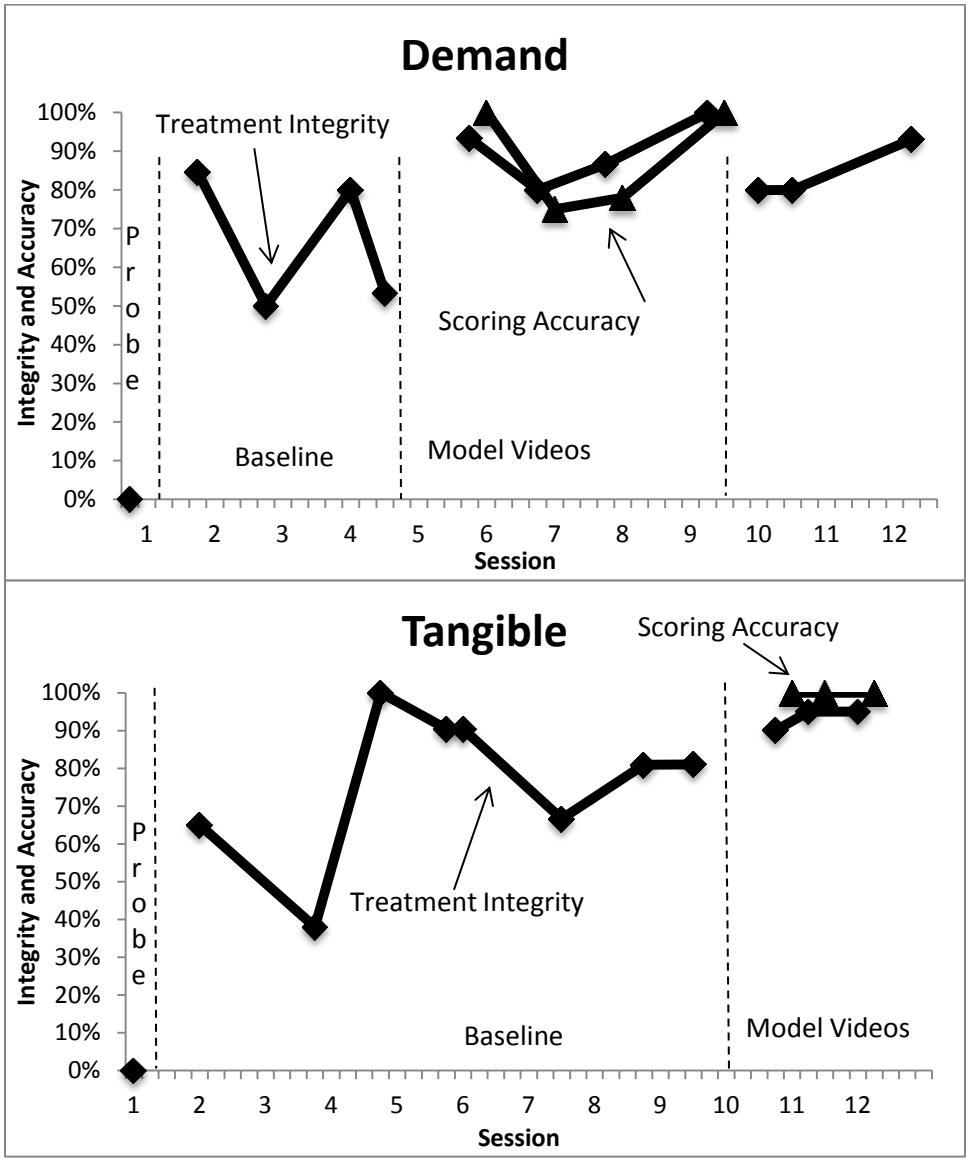
Step	Correct + Incorrect -	Misc. Comments
Relevant Materials on Desk		
Provide interaction for 7-13 seconds every 30 seconds		
If target behavior occurs while attention is being given, cease attention immediately		
Wait 30 seconds following any Target behavior to provide attention		
All target behavior is ignored		
All verbal requests to interact with participant is ignored		
All non target behavior is ignored		
End session at 5 min (+/- 5 sec)		

Appendix B

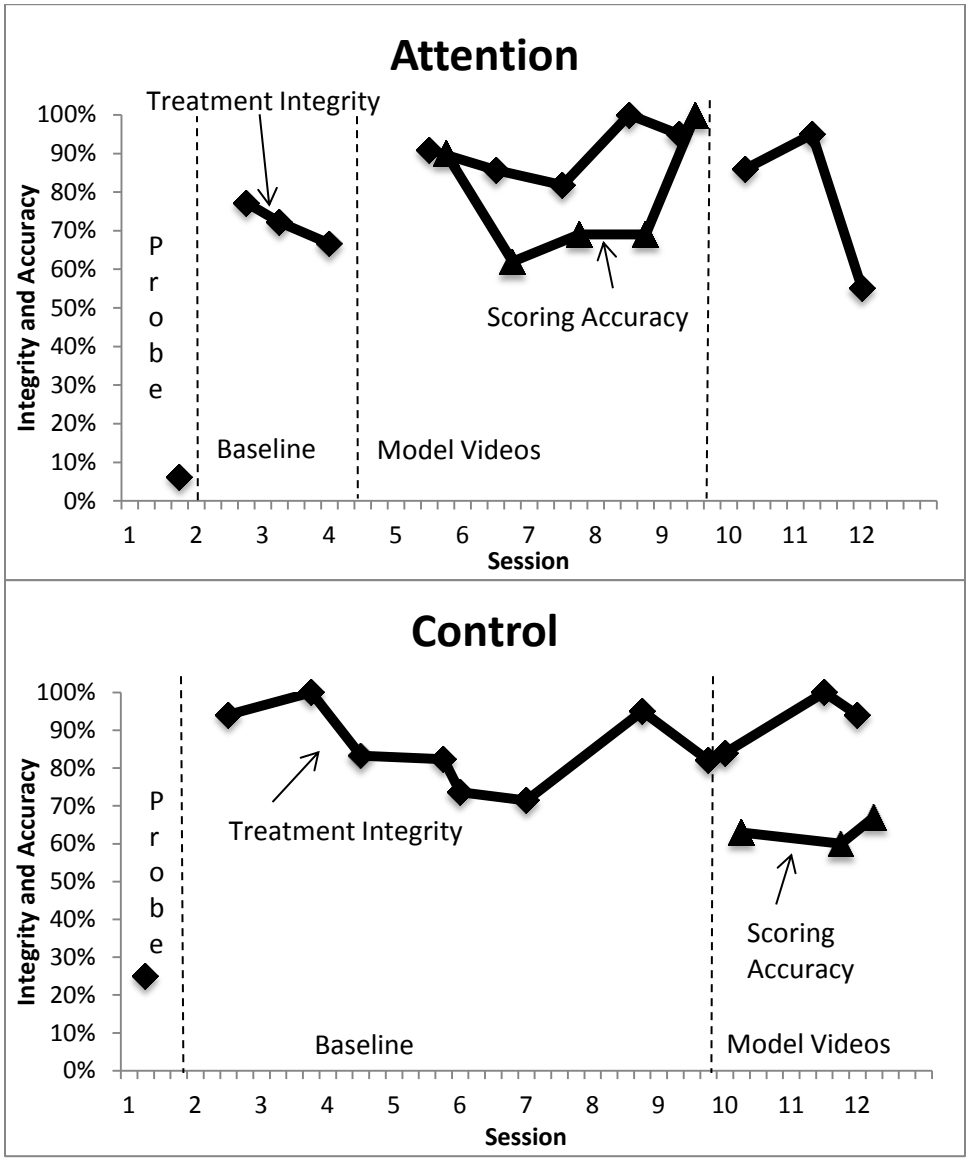
Participant Data Sets

(Data Sets 1-15)

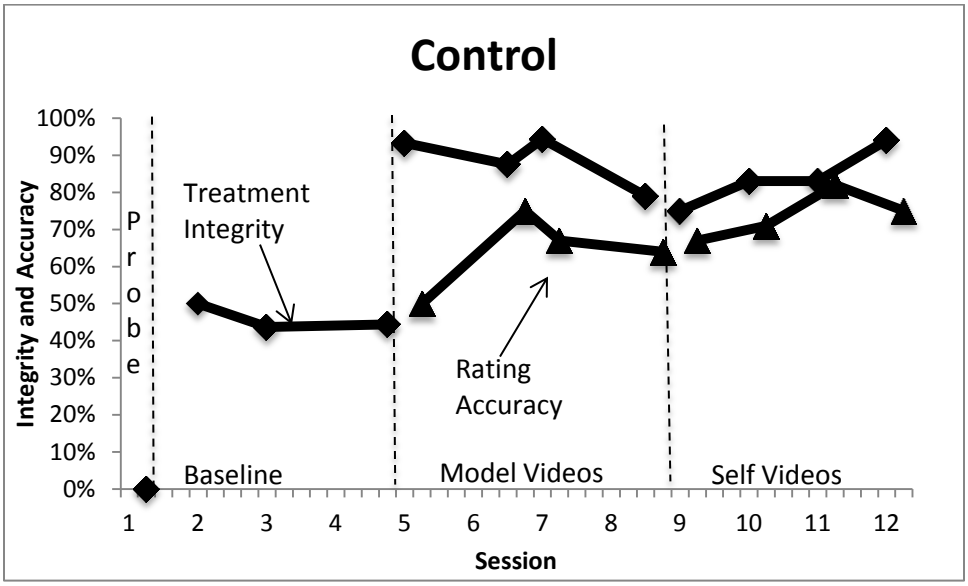
Data Set 1


Demand
Integrity and Accuracy
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Probe
Treatment Integrity
Scoring Accuracy
Baseline
Model Videos
1 2 3 4 5 6 7 8 9 10 11 12
Session
Tangible
Scoring Accuracy
Integrity and Accuracy
Probe
Treatment Integrity
Baseline
Model Videos
1 2 3 4 5 6 7 8 9 10 11 12
Session

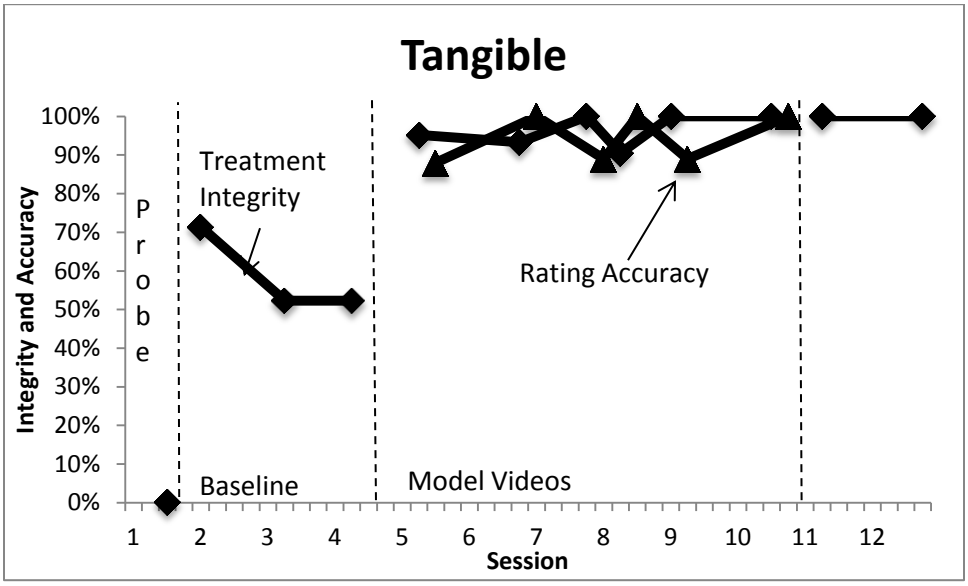
Data Set 2

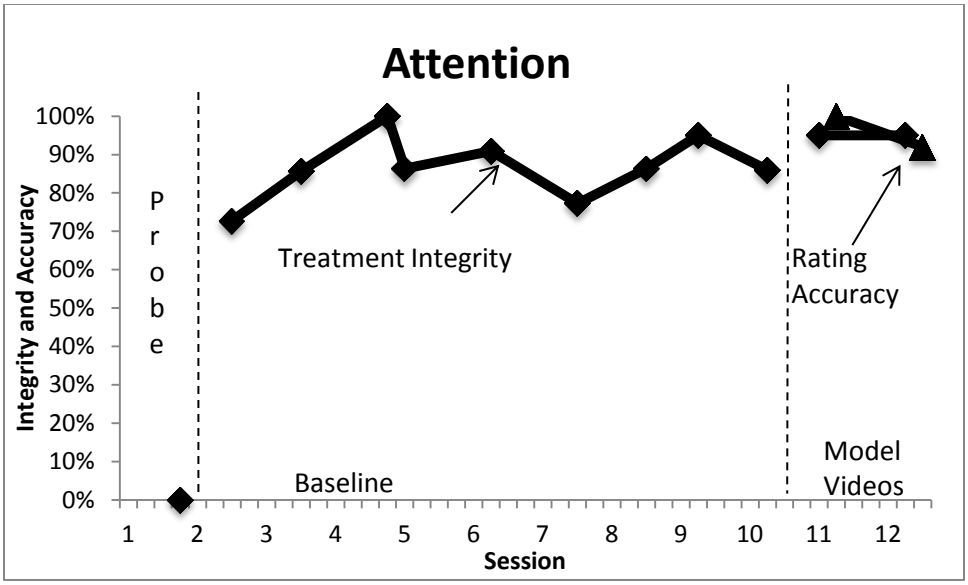

Attention
Treatment Integrity
Scoring Accuracy
Probe
Baseline
Model Videos
Integrity and Accuracy
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1 2 3 4 5 6 7 8 9 10 11 12
Session
Control
Treatment Integrity
Scoring Accuracy
Probe
Baseline
Model Videos
Integrity and Accuracy
Session

Data Set 3

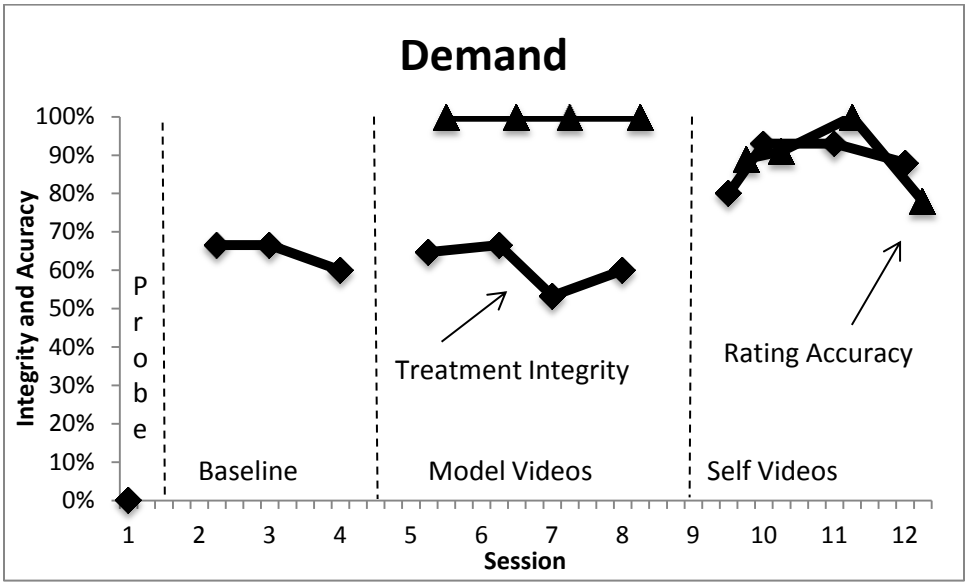

Control
Integrity and Accuracy
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Probe
Treatment Integrity
Rating Accuracy
Baseline
Model Videos
Self Videos
1
2
3
4
5
6
7
8
9
10
11
12
Session

Data Set 4

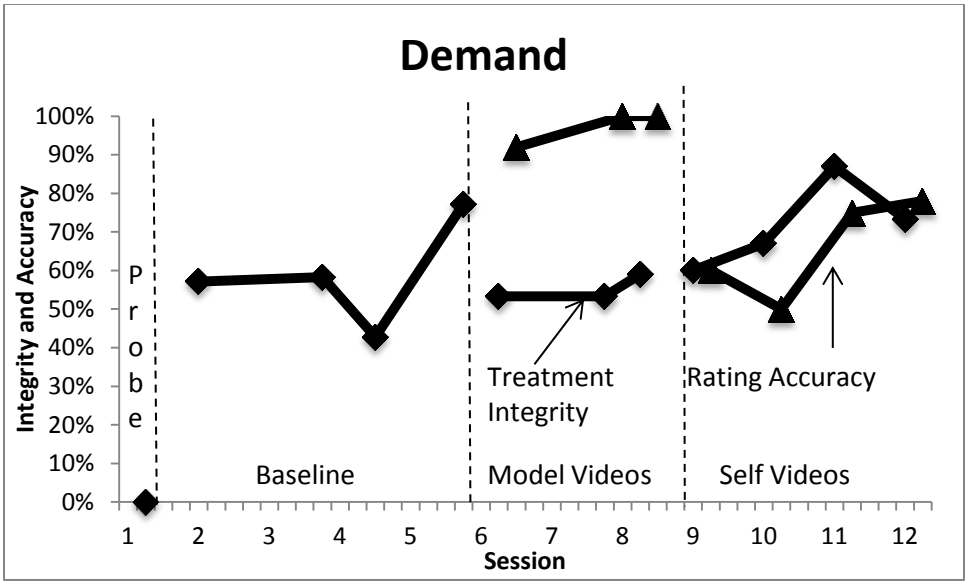

Tangible
Integrity and Accuracy
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Probe
Treatment Integrity
Rating Accuracy
Baseline
Model Videos
1 2 3 4 5 6 7 8 9 10 11 12
Session


Attention
Integrity and Accuracy
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Probe
Treatment Integrity
Rating Accuracy
Baseline
Model Videos
1 2 3 4 5 6 7 8 9 10 11 12
Session

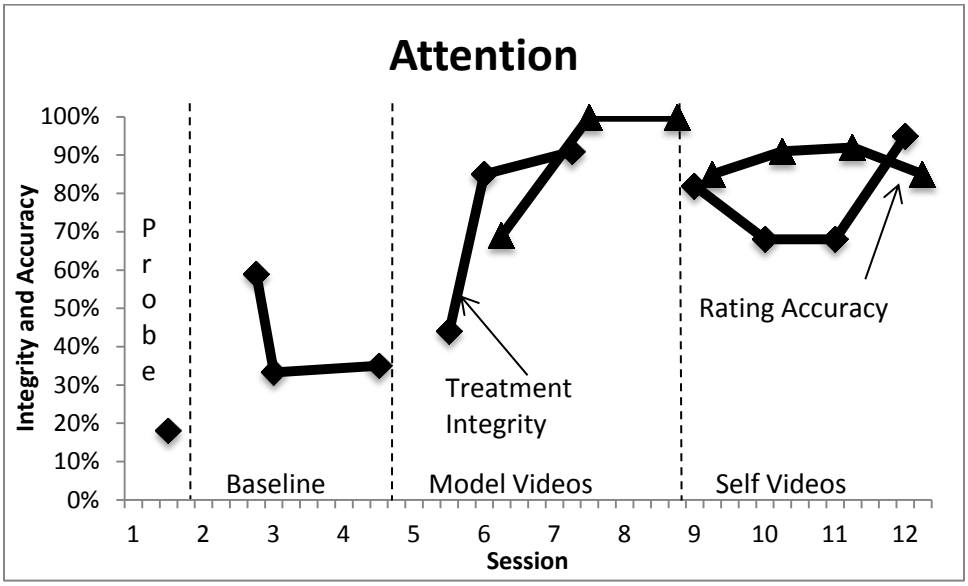
Data Set 5



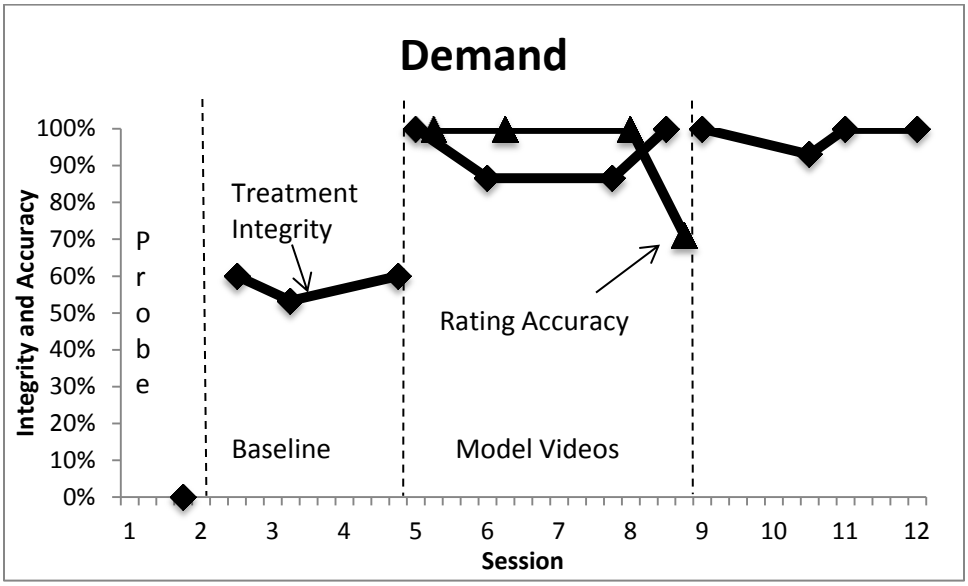
Data Set 6

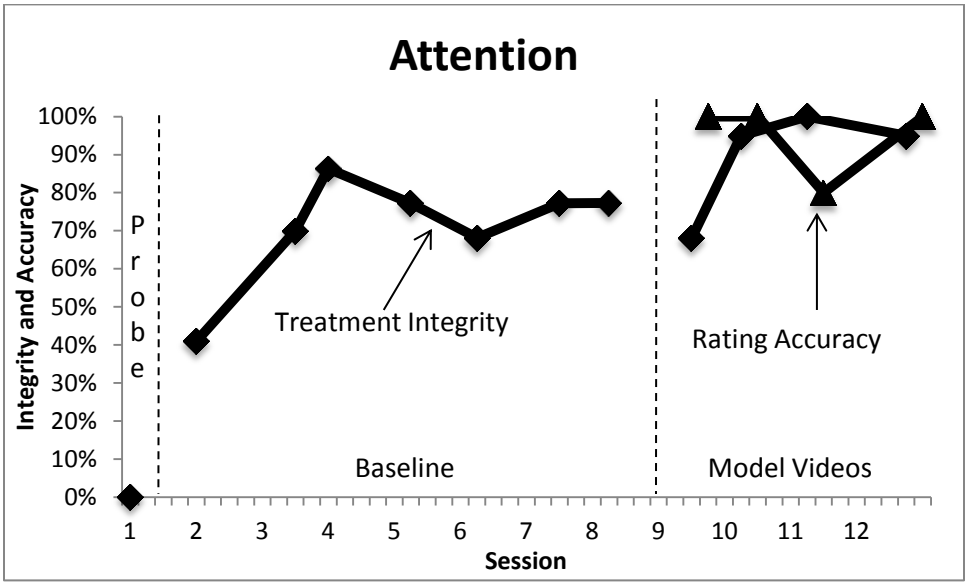


Data Set 7

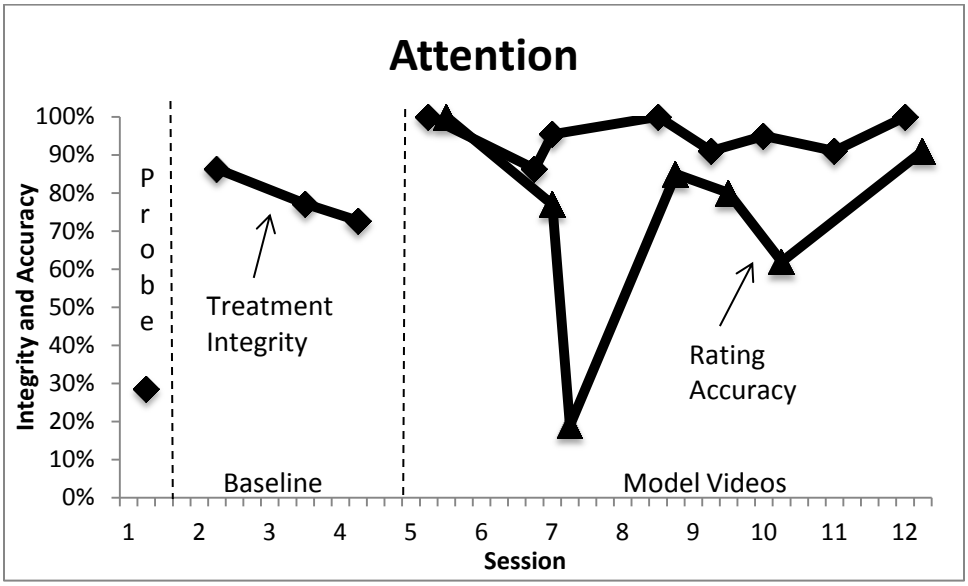

Attention
Integrity and Accuracy
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Probe
Baseline
Model Videos
Self Videos
Treatment Integrity
Rating Accuracy
1
2
3
4
5
6
7
8
9
10
11
12
Session

Data Set 8

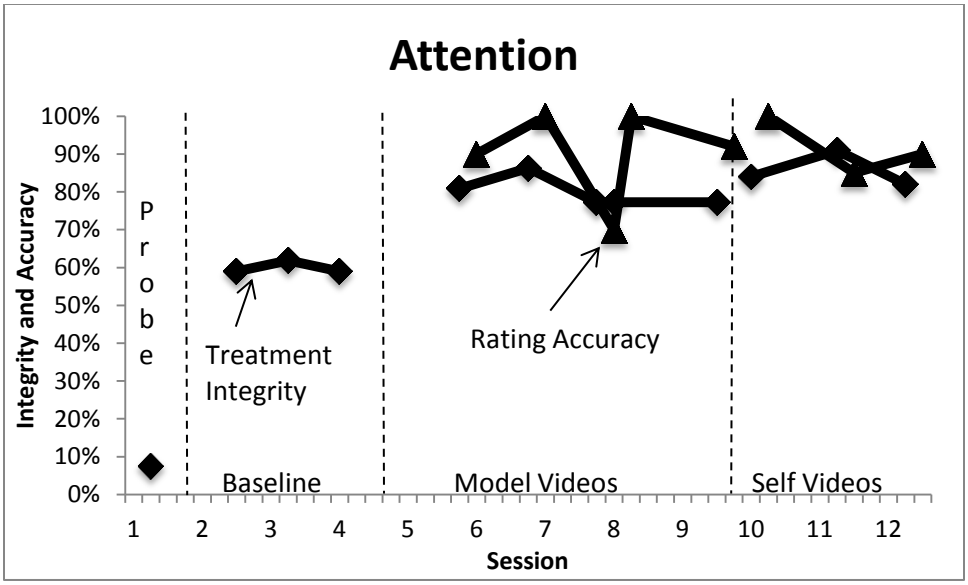

Demand
Integrity and Accuracy
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Probe
Treatment Integrity
Rating Accuracy
Baseline
Model Videos
1 2 3 4 5 6 7 8 9 10 11 12
Session


Attention
Integrity and Accuracy
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Probe
Treatment Integrity
Rating Accuracy
Baseline
Model Videos
1 2 3 4 5 6 7 8 9 10 11 12
Session

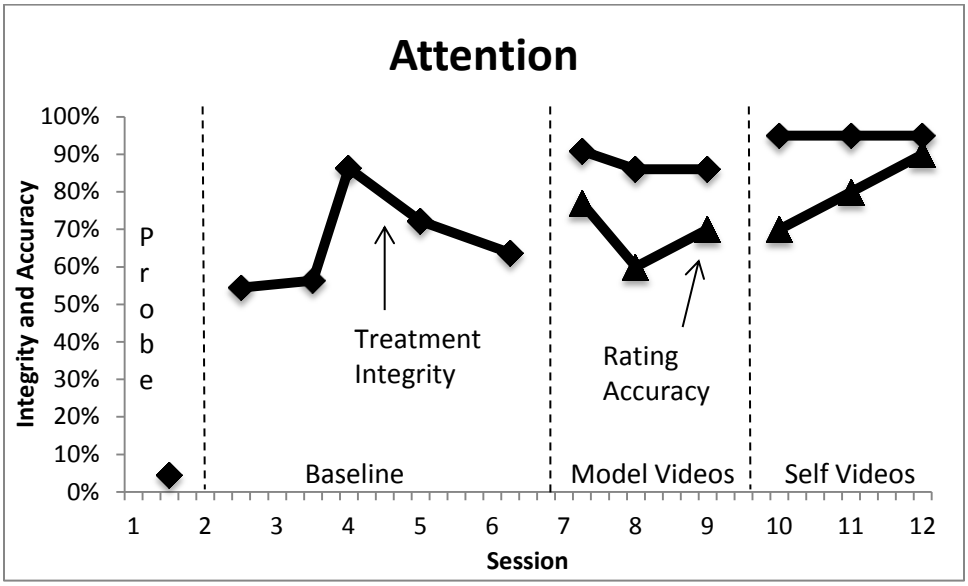
Data Set 9



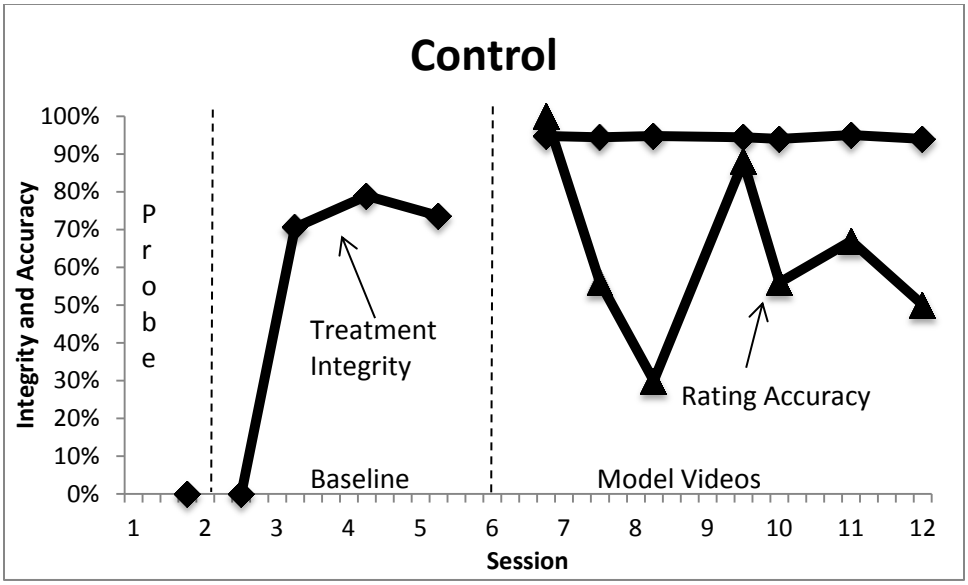
Data Set 10



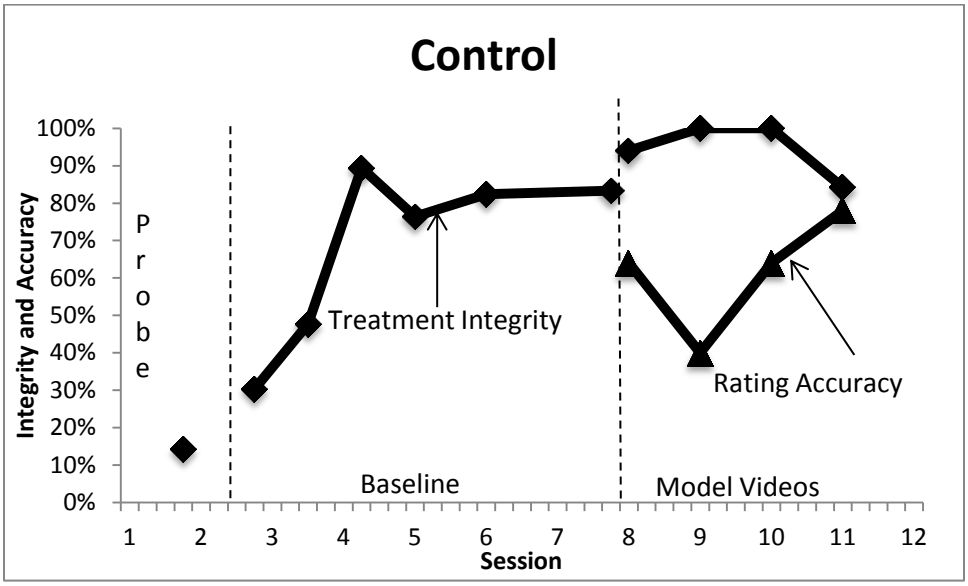
Data Set 11



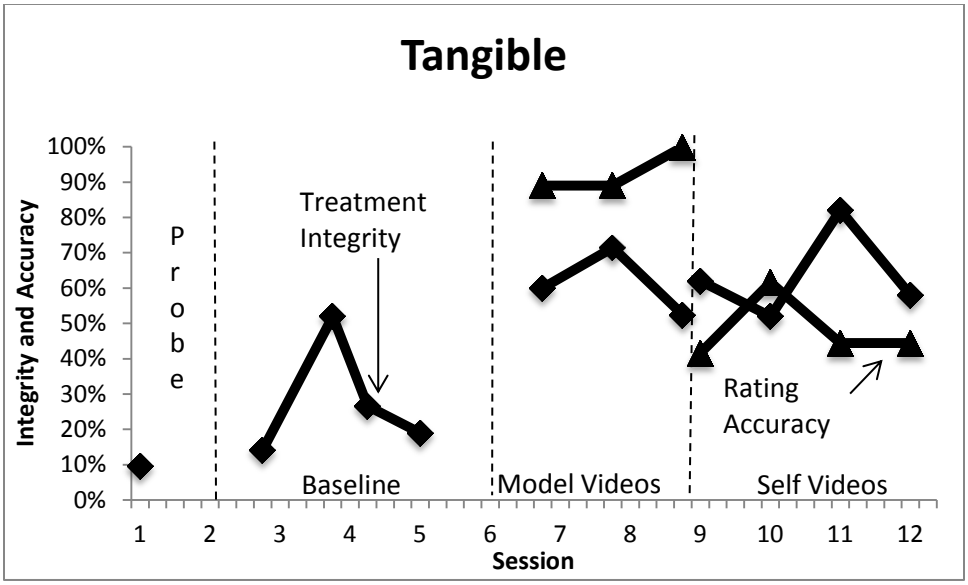
Data Set 12



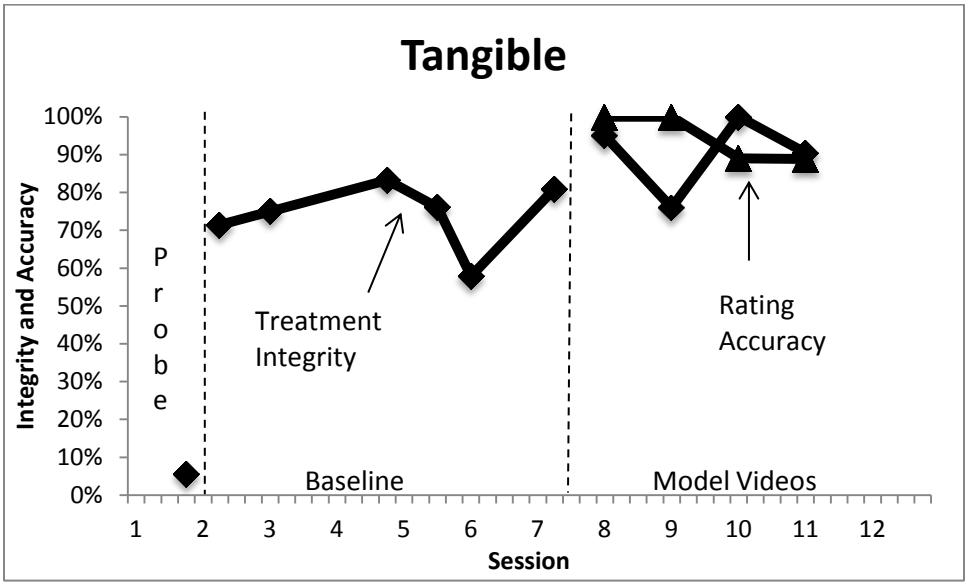
Data Set 13



Data Set 14



Data Set 15


Tangible
Integrity and Accuracy
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Probe
Treatment Integrity
Rating Accuracy
Baseline
Model Videos
1
2
3
4
5
6
7
8
9
10
11
12
Session