Improving the Reliability and Validity of Visual Inspection of Data by Behavior Analysts: An Empirical Comparison of Two Training Methods to Improve Visual Inspection and Interpretation, the Job Aid and the Conservative Dual-Criteria

Marilin Colón
IMPROVING THE RELIABILITY AND VALIDITY OF VISUAL INSPECTION OF DATA
BY BEHAVIOR ANALYSTS:
AN EMPIRICAL COMPARISON OF TWO TRAINING METHODS TO IMPROVE
VISUAL INSPECTION AND INTERPRETATION, THE JOB AID AND THE
CONSERVATIVE DUAL-CRITERIA

By
MARILIN COLÓN

A dissertation submitted to
the Department of Psychology
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Degree Awarded:
Spring Semester, 2006
The members of the committee approve the dissertation of Marilin Colón defended on March 27, 2006.

Jon S. Bailey
Professor Directing Review

___________________________
Robert Contreras
Committee Member

___________________________
Frank Johnson
Committee Member

___________________________
Bryan Loney
Committee Member

___________________________
Bruce Thyer
Outside Committee Member

The Office of Graduate Studies has verified and approved the above named committee members.
This work is dedicated to my mother, whom I have always admired for her achievements.
ACKNOWLEDGEMENTS

I would like to thank Dr. Jon S. Bailey for his support and critical feedback during the conduct of this research. I would also like to thank Cassandra Gayman, whose assistance in research preparation and data collection was so helpful. I would like to express thanks to all my committee members for their insightful contributions. Thank you also to Mike Mozzoni, Al Murphy and Kevin Murdock for their valuable expertise. I would also like to express my enormous gratitude to Dr. Maxin Reiss for all the support both professionally and personally that she has provided for me throughout this whole process. Finally, I would like to thank my husband, Matthew Normand, and my friends and family for their love and support throughout this study.
# TABLE OF CONTENTS

List of Tables vii
List of Figures viii
Abstract x
Introduction 1
  Previous Visual Analysis Research 2
  Recent Visual Analysis Research 11
  Hypotheses 20
  Study 20
Method 21
  Participants 21
  Stimulus Materials 21
  Experimental Design 31
  Data Analysis 33
Results 35
  Overall 35
  Individual Subject Data 35
  Panel of Experts 39
  Group Data 39
  Type I and Type II Errors 45
  Time 51
  Characteristics 55
  Intra-subject Agreement 56

Discussion 63
Appendices 74
  Appendix A 74
  Appendix B 76
  Appendix C 123
Appendix D 124
Appendix E 128
Appendix F 132
Appendix G 134
Appendix H 137

References 139
Biographical Sketch 142
<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Breakdown of Type I and Type II Errors by group</td>
<td>52</td>
</tr>
<tr>
<td>2</td>
<td>Total and Percentage of Type I and Type II errors made by participants across all conditions for the CDC Lines Job Aid group</td>
<td>53</td>
</tr>
<tr>
<td>3</td>
<td>Total and Percentage of Type I and Type II errors made by participants across all conditions for the Active Job Aid group</td>
<td>53</td>
</tr>
<tr>
<td>4</td>
<td>Total minutes per group to complete experiment</td>
<td>54</td>
</tr>
<tr>
<td>5</td>
<td>Total minutes to complete each condition for CDC Lines Group and Active Job Aid group</td>
<td>54</td>
</tr>
<tr>
<td>6</td>
<td>Intra-subject agreement for repeated package by CDC Lines group participants</td>
<td>60</td>
</tr>
<tr>
<td>7</td>
<td>Intra-subject agreement for repeated package by Active Job Aid group participants</td>
<td>61</td>
</tr>
<tr>
<td>8</td>
<td>Intra-subject agreement for repeated package by Control group participants</td>
<td>62</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

1. An example of the graphs and questions presented to all participants 23

2. An example of the Option A graph presented to participants during training for the CDC and Active job aid groups 24

3. An example of the Option B graph presented to participants during training for the CDC and Active job aid groups 25

4. An example of the Option C graph presented to participants during training for the CDC and Active job aid groups 26

5. An example of the mean data for the Option B graph used in the Active Job Ad training 27

6. An example of the trend data for the Option B graph used in the Active Job Aid training 28

7. An example of the variability data for the Option B graph used in the Active Job Aid training 29

8. Percent unanimous agreement with experts per packet across CDC Lines group participants 40

9. Percent unanimous agreement with experts per packet for Active Job Aid group participants 41

10. Percent dichotomous agreement with experts per packet across CDC Lines group participants 42

11. Percent dichotomous agreement with experts per packet across Active Job Aid group participants 43

12. Unanimous (left graph) and dichotomous (right graph) percent agreement per packets with the experts for Control group participants 44

13. Percentage of participant agreement for each graph 46

14. CDC Lines group participant-by-participant accuracy across experimental conditions 47

15. Average percent unanimous and dichotomous agreement with experts for each CDC Lines group participant 47
16. Unanimous and dichotomous average agreement by Active Job Aid group across experimental phases  48

17. Active Job Aid group participant-by-participant accuracy  48

18. Control group participant-by-participant accuracy across experimental conditions for unanimous and dichotomous agreement with experts  49

19. Average percent agreement of CDC Lines Job Aid participants with experts on graph characteristics for each condition.  57

20. Overall average percent agreement with experts on graph characteristics by CDC Lines Job Aid participants  57

21. Average percent agreement of Active Job Aid participants with experts on graph characteristics for each condition  58

22. Overall average percent agreement with experts on graph characteristics by Active Job Aid participants  58

23. Overall average percent agreement of Control participants  59
ABSTRACT

The present study compared the effects of two job aids on the behavior of visual data analysts. Twelve participants were randomly assigned to a CDC Lines Group, an Active Job Aid group, or a Control group and viewed a series of 56 within-subject behavioral graphs. They were asked to determine whether the graphs depicted a change that was attributable to the intervention. Results indicated that participant answers in the CDC Lines group unanimously agreed with the answers determined by a panel of experts of the graphs for 92% of the graphs during the intervention condition. For the Active Job Aid group, participants unanimously agreed with the experts for 73% of the graphs during the intervention condition. Unanimous agreement was 50% for the Control group. For dichotomous agreement with the experts, the CDC Lines group averaged 94% and the Active Job Aid group averaged 81% agreement with the experts during the intervention condition. Dichotomous agreement was 72% for the Control group. The CDC Lines Job Aid decreased both Type I and Type II errors during the intervention condition, as compared to baseline and maintenance conditions. The Active Job Aid decreased Type I errors, but increased Type II errors, as compared to baseline and maintenance conditions. The CDC Lines Job Aid improved agreement with the experts for all participants. The Active Job Aid improved agreement with the experts for only two out of the four participants. Participants in all groups were asked to identify the graph characteristics that most influenced their decision about intervention effect for each graph. The job aids did not increase the ability of the participants to correctly identify these characteristics. It is argued that the results obtained have important implications for the training and credentialing of professionals employing visual data analysis, especially behavior analysts.
INTRODUCTION

In the past few years, there has been a resurgence of research interest in the behaviors involved in visual analysis in order to empirically establish guidelines to direct such analysis. These guidelines aim to increase the reliability and validity of visual inspection. Also, in the forefront is the concern of how most effectively and efficiently to train people on these guidelines. Since within-subject data is typically graphed and analyzed visually, empirical answers to these questions would help to disseminate the use of within-subject research into other fields of psychology.

While behavior analysts have always seen the power of this methodology to identify the effects of variables on an individual’s behavior, some psychologists do not agree with its new emphasis in the clinical and basic research settings. Because of the seemingly inherent subjectivity of visual data analysis, they insist that statistical analysis is the most powerful method of identifying the significance of results. These psychologists have argued that statistical hypothesis tests are superior to other methods because they have clear guidelines describing what tests are appropriate to specific types of data and the way in which such tests are to be conducted. As a result, many different researchers can analyze the same data in the same way and obtain the same results. Unfortunately, the same cannot be said of visual analysis.

Nevertheless, visual analysis or criterion-by-inspection (Sidman, 1960) has been the technique of choice applied to behavior-analytic work since Skinner in The Behavior of Organisms (1938) emphasized that a science of behavior is about individual organisms in which statistical analysis does not belong. Since the goal is to predict and control behavior, Skinner believed it necessary to understand the difference from one individual to the next. Large between-groups statistical analyses treat variability as unwanted “noise,” thus, eliminating these valuable variations.

Applied behavior analysis is a field that strives to change socially significant behavior for the benefit of individuals and society (Baer, Wolf, & Risley, 1969). Jack Michael (1974) argued that behavior analysts must consider it a failure if their interventions cause such minimal change that they require statistical tests to be detected. Moreover, it would be detrimental to this young field since researchers might be satisfied with less powerful interventions as long as
the effects were statistically significant rather than finding the most effective interventions. Most behavior analysts would wholeheartedly agree with the following statement from a reviewer of a statistically analyzed single-subject study in a study by Gottman and Glass (1978), “Behavior modification became a technology by ignoring small effects. The sort of eyeball analysis that has been done has acted as a filter, weeding out small effects.” (p. 198). Even though this comment was supposed to be critical of the field, it embodies for what Michael (1974) suggested behavior analysts should strive.

Training new researchers and practitioners to visually analyze behavioral data is subject to great variability. Even among skilled behavioral researchers, disagreement about the effects depicted by graphed behavioral data is widespread. Recent efforts to improve this training have had mixed results. Some of the recent research has found that while participants can be taught to identify important graph characteristics, the training has not been effective in modifying their behavior when making decisions on intervention effects (Normand & Bailey, 2003).

Different experiments have tried to establish what the specific characteristics are that exert an influence on visual inference. The goal is to define these characteristics to such an extent that they can be employed to teach students and new professionals to reliably and validly identify intervention effects. As Parsonson and Baer (1978) noted, “While it is apparent that behavior analysts develop considerable skills in the visual analysis of graphic data, the means by which this skill [sic] is transmitted appears to rely on oral, rather than written, tradition.” (p. 133) In this respect, visual analysis would be taught in the same manner as statistics with established criteria for determining the significance of clinical and experimental results. The field of behavior analysis has not yet reached a consensus on what these variables are. Unfortunately, a statement made by Kadzin (1982) over 20 years ago, still applies today: “The major issue pertains to the lack of concrete decision rules for determining whether a particular demonstration shows or fails to show a reliable effect.” (p. 239) More research is necessary until these variables are identified. Until then, single-subject methodology will continue to be criticized as an unreliable methodology.

Previous Visual Analysis Research

In one of the earliest studies to evaluate visual analysis of within-subject data, Jones et al. (1978) examined the effect of serial dependence on the agreement between visual analysis
and statistical inference of behavioral data. A panel of 11 professors, researchers and graduate students were simultaneously shown 24 graphs taken from published articles in the Journal of Applied Behavior Analysis (JABA). The graphs were chosen based on some or all of the following criteria: 1) subtle changes in level between phases, 2) multiple baseline research design with several phases in which there were few and unequal number of data points in each phase and, 3) visually apparent serial dependence. After viewing each graph, the group of analysts had to decide whether there were “meaningful” changes (reliable experimental effects) in level among phases of each graph and sort them into three categories (low, moderate or high) based on the amount of serial dependence depicted in each. For each graph, the mean agreement between the judges’ decisions was compared to the result of the time series analysis.

The authors reached three main conclusions. The first was that visual inferences and time-series inferences were in agreement only 60% of the time, which they state is not much greater than chance. The second conclusion was that as serial dependency increased, agreement between the participants and the statistical tests decreased. Since behavioral data is serially dependent, it is likely that agreement between visual analysis and statistical tests will always be low. Lastly, when the time-series tests suggested statistically reliable changes in behavior, agreement between them and the judges was the poorest with agreement only 52% of the time. The authors suggest that statistically reliable experimental effects may be more often overlooked by visual appraisals of data than non-meaningful effects.

The authors suggest that time-series inferences should be utilized as a supplement to visual inferences because of the reliability of the time-series analyses and their ability to capture subtle changes in the data. However, because serial dependency is a prevalent characteristic of experimental data in operant studies, differences in agreement between these tests should be expected. The authors do not resolve the issue of which analysis to accept when there are

---

1 Serial dependency means that temporally adjacent scores tend to be related to, or predictive of one another. This common property of most behavioral scores violates the independence assumption underlying most common statistical methods.

2 Time-series analysis is a statistical procedure that accommodates for serial dependency and is used to make inferences about change in level and trend among several phases of an individual-subject experiment.
differences but based on their assessment that visual analysis suffers from low reliability, it is assumed that they would suggest relying ultimately on the results of the statistical tests.

Since reliability among judges was low, DeProspero and Cohen (1979) studied whether agreement in visual judgment could be reliably attributed to certain graph characteristics. The authors simulated 36 “ABAB reversal design” graphs and divided them into nine packets. Each of the packets had a graph illustrating each of the four factors thought to represent characteristics important for visual analysis: degree of mean shift, pattern of mean shift, variation between phases and trend. One of the nine packets was randomly mailed to each of 250 participants, who were either editorial board members or guest reviewers of the JABA or the Journal of the Experimental Analysis of Behavior (JEAB).

The reviewers had to rate the graphs in their packet from a low of zero to a high of 100 on how satisfactory a demonstration of experimental control they considered it to be. DeProspero and Cohen (1979) found that there was only 60% agreement between the reviewers, replicating the finding that the visual analysis of judges was not particularly reliable. However, the method of indexing confidence levels could have lowered interrater agreement considering the breadth of the scale.

In addition, the reviewers were asked to list those factors that were instrumental in their decision-making. The authors subjectively summarized the characteristics identified by self-report into four groups: topography of result, format of presentation, intra-experimental concerns and extra-experimental concerns. They concluded that these factors appear to affect judgment in concert rather than singly. Moreover, the reviewers seemed to weigh the factors differently when making their decisions, which could account for the variability in their responses. As in the previous study, the authors recommended that visual analysis be utilized in conjunction with statistical procedures because of the reliable nature of statistical tests. DeProspero and Cohen (1979) also proposed that job aids be developed to assist behavioral researchers with the visual analysis of data in an effort to reduce the variability that they encountered. It would be 24 years before a study incorporating this idea would be published.

A few years after DeProspero and Cohen (1979) published their study, Wampold and Furlong (1981) proposed that the reason that the previous studies had found such shortcomings with visual analysis was due to the fact that the field lacked a theoretical model that explained
the process. In their study, they attempted to provide a theoretical framework from which to examine the process of visual inference and to empirically test hypotheses generated by their framework.

Their framework, based on schema theory\(^1\), suggested that given a class of graphs, a prototypical graph exists that is representative of all graphs. Different graphs are examplars that range from those that are similar to the prototype to those that are more dissimilar to the prototype. Individuals, then, classify new graphs by comparing it to the prototype. Based on this theory, they hypothesized that subjects trained primarily with prototypes and small-distance exemplars (such as, behavior analysts) would attend primarily to large absolute changes and ignore variation in classifying graphic data. On the other hand, subjects treated primarily in advanced statistical procedures would attend more to the size of change in relation to variation.

Two groups of participants were selected to test this hypothesis. One group was comprised of 14 graduate students, who had completed a course on within-subject research designs. The second group was composed of 10 graduate students who had completed two of three quarters of a statistical analysis course. Both groups were asked to sort 36 AB graphs with no ordinates or abscissas into piles, such that each pile contained graphs that demonstrated similar “effects.” The graphs depicted three different phase changes: level shift, trend shift, and both level and trend shift. Their results showed that even though neither group performed the task exceptionally well, the subjects trained in statistics were better able to differentiate the intervention effects than were the subjects trained in visual inference. However, this sorting task was very uncharacteristic to one usually performed by visual analysts so the results may be very uncharacteristic of typical visual analysts. In addition, even though the graduate students had had a course in single-subject design, the authors acknowledged that their experience in the field was small compared to individuals who are professionally involved with single-subject designs. Experts may have performed better in this type of sorting task. Even though this

---

\(^1\) Schema theory states that given a class of objects, a prototype of that object exists such that the prototype represents the central tendency of the objects. Each object in a class is referred to as an examplar of that class or of that prototype and exemplars can range in similarity to the prototype from small-distance exemplars (similar ) to large-distance exemplars (dissimilar). The theory proposes that individuals abstract the prototype for each class and compare new exemplars to each of the prototypes. In addition, the individual will also extract information about the variability of the examplar from its prototype.
unlikely based on the uncharacteristic nature of the task, it does lead researchers to view these conclusions with caution.

Furlong and Wampold (1982) addressed this issue in a follow-up study a year later. They examined the ability of experts in applied behavior analysis to view graphs, classify them as examples of common intervention effect patterns, and to determine how variation in the data affected these judgments. This was one of the first studies to try to establish a group of experts to judge intervention effects. The participants were ten members of the JABA editorial board, who had extensive experience with single-subject methodology.

The graphs used as stimulus materials were created by a computer program that generated intervention effects reflecting a change in level, trend or both level and trend. As in the previous study, the graphs were drawn on index cards without ordinates or abscissas. A set of 36 graphs of this hypothetical data were mailed to each participants and the experts, again as in the previous study, were requested to free-sort into classes that demonstrated “similar effects.” This lack of restriction was intended so that individual visual inferences could be self-reported. The authors found that as in their previous study, most of the experts were able to identify intervention effect patterns but attended to the absolute size of the effect more than the relationship between the size of the effects and variation, thus naming graphs that were mathematically equivalent as being different. In addition, Furlong and Wampold (1982) concluded that the variable manner in which experts view and interpret graphs is responsible for the lack of interjudge agreement reported in previous research. (DeProspero & Cohen, 1979; Jones et al., 1978)

This variable manner, could not be further analyzed by the experimenters because of the survey method employed to collect the data. As in Wampold and Furlong (1981), the uncharacteristic nature of the task to sort graphs into piles seems of little value without evidence of its relationship to accurate data interpretation. More recent studies, such as Normand and Bailey (2002), suggest that the ability to identify data characteristics does not correlate with accurate decision-making.

Furlong and Wampold (1981) themselves proposed three further limitations of their study. First, only two variables, variance and intervention effects were examined. Second, they admit to the limitations imposed by using hypothetical data, which might introduce some
artificiality into the study. Lastly, they acknowledged that the AB design is not the best design for demonstrating experimental control.

Research addressing the controversy of relying solely on visual inspection for graphed data to assess changes in subject performance continued. This controversy centered around two issues: 1) the distinction between changes that are clinically significant versus those found to be statistically significant (Michael, 1974), and 2) the reliability of visual analysis. All of the previous studies provided evidence of the inconsistency and contradictions of visual inference. This accumulated evidence prompted some researchers to advocate the use of adjunctive procedures to supplement visual analysis (DeProspero and Cohen, 1979; Jones et al., 1978).

One such procedure that was advocated to supplement visual analysis was the split-middle method of trend estimation or celeration line approach. This approach consists of fitting a trend line to baseline data and projecting it to the treatment phase in an effort to predict performance. If a change in phases is present, the proportion of data points above (or below) the trend line will be different than the proportion during the baseline phase. Since the celeration is not a statistical test, but provides statistical probability and is fairly easy to compute, its advocates proposed its use as a supplement to visual inference. A study by Bailey (1984) suggested that the use of this method improved interrater agreement associated with visual analysis. In 1990, Ottenbacher published a study with the purpose of further developing the understanding of the analytic process by exploring the agreement obtained by visual analysis to that of the split middle method of trend estimation.

In his study, Ottenbacher (1990) had 30 subjects analyze 24 -AB graphs composed of either hypothetical or actual data and rate on a 6-point scale how much they agreed with the statement, “There is a significant change in performance across the two phases.” The hypothetical graphs were constricted to represent a variety of data patterns and varied across six factors believed to influence the interpretation of single subject data (Parsonson & Baer, 1986). The six graphic factors were: 1) degree of mean shift across phases, 2) degree of variability across phases, 3) change in slope across phases.

The results of the study revealed poor agreement between visual judgments and the results of the statistical inferences based on the split-middle method of trend estimation. When
this celeration line approach was used as the “gold standard” because of the consistency of its results, the findings also indicated low sensitivity, specificity, and predictive ability for visual analysis. However, Parsonson and Baer (1986) argue that the insensitivity associated with visual analysis of graphed data is one of the primary advantages of visual interpretation.

The methods employed may have had effects that affected the results. In the first place, the selection criteria introduced the potential bias of varying competencies relating to visual analysis since the participants had no formal training in visual analysis. In addition, as in other previous studies, the graphs that the participants rated were mostly comprised of hypothetical data in an AB design, the limitation of which have been critiqued in previous studies. Lastly, the use of the method of trend estimation was selected from numerous other statistical methods that could have been used for comparison purposes, such as least squares General Linear Model (GLM) and interrupted time series (ITSE), so its “gold standard” determination is questionable. Ottenbacher (1990) acknowledges that more research is needed to establish a “gold standard” of data analysis for single subject research. As stated previously, the lack of formal decision rules for visual analysis has had a restrictive influence on research efforts.

Richards et al. (1997) replicated and extended Ottenbacher’s (1990) thread of research by examining the effects of subject and rater characteristics. As in Ottenbacher’s study, two of their goals were to determine the reliability and validity of visual analysis by comparing the accuracy of visual inference with a statistical procedure, the split-middle method of trend estimation. In addition, Richards et al. (1997) wanted to specifically determine if the expertise of the raters affected accuracy, along with the effect on accuracy of the subjects’ characteristics, label and gender.

The authors wanted to examine the effect that external factors have on the accuracy of visual analysis. Even though DeProspero and Cohen (1979) had suggested that rater expertise might be one of those variables, very few studies, such as by Wampold and Furlong (1982) had published research concerning expert performance on visual inspection. The “novice” group of participants was comprised of undergraduate students who had not yet received instruction on visual analysis of graphed data. The “experts” were graduate students and in-service teachers who reported that they had experience in analyzing graphed data. The authors employed the
same stimulus material as Ottenbacher (1990), except that trend lines were drawn on both phases, as opposed to just the intervention phase.

Richards et al. (1997), surprisingly, found no differences between the two groups and that they tended to make the same errors in analyzing the graphs, frequently disagreeing with the results of the split-middle method. Both groups had a high percentage error rate with the tendency to assume a change was significant when it was not (75%) and to state there was no significant changes when statistically there were (44%). However, the following precautions should be taken when interpreting their data. First, it is not clear how long the “experts” had been analyzing data so they may not have been too different from the “novice” participants. Second, the participants only had one minute to make their decisions on whether changes in phases were significant. This may have affected how they analyzed the data. Moreover, the participants had to rate the significance on a five-point scale, which may make dichotomous scorings hard to ascertain. Finally, as in Ottenbacher’s (1990), the use of the split-middle method as a standard on which to base whether or not there is significant phase difference is still to be established.

Other factors thought to affect decision-making in visual analysis by Richards et al. (1997) were the characteristics of the individuals being observed. Therefore, only some of the individuals were labeled as having behavior disorders. The age of the individual remained the same, but the gender, male or female, also varied. Various combinations of label and gender were randomly assigned to subjects. The rationale for using these characteristics is that, according to Richards et al. (1997), they are ones that in the special education literature have been reported as affecting the expectations and attitudes towards them.

Neither the gender nor educational labeling of the students led to significantly differential ratings. Nevertheless, the authors report that the mean ratings were higher when the rate thought the subject was male or when the rater thought the subject had a behavior disorder. Richards et al. (1997) concluded that there might be a tendency to assume that behavior has significantly changed when the subject is male or has a behavior disorder. However, this seems to be an exaggerated statement made by the authors since the statistical analysis yielded no significant results.
As in previous studies, the authors recommend that the visual analysis of data not be used as the sole method of determining significant behavior change. This decision is partly based on the determination that visual analysis resulted in incorrect decisions fewer times than if it had been left up to chance when using statistical significance as the standard for change. Furthermore, the inaccuracy of the raters included both groups and was not significantly related to the characteristics of the subjects. The authors conclude that it cannot be assumed that individuals can automatically judge significant changes in graphic data. It has taken over 20 years of studies to conclude that, just as people need to learn how to use statistical tests, that training in visual analysis is necessary to correctly interpret single subject data. Instead of adopting more objective methods of determining behavior change, it might be more beneficial to let competently, thoroughly trained behavior analysts make decisions about the effect of interventions.

Most of the research on visual inspection up to this point had focused almost exclusively on the effects of manipulating stimulus characteristics of the data on the final judgments made by raters exposed to them. Experimenters were forced to infer to which dimensions of the data the participants had reacted. Austin and Mawhinney (1999) decided to expand the research by collecting information from the participants as to which characteristics they believed they were responding. Concurrent verbal reports were collected as the subjects were presented with actual data one point at a time to determine when an intervention had been implemented. They reasoned that this process of data analysis was similar to how researchers and clinicians evaluate visual information in real time. By using actual data as the experimental data, the authors also had an absolute answer with which to compare their results, instead of comparisons to statistical tests.

In order to analyze the verbal reports, the authors used protocol analysis (Ericsson & Simon, 1993), a method used in cognitive psychology to collect and analyze verbal protocols. Ericsson and Simon (1993) recommend comparisons to a thorough task analysis. Since a formal task analysis has not been adopted by the field of behavior analysis, the first author (John Austin) compiled information and guidelines accumulated from different sources. These sources included guidelines by Parsonson and Baer (1978), information garnered from a highly
skilled and expert behavior analyst and results of the survey conducted by DeProspero and Cohen (1979) in their study.

Austin and Mawhinney (1999) collected verbal protocols from two “novices” and two “experts” in behavior analysis as they viewed each data point. After they had viewed all points, the participants were also required to make a decision as to whether there was evidence of an intervention and at what point the intervention had been implemented. Their reports were coded and analyzed and differences were found between the protocols of the novices and experts. The conclusions reached were: 1) experienced analysts tended to spend more time analyzing the data, from two to eight seconds more per data point 2) experienced analysts generated more data analysis-relevant segments, 3) experienced analysts were generally closer to the actual onset point of the intervention (data were not analyzed statistically so statistical significance was not determined and 4) experienced analysts considered each data point in the larger context of the surrounding data and the data set as a whole. The authors encouraged these results to be replicated in order to generalize the results. Unfortunately, recent studies have failed to replicate findings of differences between novices and experts, although the dependent variables have differed.

Recent Visual Analysis Research

Normand and Bailey (2003) examined the effects of a job aid on the performance of participants in visually analyzing data, as had been proposed by DeProspero and Cohen (1979) but never formally investigated.

Nineteen participants were randomly assigned to an active job aid group, a passive job aid group, or a control group to view a series of 60 within-subject graphs. None of the participants held higher than a bachelor’s degree in psychology or a related field. Additionally, no participant had worked for more than three years in a profession for which the visual analysis of behavioral data was a regular activity. Because of these criteria, this ensured that participants would be considered “novice” behavior analysts in terms of judging behavioral data.

The participants were asked to determine whether the graphs depicted a change that was attributable to the described intervention. Fifty-one within-subject behavioral graphs depicting ABA designs (A=Baseline, B=Intervention) were selected from research articles published in JABA. Each graph was accompanied by a short paragraph describing the type of participant
used, the type of behavior depicted, and the setting used for the study from which the graph was taken.

To aid in their analysis of whether the graphs depicted a change attributable to the intervention, some randomly selected participants were given one of two types of job aids, a passive job aid and an active job aid. While the two job aids had the same instructions and steps, the active job aid required that an answer be produced and recorded next to each step. The job aids were based on a visual analysis flowchart developed from two visual analysis examplars that participated in a previous study by Normand and Bailey (2002). As they viewed each graph, and when applicable, as they worked through each job aid, the participants were instructed to talk aloud.

Although five possible answers were available for each graph, the answers were collapsed into two discrete categories to allow for dichotomous scoring. The first category was labeled “Effect” and the second category was labeled “No Effect.” The answers of the participants were compared with those of a panel of three Ph.D. level behavior analysts, each with at least 20-years professional experience that independently reviewed each graph.

The experts unanimously agreed on the individual multiple-choice answers (i.e., a, b, c, or d) for 40% (24 out of 60) of the graphs. They most often agreed unanimously for the option that the intervention did not have an effect on the behavior. The experts were in total agreement for the dichotomous scoring (i.e., change or no change) for 67% of the graphs, most often agreeing on a decision of “no change” (65%, N=40). Additionally, participants were not in better agreement with the experts for those 40% of the graphs on which the experts were in unanimous agreement among themselves as to the specific interpretation (i.e., choice a, b, c, d, or e). Results indicated that a participant’s answers agreed with the answers determined by a panel of experts for fewer than 50% of the graphs, replicating the results obtained in many of the previous studies. Participants were more likely to report an intervention effect (72% of the graphs) than were the panel of experts (30% of the graphs), making experts more conservative in their decision-making. It should be noted that since the experts were not in unanimous agreement, the performances of participants should not be surprising.

The job aids did not improve participant agreement with the panel of experts. In fact, overall agreement with the panel of experts was low for both the active (43%) and passive job
aid (48%) groups. Agreement levels generally decreased across conditions, including the control group.

The verbal protocol analysis revealed that the job aids did increase attention to important graph characteristics (i.e., trend and variability) but the participants appeared to lack the necessary repertoire to modify their decisions based on the observed characteristics. In addition, the job-aid participants failed to accurately identify the relevant characteristics evident in the graphs they viewed; trends were often missed and variability underestimated. The data suggest that participants made their decisions based primarily on the degree of mean shift across conditions to the exclusion of other data characteristics, such as trend and variability. It seems that a few participants had received sufficient instruction to enable them to accurately interpret graphically displayed data. However, the participants did not receive any training on the job aids, feedback on their decisions or consequences for their conclusions. It may be that any one of these variables affects the ability of job aids to increase agreement with experts’ analyses.

Although a brief description of the depicted data was included for each graph, this might have actually hindered accurate interpretation on the part of the participants. For some participants, the clinical details seemed to distract them from the actual data, as evidenced by frequent statements about the validity of the intervention in question, about the type of participant studied, the measures used, etc. These results were opposite of those found by Richards et al. (1997), in which subjects’ characteristics did not affect decision-making. The differences might have resulted because more context information was provided in this study.

The most important limitation of prior studies that have compared the results of visual inspection and statistical analyses on the same data sets is that when the two methods produce discordant interpretations, it was not possible to determine which interpretation was correct (Matyas & Greenwood, 1990). Some studies have attempted to overcome this limitation by having visual inspectors interpret graphs for which the “correct” interpretation was known beforehand, rather than using a particular statistical procedure as “gold standard.” As has been stated throughout this chapter, there is no agreement on which statistical procedure is best to use with single-case design or if it is even appropriate to compare it with a statistical design.

One such study by Matyas and Greenwood (1990) had graduate students interpret graphs that had been created from a first order autocorrelation model. This allowed the authors to
program whether there was a treatment effect and whether the data were serially dependent. The authors then used the data to estimate the type of Type I and Type II errors made by the participants. The authors found that judges made a statistically significant higher percentage of Type I errors when compared to Type II errors, especially when the autocorrelation was greater than zero and the amount of random error was three standard deviations or greater.

However in the Matyas and Greenwood study (1990) only one graph was created from each set of autoregressive parameters, leaving open the possibility that some of the graphs were not representative of the model parameters used to create those graphs. Fisher et al. (2003) tested a data set from the previous study and determined that it was not a representative and as a result over inflated the percentage of Type I errors made by the judges.

Fisher’s study (2003) attempted to correct this error by using Monte Carlo procedures to evaluate visual inspection guides. Fisher’s goals were to further refine the training methods, structured criteria and visual aids needed to improve the reliability and validity of visual inspection methods.

In the first study, Fisher et al. (2003) developed a refined visual inspection method. First, graphs were created from a first-order autoregressive model. (These graphs were not included in the published article.) Using the split-middle method (SM), a regression line from the baseline data was superimposed on each treatment phase. By varying the effect size, they were able to examine patterns of Type I and Type II errors made by the SM method. Through visual inspection, the authors determined that the SM method made Type I errors when the baseline data was either on an upward or downward trend, but the slope in baseline was due to sampling error and not programmed by the autoregressive model. In order to compensate for this deficiency, the authors added a second criterion line that was generated from the baseline mean and superimposed on the treatment phase. Based on the binomial test, the same number of data points had to fall above (or below) both lines, hence creating the dual-criteria (DC) method. Kadzin (1982) had recommended that the SM method could be applied to data and the number of treatment points that fell above (or below) the superimposed trend line counted. Then, the binomial formula could be used to calculate the probability of that number occurring by chance. This coupling of the binomial formula with the SM method would probably increase the reliability of visual inspection, whereas Bailey (1984) found that the SM line alone
produced only modest gains. The problem, as pointed out by Fisher et al. (2003), is that the accuracy of the binomial formula has been found to decrease if there is serial dependence in the data series (Crosbie, 1987), thereby making the validity questionable. However the DC method relies on a pre-specified number of points, based on the binomial test, falling above the trend line. Since the formula creates data with serial dependence, this suggests an error in judgment by the authors.

A more conservative version of the DC method (the CDC method) was created after the DC method was tested at several different levels of autocorrelation and it was found that the rate of Type I errors was unacceptably high. This might be due to the faulty use of the binomial test with serially dependent data. The authors state that they raised the criterion lines several different values until they judged that .25 standard deviations represented a reasonable compromise between Type I and Type II errors. The number of errors is not reported and it seems a subjective decision was made.

The authors then created 30,000 data sets from 20 different combinations of effects and autocorrelation of their autoregressive model by using Monte Carlo simulations and programmed evaluations by each of the methods; the two new methods, the DC and CDC methods, the original SM method and two statistical methods commonly recommended for use with single-case designs, GLM and ITSE. (No samples of these graphs were included in the published article.)

The authors then compared the rates of both Type I and Type II errors. Their results indicated that the CDC method was the only interpretative procedure that controlled Type I error rates for short data series with and without autocorrelation and maintained reasonable power levels.

Even though the results of the first study showed that the CDC method was the preferred one of those tested, Fisher et al. (2003) ran these studies concurrently and the remaining studies were done with the DC method. Since the target responses are identical in both methods, the results of the two studies should still be applicable to either method. Nevertheless, this claim still needs to be validated and the results of both of these studies replicated, using the CDC method.
In the second study, five staff members were trained to apply the DC method to 20 A-B graphs and their interpretations were analyzed for improved accuracy due to the training. The behavior therapists had at least four months of daily exposure to visual inspection and interpretation of single-case designs, however, their background in behavior analysis was not provided. The graphs were generated using the same autoregressive model as in the first study, but included a modification to produce graphs of different lengths. (No samples of these graphs were included in the published article.)

Since the purpose of the training was to teach how to correctly implement the DC method, graphs were eliminated if interpretation with the DC method resulted in an error; that is, the DC method concluded that there was effect when the autoregressive model was programmed with an effect size of zero. As in previous studies, Fisher et al. (2003) tried to use graphs that had more subtle differences and they eliminated graphs with obvious treatment effects, such as large level changes, little variance, no trending, and large mean shifts. These four factors have been recognized in previous studies as being important variables in visual analysis. A majority of the studies have chosen to not include those graphs with obvious changes and to concentrate on those with more subtle differences. It may be that a more successful training program would take learners from the easiest to most difficult interpretations so that they learn to start recognizing how these variables affect their decision-making.

Once all of the graphs that met the above criteria were eliminated, graphs were divided into four groups. The first group of graphs contained those that met the DC criteria. The other three groups consisted of graphs that did not meet the criteria. The graphs in the second group met only one of the criteria; the third group fell one point short of meeting one of the criteria; and, the fourth group included all the other graphs that did not fall in the previous three groups. Each participant reviewed packets with 20 graphs, which consisted of 10 graphs from the first group and 10 graphs from a set combination of the other three groups. This was done to ensure that each packet contained an equal number of graphs with treatment effects and no treatment effects and that the difficulty level of each packet was fairly equal.

Fisher et al. (2003) included graphs with a programmed effect size of .5 of a standard deviation. The effects were so subtle that in real life they are probably unlikely to be judged to have socially significant differences. Since behavior analysts are looking for robust changes,
training staff to detect such subtle differences is a misuse of training time and resources. The authors do point out that small but reliable effects are sometimes important, such as small decreases in mean blood pressure which can produce clinically important health benefits. This problem can be solved by adjusting the ordinate scale on which the data is graphed. In this way, small differences can readily be viewed during visual analysis. It bears repeating that a small effect that is statistically significant may not have a social or clinical significance to the participant.

The five participants received training between baseline and treatment on how to analyze the graphs with the DC criteria lines. In the training, which lasted an average of 15 minutes, they were presented with a training graph to expose them to the DC method, a graph which met the criteria and one graph which did not meet the criteria, so that they were exposed to correct “yes” and “no” responses. The data sheet on which they recorded their answers (Was there a reliable treatment effect?) had a table showing the number of points that had to be above both criterion lines for different treatment lengths and the rules for using these criterion lines. No mention is made of whether participants were allowed to ask questions or received feedback. In addition, it is not clear whether the training was standardized for all participants.

In baseline, the participants had an average of 55% correct interpretations. This number is somewhat higher than the percentage agreement in previous studies (Richards et al., 1997; Normand & Bailey, 2003). However, no breakdown is provided for which treatment effects the most errors were made. The participants did state that they had difficulty making interpretations when one or more points overlapped with the criterion lines, making it difficult to determine whether the center of the data point fell above or below the line. This situation is more likely to occur with small treatment effect sizes than larger ones. There is also no breakdown as to the agreement between participants for each graph. This information would be helpful in identifying characteristics that make graphs difficult to interpret.

After training, the mean average for correct interpretations increased to 94%, the highest agreement level reached in a study. In addition, the DC method also decreased both Type I and Type II errors from a mean of 24% and 20% to .25% and 6%, respectively. The DC methods shifted the participants’ biases toward more conservative types of errors.
In the third study, Fisher et al. (2003) investigated whether the DC method could be presented to train a large group quickly and effectively. The participants were 87 adults, randomly assigned to two groups, attending a workshop on behavior analysis at an annual meeting of a state chapter of the Association for Behavior Analysis. Even though no demographics were provided on the participants, this author assumes that most of the participants were behavior analysts and many were probably certified behavior analysts. This makes the participant pool a little different than the previous study, since as a whole, these participants probably had more experience in visual inspection. Gathering some personal information would have helped to glean important characteristics of these participants.

The 60 graphs that were used were created in the same manner as in the previous study, but in this case, graphs with obvious treatment effects were not eliminated. It may be that the graphs in this study were easier to interpret than the graphs of the five participants in Study Two. No mention was made as to whether the graphs were presented in order of difficulty. Both groups interpreted the first set of 20 graphs together. The graphs, which were in a PowerPoint™ presentation format, were projected onto a screen and participants were asked to mark their data sheets as to whether or not a treatment effect was reliable. The participants had unlimited time to judge each graph, but the authors reported that sessions generally lasted about five to 15 minutes.

After this baseline session, Group B was excused for a break while Group A remained behind to receive training. The training was identical to the training provided in the previous study, except for the fact that the graphs were projected on a screen instead of in paper format. Immediately after training, Group A viewed another set of graphs with the DC criterion lines on them. They then were dismissed for a break while Group B returned and received a second baseline session, after which they received training. Then, Group A returned and both groups were presented with a final treatment session. It is unclear why a reversal design was not employed on the group as a whole.

The initial baseline mean for both groups was 71%. This percentage is higher than the percentage obtained from the participants in Study Two, reinforcing the assumption that these participants were more experienced in visual analysis. During the first treatment session for Group A, while Group B received a second baseline, there was a marked increase in correct
interpretations for Group A to 96% (from 72%) and a slightly marginal increase in Group B to
73% (from 70%). In the final treatment session for both groups, Group A remained at high
levels (98%) and Group B increased markedly to 93%. The authors also note that during the
initial baseline only one participant achieved 90% mastery level, after training, 73 (84%) of the
participants achieved this mastery level. No calculations were made as to the number of Type I
and Type II errors made by the participants either before or after training. The authors
concluded that the DC method can be used to train large groups in a short period of time,
improving mean level in correct interpretations and attaining mastery performance. It would be
interesting to investigate the reasons why some participants did not achieve mastery after
training.

In the discussion section, the authors address the possible critique that the DC and CDC
methods are just a hybrid of visual inspection and statistical methods. The authors state that the
DC method was created by visual inspection and interpretation of the pattern of errors made by
the SM method. The use of the binomial test was only a starting point to empirically test the
performance of the DC method using Monte Carlo simulation, which led to the creation of the
CDC method. However, the inclusion of values from the binomial test was not critical to the
development or refinement of these interpretive methods. They assert that the numerical criteria
could have been created by having a panel of experts sort the graphs into effect categories
(effect, no effect) and, then, setting the criteria so that the DC method closely approximated the
decisions of the panel. This would have provided more social validity to their method,
especially since studies have found that it is difficult even for experts to agree.

The authors also caution about using the CDC method as an endpoint for training on
visual inspection. Instead, they suggest that it has the potential to be a useful first step. The
DC or CDC methods do not provide behavior analysts with the underlying concepts that lead to
rejection or acceptance of treatment effects. It is probable that after removing the criterion
lines, interpretation levels would return to baseline levels. While improving performance in
visual inspection and increasing agreement is obviously a goal, it is important for behavior
analysts to be able to identify the variables that affect their decision-making.
Hypotheses

From the results of previous research and pilot work, I believe, both job aids will increase accuracy of participants, they both will increase inter-judge agreement and they both will reduce Type I errors. Participants will be more likely to use the Active job Aid methods after they have been removed. Even though the CDC method will be easier to train and take less time to train, the Active Job Aid will have a longer-lasting effect on the participants’ agreement with the experts.

Study

My study expanded the visual analysis research in several important ways. First, it evaluated whether systematically training participants to analyze visually displayed within-subject data was effective. Some of the training allowed participants to not only recognize important graph characteristics, such as mean shift, trend and variability, but made participants utilize them to make accurate interpretations about the data. Data analyzed included various dependent variables, such as time to train, agreement between participants and experts and number and type of errors committed. Second, this study replicated the methodology of previous studies and compared their effectiveness. As suggested in the study by Fisher et al. (2003), some participants used the CDC method to interpret single-case designs. However, unlike Fisher et al. (2003), these participants analyzed graphs from published reports as opposed to graphs created from hypothetical data. In addition, graphs employed the reversal design, which demonstrate more experimental control than the A-B design used in the Fisher et al. (2003) study. Other participants employed the active job aid methodology employed in Normand and Bailey (2003). The job aids utilized in that study did not improve participants’ agreement with a panel of experts. Even though the participants were required to fill-in answers for the active job aid, they were not required to utilize the results to make their decisions, as they were in this study. Finally, this study evaluated which method was easier to teach, and which one had the most durable effects on the behavior of the participants.
METHOD

Participants

Twelve participants were selected, all of whom had completed at least one course during which within-subject research methodology was taught. Even though the range of degrees varied, no participant held higher than a master’s degree in psychology or related field. Additional qualifications included that the participant was not a Level I Board Certified Behavior Analyst; that is all participants, if they were board certified, had held this certification for less than three years. All participants were asked to participate in two sessions, each lasting no longer than 60-minutes. At the end of the experiment, each participant was given a post-study questionnaire asking them to indicate, among other things, whether they believed the job aid helped them to accurately analyze the data (see Appendix G).

Stimulus Materials

Forty-four within-subject behavioral graphs depicting ABA designs (A=Baseline, B=Intervention) were selected from research articles published in JABA. To have met the criteria for selection, the graphs must: (a) have at least two baseline phases and one intervention phase, (b) contain at least five data points per phase, (c) depict only one data series, (d) have each phase clearly labeled, and (e) have the x and y axes clearly labeled and scaled. Some of the published graphs were modified from their original format to create an ABA design. For example, a graph that had more than one intervention depicted between baseline phases had one of the intervention or baseline phases removed so that only the ABA depiction remained. In addition, although all of the identifying information was removed from the graphs, most of the graphs retained the same scale. If the scale was altered, it was because it was necessary for the inclusion of the CDC lines. The graphs were a subset of the graphs utilized in Normand and Bailey (2003).

The selected graphs were enlarged or reduced so that all were as similarly sized as possible. Each of the graphs were presented to participants on a single sheet of paper, with a graph at the top and a multiple-choice question at the bottom asking the participants to indicate his or her decision as to whether a change in behavior that is attributable to the intervention was present on the graph above. The participants were also asked what graph feature(s) was/were
the most important when making their decision. They were able to circle all the answers that applied from the choices of mean, variability and trend. (See Figure 1 for an example.)

Each participant viewed 14 packets with four sets of graphs in each packet. Each packet consisted of two graphs depicting the A option, one graph depicting the B option and one graph depicting the C option. The fourth packet consisted of the same four randomly chosen graphs. The fourth packet was represented throughout each phase so that a direct comparison could be made as to whether participants’ decisions changed with the presence of either job aid.

Four participants (chosen randomly) were given the Active Job Aid (See Appendix E) to help in their analysis of some of the graphs. Briefly, the job aid required that an answer be produced and recorded for each step. The job aid is based on a visual analysis flowchart developed from two visual analysis exemplars that participated in a previous study on visual analysis (Normand & Bailey, 2002). The job aid used in this study is an altered version of the active job aid in Normand and Bailey (2003).

Each of the participants who were given the job aid received instructions and a standardized one-on-one training between baseline and treatment on how to analyze the graphs with the job aid. After completing a group of graphs without any job aid (the number of groups varied in order to conform to the multiple-baseline design of the study), the subject encountered a note in their package to call the experimenter over to receive training before they continued evaluating graphs. At this point, the experimenter introduced the job aid and three training graphs to expose them to a graph which met Option A criteria (the intervention caused a change in behavior), a graph which met Option B criteria (the intervention did not cause a change in behavior) and one graph which met Option C criteria (can’t make a decision based on the available information), so that they were exposed to all three responses on the data sheet (See Figures 2, 3, and 4). The experimenter asked the participant to answer each step of the job aid and gave feedback as to the correctness of his or her response. For example, after the participant drew the mean for baseline, intervention and return to baseline as indicated in Step 1, the experimenter placed a transparency over the graph that showed the actual means. If the participant’s means differed largely from the actual means, the participants were given a new blank graph and asked to draw the means again. This step was repeated until the participant drew means that were within an acceptable range of the actual mean. The same process was
1. Please circle a, b, or c based on your interpretation of the data above.
Based on the data above, I _____ (Circle only one answer)
   a. believe that the intervention caused a change in behavior.
   b. believe that the intervention did not cause a change in behavior.
   c. can’t make a decision based on the available information.

2. Based on the data, which feature(s), if any, was/were the most important when making your decision about the effect of the intervention? Keep in mind that it may have been lack of a feature that influenced your decision. (Circle all that apply.)
   a. Mean Shift
   b. Variability
   c. Trend

Figure 1. An example of the graphs and questions presented to all participants.
1. Please circle a, b, or c based on your interpretation of the data above.
   Based on the data above, I ____ (Circle only one answer)
   
   a. believe that the intervention caused a change in behavior.
   b. believe that the intervention did not cause a change in behavior.
   c. can’t make a decision based on the available information.

2. Based on the data, which feature(s), if any, was/were the most important when making your decision about the effect of the intervention? Keep in mind that it may have been lack of a feature that influenced your decision. (Circle all that apply.)
   
   a. Mean Shift
   b. Variability
   c. Trend

*Figure 2.* An example of the Option A graph presented to participants during training for the CDC and Active Job Aid groups.
1. Please circle a, b, or c based on your interpretation of the data above.

   Based on the data above, I _____ (Circle only one answer)
   
   a. believe that the intervention caused a change in behavior.
   b. believe that the intervention did not cause a change in behavior.
   c. can't make a decision based on the available information.

2. Based on the data, which feature(s), if any, was/were the most important when making your decision about the effect of the intervention? Keep in mind that it may have been lack of a feature that influenced your decision. (Circle all that apply.)

   a. Mean Shift
   b. Variability
   c. Trend

Figure 3. An example of the Option B graph presented to participants during training for the CDC and Active Job Aid groups.
1. Please circle a, b, or c based on your interpretation of the data above. 
   Based on the data above, I _____ (Circle only one answer)  
   
   a. believe that the intervention caused a change in behavior.  
   b. believe that the intervention did not cause a change in behavior.  
   c. can’t make a decision based on the available information.

2. Based on the data, which feature(s), if any, was/were the most important when making your decision about the effect of the intervention? Keep in mind that it may have been lack of a feature that influenced your decision. (Circle all that apply.)

   a. Mean Shift  
   b. Variability  
   c. Trend

*Figure 4.* An example of the Option C graph presented to participants during training for the CDC and Active Job Aid groups.
Figure 5. An example of the mean data for the Option B graph used in the Active Job Aid training. A transparency was made of this graph. During training, when the participant drew the means, this transparency was placed on top of the training graph to compare the actual means with the drawn means of the participant.
Figure 6. An example of the trend data for the Option B graph used in the Active Job Aid training. A transparency was made of this graph. During training, when the participant drew the trend, this transparency was placed on top of the training graph to compare the actual trends with the drawn trends of the participant.
Figure 7. An example of the variability data for the Option B graph used in the Active Job Aid training. A transparency was made of this graph. During training, when the participant drew the variability, this transparency was placed on top of the training graph to compare the actual variability with the drawn variability of the participant.
repeated with the steps requiring the participant to draw the trends and variability (See Figures 5, 6, and 7). The length of training depended on how many trials the participant required before the mean, trend and viability lines were drawn correctly and each step of the job aid was completed.

Once the participant completed the job aid, he or she filled out the original data sheet presented, using the answers provided by the job aid. After the training was completed, the participant was instructed to continue to the next group of graphs and asked to fill out the job aid for each one. After the training, the participant did not receive any feedback by the experimenter.

Four participants (chosen randomly) viewed graphs with conservative dual criteria (CDC) lines superimposed on the intervention data. The participants were required to count the number of data points above both lines and compare them to a control number, based on the number of intervention data points, to determine whether there was a change attributable to the intervention. The CDC lines were developed by Fisher et al. (2003) in a study in which these lines were used to increase the reliability of participants in visual analysis of data. The participants viewed each graph with the CDC lines as two separate graphs, side-by-side. The first graph contained the first baseline phase and the intervention. The second graph contained the intervention and the second baseline phase. (See Appendix E) The participants applied the criteria to the set of graphs. In order to reach a conclusion that the intervention caused a change in behavior, both graphs must have met the criteria. If both of the graphs did not meet the criteria or the first graph did not meet the criteria, the participants were trained to conclude that the intervention did not change the behavior. If only the first graph met the criteria, the participants were trained to conclude that there was not enough information to determine whether or not there was a treatment effect.

In addition to standardized general instructions, the participants who were in the CDC Lines group received training between baseline and treatment on how to analyze the graphs with the CDC criteria lines. In the training, they were presented with three training graphs to expose them to the CDC method, a graph which met the criteria, one graph which did not meet the criteria and one that halfway met the criteria so that they were exposed to Option A (the intervention caused a change in behavior), Option B (the intervention did not cause a change in
behavior), and Option C (can't make a decision based on the available information) responses. (See Figures 2, 3and 4.) Along with the data sheet on which they recorded their answers (Was there a reliable treatment effect?), there was a table showing the number of points that had to be above or below both criterion lines for different treatment lengths and the rules for using these criterion lines. (See Appendix E) The experimenter asked the participant to answer each step of the CDC graph and gave feedback as to the correctness of his or her response. Once the participant completed the CDC aid, he or she filled out the original data sheet presented, using the answers provided by the CDC aid. After the training was completed, the participant was instructed to continue the next group of graphs and was asked to fill out the CDC aid for each one. After the training, the participant did not receive any feedback from the experimenter.

Additionally, a panel of three Ph.D. level behavior analysts with at least 20-years of professional experience, independently reviewed each graph to 1) verify whether or not a change was depicted, and 2) indicate the feature (e.g. trend, variability, mean shift) that they felt was the feature that most influenced their decision. Initially, only graphs on which all three experts agreed, and met CDC and Active Job Aid criteria were selected, for a total of 36 graphs. In order to increase the amount of data to be analyzed, eight graphs were added and divided into two more packets. These two sets of graphs contained graphs on which two out of the three experts agreed on the effect of the intervention as shown in Appendix C. Each of these sets contained graphs on which experts dichotomously agreed that the intervention did not cause a change in behavior.

Experimental Design

Although only 44-graphs were be selected, four graphs were re-presented throughout each phase of the study so that a direct comparison could be made as to whether participants’ decisions changed with the presence or absence of a job aid or of the CDC lines or due to practice (for the Control Group). Therefore, the participants viewed a total of 56 graphs.

The selected participants were randomly assigned to one of three groups: Active Job Aid, CDC Lines or Control. The Active Job Aid and CDC Lines groups used the Active Job Aid and the CDC Lines Aid, respectively, as described above. Those participants in the Active Job Aid and CDC Lines groups viewed the graphs according to an ABA design.
Independent Variable Manipulation

Baseline (A phase)

The first phase for the Active Job Aid group involved each participant viewing sets of graphs without the job aid. The first phase for the CDC Lines group involved each participant viewing sets of graphs without the CDC lines. The first participant in each group viewed four sets of four graphs each. The other participants viewed additional sets in order to adhere to the multiple-baseline design of the experiment.

Training (B phase)

The second phase for the job aid group involved each participant viewing sets of graphs while using the Active Job Aid. The second phase for the CDC Lines group involved each participant viewing sets of graphs with the CDC lines. The first participant in each group viewed four sets of four graphs each. The other participants viewed additional sets in order to adhere to the multiple-baseline design of the experiment.

Maintenance phase (A phase)

The second A phase for the Active Job Aid group involved each participant viewing sets of graphs without the job aid, as in the first A phase. The second A phase for the CDC Lines group involved each participant viewing sets of graphs without the CDC lines, as in the first A phase. All participants in each group viewed four sets of four graphs each.

Those participants in the Control Group viewed all 56 graphs without the use of any extraneous aids. The purpose of this group was to account for the possibility that performance might improve over time because of practice, independent of any intervention.

Dependent Variables

Multiple-choice Answers

The answers for the multiple-choice questions about effect were categorized as follows. For unanimous agreement, the average percentage of correct interpretations in baseline, treatment and maintenance as compared to expert agreement was calculated for each participant. Based on the unanimous scoring, the percentage of correct interpretations was displayed visually for each of the 14 sets of graphs for each group, consisting of four participants each. The data was graphed to conform to a multiple-baseline across participants design for each of the three phases described above.
For dichotomous agreement, although three possible answers were available, the answers were collapsed into two discrete categories to allow for dichotomous scoring. The first category was labeled “Effect” and was comprised of answer A (increase or decrease due to intervention, causing a change). The second category was labeled “No Effect” and was comprised of answers B and C (did not cause a change or unable to make a decision, respectively). Based on the dichotomous scoring, the average percentage of correct interpretations in baseline, treatment and maintenance as compared to expert agreement was calculated for each participant.

Based on the dichotomous scoring, the percentage of correct interpretations was displayed visually for each of the 14 sets of graphs for each group, consisting of four participants each. The data was graphed to conform to a multiple-baseline across participants design for each of the three phases described above.

The dichotomous scoring also allowed the relative rates of Type I and Type II errors to be easily determined. Such an analysis assessed whether visual analysis is a conservative method of data analysis, as has been suggested by many (e.g., Baer, 1977; Parsonson & Baer, 1992; Michael, 1974), or whether it is actually less conservative as suggested by Matyas and Greenwood (1990). This scoring also allowed an assessment of which type of aid leads to more conservative choices, if any.

The percentage of correct interpretations based on the dichotomous scoring also allowed comparison of the job aids ability to improve performance. Additionally, the return to baseline phase allowed a determination of whether the job aids produce lasting changes.

**Time To Train**

The time spent by the researcher training each participant in the Active Job Aid group and the CDC Lines group was recorded. An average time to train each condition was also calculated. For the Active Job Aid group, the number of trials needed to correctly draw the mean, trend and variability for the training graphs was also calculated.

**Data Analysis**

All categorical data from survey question one was scored unanimously and dichotomously (Change or No Change) as described above and then plotted graphically and analyzed visually. Additionally, the following variables were also be analyzed.
Panel of Experts.

The experts ratings of “easy” and “difficult” were compared with the correct and incorrect interpretations of the participants. The analysis evaluated whether the graphs that were considered “easy” by the experts were interpreted correctly more those that were rated as “difficult.”

Participant Agreement.

Overall agreement with the panel of experts was computed across all conditions (baseline, intervention and maintenance for each group, the Active Job Aid group, the CDC Lines group and the Control group. Individual performances from each group were also computed.

Type I and Type II Errors.

The total number and type of each error, in percentage form, that the participants made was computed.

Job Aid Comparisons

The time to train participants individually and as a group was compared. Percentage of correct interpretations was compared across groups of participants to determine which job aid, if any, had a greater effect on the scores of the participants. Percentage of scores was also compared to determine which job aid, if any, provided a lasting behavior change once the job aid was removed. Anecdotal reports from participants were also reported as to ease of employing job aid and likeability of job aid.
RESULTS

Overall

As predicted, training with the CDC Lines Job Aid was successful in changing the behavior of the participants and improving agreement with the experts. However, training with the Active Job Aid did not reliably improve agreement with the experts across all participants, as expected. After training for the Active Job Aid was completed, and no further feedback was provided to the participants, they were unable to accurately identify the features that were the most influential in their decision-making. While this was expected for the CDC Lines Job Aid, it was surprising that after being directed towards these traits by the Active Job Aid, participants were not very accurate in their identifications.

Agreement with the experts by the control group was lower than the agreement for the two job aids as expected. Even though the Active Job Aid took longer to train, it did not improve agreement with the experts, as would have been expected. In actuality, the Active Job Aid doubled the number of Type II errors during the intervention condition. Both job aids did decrease the number of Type I errors during intervention phase as hypothesized. As predicted the CDC Lines job Aid was easier to train (at least in length of time), however it’s effects were non-lasting. It was assumed, that the Active Job Aid would have produced more durable effects, but this was only the case for two out of the four participants.

Individual Subject Data

Unanimous Agreement

CDC Lines group. Unanimous agreement required the participant and expert to agree on their choice (A, B or C) of what effect the intervention had on the behavior. Individual performances per packet for unanimous agreement with the experts are displayed in a multiple-baseline format in Figure 8.

Participant 104 had a stable baseline with a small drop in the last data point, producing a short downturn. Intervention data were also stable, except for Packet 7 in which there was a sharp drop. This drop decreased the mean shift from baseline to intervention and increased the variability of the data in intervention. Data in the maintenance condition were also stable and decreased below baseline levels.
Participant 102 had a variable baseline. After the CDC lines were introduced the data became fairly stable with a reduction in the variability of the data. When the CDC lines were removed, data again became more variable.

Participant 103 had a variable baseline, with an overall downtrend. After the CDC Lines Job Aid was introduced, the data became very stable. When the job aid was removed, the data again became more variable and also had a downtrend, but a greater average than baseline.

Participant 101 also had variable baseline data with a downtrend. After the CDC Lines Job Aid was introduced percent agreement increased and became less variable than baseline. After the job aid was removed the data again became very variable; the data increased for one packet and then sharply dropped, but again increased in the last data point.

The CDC Lines Job Aid increased performance to 100% unanimous agreement in all participants, except for packets 6 and 7 in Participants 102, packet 9 in Participant 104, and packet 10 in Participant 103. After the CDC Lines Job Aid was removed, unanimous agreement immediately decreased in participants 102,104, and 101. Participant 103 continued to achieve 100% agreement for the next packet before percent agreement decreased.

The CDC Lines job aid increased overall unanimous agreement with the experts. For three out of four participants (101, 102, and 103), the agreement increased above the percent agreement in baseline and actually decreased the range of scores. While in the other participant (104) the CDC Lines Job Aid did increase the range of scores, it also increased the number of packets in which the participants had 100% agreement with the experts.

*Active Job Aid group.* Individual performances per packet for unanimous agreement with the experts are displayed in a multiple-baseline format in Figure 9. For all participants, baseline data were variable with a downtrend. Unanimous agreement data did not increase when the job aid was introduced except for participant 204. For the other three participants, intervention data remained variable within the same range as baseline. After the job aid was removed, the mean remained the same for Participants 203 and 201. For Participant 204, the mean continued to increase even after the job aid was removed. In contrast, the mean for agreement for Participant 202 continued to decrease across all conditions.

The job aid did not increase over all unanimous agreement with the experts. For three out of four participants, the agreement did not increase above the percent agreement in baseline
and actually increased the range of scores. In addition, the job aid did not increase the number of packets in which the participants had 100% agreement with the experts.

Control group. Individual performances per packet are displayed in Figure 12. The unanimous agreement scores ranged from 0% per packet to 100% per packet. However, only Participant 4 achieved 100% agreement with the experts and only on one packet.

All participants had variable data. Participants 1 and 4 show a downtrend in their data. Participant 2 had the highest mean of all Control group participants and had a slight upward trend throughout the data. Participant 3 also had variable data but remained fairly level.

Dichotomous Agreement

CDC Lines group. Dichotomous agreement required the participant and expert to agree on whether or not the intervention had an effect on the behavior. Individual performances per packet for dichotomous agreement with the experts are displayed in a multiple-baseline format in Figure 10. Participant 104 had a more variable baseline with the percent agreement decreasing after the second packet and in a downtrend. The CDC Lines Job Aid had the same effect in intervention with Participant 104. For all participants, data in the maintenance condition were also more variable as compared to the intervention condition and the mean was lower than baseline.

Participant 102 had a variable baseline with a sharp drop at the end. After the CDC Lines Job Aid was introduced data became fairly stable. Variability once again increased in the maintenance condition.

Participant 103 had a more stable baseline than the unanimous agreement baseline, but still with a downward trend. After the CDC Lines Job Aid was introduced, the dichotomous data became very stable. When the job aid was removed, the data again became more variable and also showed a downtrend, but a greater average than baseline.

Participant 101 also had variable baseline data with a downward trend. After the CDC Lines Job Aid was introduced percent agreement increased and became very stable. After the job aid was removed the data again became very variable; the data remained the same as intervention for one packet and then dropped, but again increased in the last data point.

The CDC Lines Job Aid also increased dichotomous agreement with the experts for all four participants. Dichotomous agreement was only below 100% in three packets (6, 7, and 9)
across participants. For all participants, average dichotomous agreement decreased after the CDC Lines Job Aid was removed. For Participants 102 and 104, average dichotomous agreement dropped lower than baseline levels, while for Participants 101 and 103, it decreased from intervention levels but remained above baseline levels. All four participants had more packets in which the dichotomous agreement with the experts was 100% after the job aid was introduced or before it was removed.

*Active Job Aid group.* Individual performances per packet for dichotomous agreement with the experts are displayed in a multiple-baseline format in Figure 11. All participants also had a variable baseline with a downtrend. The job aid did not decrease variability on the intervention data for Participants 203 or 202. The job aid decreased the variability in the intervention data for Participants 204 and 201, whose intervention data were stable. For all participants except for Participant 202, data in the maintenance condition were as variable as compared to baseline.

For Participants 203 and 204, the job aid increased dichotomous agreement. There was a further increase in agreement when the job aid was removed. Participant 202, on the other hand, had a decrease in dichotomous agreement throughout each phase. Participant 201 had a high baseline dichotomous agreement rate, which increased further when the job aid was introduced. Following the removal of the job aid, dichotomous agreement also decreased when the job aid was removed but was higher than baseline level.

The job aid increased dichotomous agreement with the experts for three out of the four participants. In those three participants, dichotomous agreement was only below 75% in one packet (5) for Participant 203. For these three participants, dichotomous agreement remained high even after the job aid was removed with no participant having less than 75% agreement in all packets. All three participants had more packets in which the dichotomous agreement with the experts was 100% after the job aid was introduced.

However, for Participant 202, the job aid had a variable effect on dichotomous agreement. The participant had only two packets in which 100% agreement with the experts was achieved in the intervention and maintenance conditions. In addition, the participant’s agreement with the experts also dropped to 25% for two packets after the intervention condition.
**Control group.** Individual performances per packet for dichotomous agreement are displayed in Figure 12. All of the participants’ scores increased when calculated as dichotomous agreement. There were 11 total packets in which dichotomous agreement with the experts was 100%. Participants 1 and 4 still showed a downward trend in their data. Participant 2 had a more stable set of data with the variability reduced. Participant 3 also had variable data and a slight upward trend could be seen.

**Panel Of Experts**

Experts agreed unanimously with 36 of the 44 graphs (82%) in this study. For the other eight graphs, two out of the three experts agreed with the intervention effect. Therefore, dichotomous agreement was 100% for all graphs.

As the experts viewed each graph, they were also asked how easy or difficult that graph was to judge. With respect to the ratings of easy or difficult, the experts unanimously agreed for 52% of the graphs. For the other graphs (48%), a majority decision was reached for the ease and difficulty of the graph. That is, at least two out of the three experts agreed on each graph. It was evident that the panel of experts’ ratings of difficult or easy for each graph did not correspond with the percentage of participants who interpreted the graph correctly. As seen in Figure 13, percent agreement was not greater for the “easy” graphs; in fact, graph 48, which was rated as “easy” by the experts had the lowest percent agreement by participants (22%). Participants had 100% agreement with many graphs rated as difficult by the experts.

**Group Data**

**Unanimous Agreement**

**CDC Lines group.** Overall unanimous agreement with the panel of experts was 68% (range: 55% -92%) across all conditions. Agreement levels increased from 55% in baseline to 92% following the introduction of the CDC Lines Job Aid. A reduction similar to baseline levels was evident in the reversal condition (58%) as shown in Figures 14 and 15.

**Active Job Aid group.** Overall agreement with the panel of experts on unanimous agreement was 66% (range: 59% -73%) across all conditions. Agreement levels increased from 59% in baseline to 73% following the introduction of the Active Job Aid. Mirroring the CDC Lines Job Aid, a slight reduction in agreement level was evident in the reversal condition (67%) as shown in Figures 16 and 17.
Figure 8. Percent unanimous agreement with experts per packet across CDC Lines group participants.
Figure 9. Percent unanimous agreement with experts per packet across Active Job Aid group participants.
Figure 10. Percent dichotomous agreement with experts per packet across CDC Lines group participants. The open circles on the graph of Participant 104 indicate where the CDC Lines Job Aid provided the correct answer but the participant made a choice making error.
Figure 11. Percent dichotomous agreement with experts per packet across Active Job Aid group participants.
Figure 12. Unanimous (left graph) and dichotomous (right graph) percent agreement per packet with the experts for Control group participants.
Dichotomous Agreement

**CDC Lines group.** Overall dichotomous agreement with the panel of experts was relatively high for the CDC Lines group job aid group with an average of 77% (range: 69% - 94%) agreement across all conditions as shown in Figure 20. Agreement levels increased from 69% in baseline to 94% following the introduction of the CDC Lines job aid. Agreement level decreased to baseline levels in the reversal condition (69%) as shown in Figures 14 and 15. The dichotomous agreement with the experts was greater in each condition than the unanimous agreement.

**Active Job Aid group.** Overall agreement with the panel of experts on dichotomous agreement was relatively high for the active job aid group with an average of 75% (range: 63% - 82%) agreement across all conditions. Agreement levels increased from 63% in baseline to 81% following the introduction of the active job aid. Unlike the CDC Lines Job Aid pattern, agreement level remained stable in the reversal condition (82%). As shown in Figures 16 and 17, the dichotomous agreement with the experts was greater in each condition than the unanimous agreement data.

**Control group.** As shown in Figure 18, overall unanimous agreement with the experts for the Control group was pretty low at 50% (range: 40% - 54%). Overall unanimous agreement with the panel of experts for the control group was similar to that for the baseline and return to baseline conditions of the job aid groups. Dichotomous agreement with the experts was also relatively high, averaging 72% (range: 58% - 89%) across participants.

**Type I and Type II Errors**

Overall, participants were equally likely to detect non-existent effects (Type I errors) as to fail to detect existent effects (Type II errors). As shown in Table 1, Type I errors and Type II errors each accounted for 50% of all errors (N=157). Although, the CDC Lines Job Aid decreased the number of errors (N=47) as compared to the control group (N=60), the percentage of Type I errors increased to 55%. Type II errors accounted for 45% of the errors. Therefore, the CDC Lines Job Aid group made more Type I errors (i.e., detected non-existent effects). Only 11% of all errors were made during the intervention condition, as compared to 51% in the baseline condition (51%) and 38% in the
Figure 13. Percentage of participant agreement with experts for each graph. The gray bars indicate that the panel of experts rated the graph as “easy.” The white bars indicate that the panel of experts rated the graph as “difficult.” See Appendix C for arrangement of graphs into each packet.
Figure 14. Average percent unanimous and dichotomous agreement with experts for each CDC Lines group participant.

Figure 15. CDC Lines group participant-by-participant accuracy across experimental conditions for unanimous agreement (left graph) and dichotomous agreement (right graph) with experts.
Figure 16. Unanimous and dichotomous average agreement by Active Job Aid group across experimental phases.

Figure 17. Active Job Aid group participant-by-participant accuracy across experimental conditions for unanimous agreement (left graph) and dichotomous agreement (right graph) with experts.
Figure 18. Control group participant-by-participant accuracy across experimental conditions for unanimous and dichotomous agreement with experts.
maintenance condition. As can be seen in Table 2, only 8% of the Type I errors and 14% of Type II errors occurred during the intervention phase.

Although the Active Job Aid decreased the number of errors (N=50) as compared to the control group (N=60), Type II errors only accounted for 40% of the errors. Therefore, the Active Job Aid group also made more Type I errors. Type I errors accounted for 60% of all errors (N=50). Twenty-six percent of the errors were made in the intervention condition, as compared to 52% in the baseline condition and 22% in the maintenance condition. However, only 10% of the Type I errors occurred during the intervention phase. As can be seen in Table 3, 50% of the Type II errors were made during the intervention condition. The CDC Job Aid decreased Type II errors more than the Active Job Aid did.

The Control group made more total errors than the CDC group or the Active Job Aid group (N=60). As can be seen in Table 1, Type I errors accounted for 38% (N=23) of all errors and Type II errors accounted for 62% (N=37) of all errors. Overall, the job aid groups made a lesser number of errors as compared to the control group, but made a larger amount of Type I errors. However, most errors were made during the baseline and maintenance phases. The job aids changed the likelihood of detecting non-existent effects. The CDC Lines Job Aid also decreased the likelihood of failing to detect existent effects. The Active Job Aid, however, increased the likelihood of failing to detect existent effects. Both job aids decreased the number of total errors.

The CDC Lines Job Aid and the Active Job Aid both increased sensitivity (the conditional probability that case x is correctly classified. However, the Active Job Aid increased specificity (conditional probability that case x is not correctly identified) while the CDC Lines Job Aid decreased it. Thus, the CDC Lines Job Aid increased both the positive (probability that a case is x if a classifier classifies it as x) and negative predictive power (the probability that case x is not x if the classifier does not classify it as x) of the participants, while the Active Job Aid only increased positive predictive power.
**Time**

The average time for the Control participants to complete all packets was 34 minutes (range: 33-35). All participants in this group completed 14 packets, each packet containing four graphs for a total of 56 graphs. Thus, participants analyzed 1.7 graphs per minute.

The CDC Lines group completed the packets in an average of 57 minutes (range: 40-89). The participants in this group averaged 13 packets, for a total of 52 graphs. Participants analyzed .9 graphs per minute. The average time spent analyzing graphs in the baseline condition was 16 minutes and in the return to baseline condition was 13 minutes. The participants analyzed 1.2 graphs per minute in both conditions. Participants in the CDC Lines group spent an average of 29 minutes analyzing the graphs in the intervention condition for an average of .6 graphs per minute. The average training time for this group was eight minutes. There were three graphs in the training portion, so participants averaged .4 graphs per minute.

The Active Job Aid group completed the packets in an average of 59 minutes (range: 50-69). The participants in this group averaged 13 packets, for a total of 52 graphs. Participants analyzed .9 graphs per minute. The average time spent analyzing graphs in the baseline condition was 14 minutes and in the maintenance condition was 10 minutes. The participants analyzed 1.5 graphs per minute in the baseline condition and 1.6 graphs per minute in the maintenance condition. Participants in the Active Job Aid group spent an average of 36 minutes analyzing the graphs in the intervention condition for an average of .5 graphs per minute. The average training time for this group was 28 minutes. There were three graphs in the training portion, so participants averaged .1 graphs per minute.

Overall, control participants analyzed more graphs per minute than either the CDC Lines group or Active Job Aid group. As shown in Tables 4 and 5, in the CDC Lines group, the addition of the job aid extended the time participants spent analyzing each graph in the intervention condition so that about twice the time was spent looking at each graph as compared to the other two conditions. As shown in Table 4 and 5, in the Active Job Aid group, the addition of the job aid extended the time participants spent analyzing each graph in the intervention condition so that three times as much time was spent looking at each graph as compared to the other two conditions. The rate for each intervention condition for both groups
was roughly equal. The CDC Lines group was trained in one-third less time than the Active Job Aid group.

**TABLE 1**

*Breakdown of Type I and Type II Errors by group*

<table>
<thead>
<tr>
<th>GROUP</th>
<th>TYPE I</th>
<th>TYPE II</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDC JOB AID</td>
<td>26</td>
<td>21</td>
<td>47</td>
</tr>
<tr>
<td>ACTIVE JOB AID</td>
<td>30</td>
<td>20</td>
<td>50</td>
</tr>
<tr>
<td>CONTROL</td>
<td>23</td>
<td>37</td>
<td>60</td>
</tr>
<tr>
<td>TOTAL</td>
<td>79</td>
<td>78</td>
<td>157</td>
</tr>
</tbody>
</table>
TABLE 2
Total and Percentage of Type I and Type II errors made by participants across all conditions for the CDC Lines Job Aid group

<table>
<thead>
<tr>
<th>CDC LINES</th>
<th>TYPE I</th>
<th></th>
<th>TYPE II</th>
<th></th>
<th>TOTAL</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#</td>
<td>%</td>
<td>#</td>
<td>%</td>
<td>#</td>
<td>%</td>
</tr>
<tr>
<td>BASELINE</td>
<td>10</td>
<td>38</td>
<td>14</td>
<td>67</td>
<td>24</td>
<td>51</td>
</tr>
<tr>
<td>INTERVENTION</td>
<td>2</td>
<td>8</td>
<td>3</td>
<td>14</td>
<td>5</td>
<td>11</td>
</tr>
<tr>
<td>MAINTENANCE</td>
<td>14</td>
<td>54</td>
<td>4</td>
<td>19</td>
<td>11</td>
<td>38</td>
</tr>
<tr>
<td>TOTAL</td>
<td>26</td>
<td>55</td>
<td>21</td>
<td>45</td>
<td>47</td>
<td>100</td>
</tr>
</tbody>
</table>

TABLE 3
Total and Percentage of Type I and Type II errors made by participants across all conditions for the Active Job Aid group

<table>
<thead>
<tr>
<th>ACTIVE JOB AID</th>
<th>TYPE I</th>
<th></th>
<th>TYPE II</th>
<th></th>
<th>TOTAL</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#</td>
<td>%</td>
<td>#</td>
<td>%</td>
<td>#</td>
<td>%</td>
</tr>
<tr>
<td>BASELINE</td>
<td>19</td>
<td>63</td>
<td>7</td>
<td>35</td>
<td>26</td>
<td>52</td>
</tr>
<tr>
<td>INTERVENTION</td>
<td>3</td>
<td>10</td>
<td>10</td>
<td>50</td>
<td>13</td>
<td>26</td>
</tr>
<tr>
<td>MAINTENANCE</td>
<td>8</td>
<td>27</td>
<td>3</td>
<td>15</td>
<td>11</td>
<td>22</td>
</tr>
<tr>
<td>TOTAL</td>
<td>30</td>
<td>60</td>
<td>20</td>
<td>40</td>
<td>50</td>
<td>100</td>
</tr>
</tbody>
</table>
TABLE 4

*Total minutes per group to complete experiment*

<table>
<thead>
<tr>
<th>GROUP</th>
<th>MINUTES</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDC JOB AID</td>
<td>57</td>
</tr>
<tr>
<td>ACTIVE JOB AID</td>
<td>59</td>
</tr>
<tr>
<td>CONTROL</td>
<td>34</td>
</tr>
</tbody>
</table>

TABLE 5

*Total minutes to complete each condition for CDC Lines group and Active Job Aid group*

<table>
<thead>
<tr>
<th>CONDITION</th>
<th>CDC</th>
<th>ACTIVE</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASELINE</td>
<td>16</td>
<td>14</td>
</tr>
<tr>
<td>TRAINING</td>
<td>8</td>
<td>28</td>
</tr>
<tr>
<td>INTERVENTION</td>
<td>29</td>
<td>36</td>
</tr>
<tr>
<td>MAINTENANCE</td>
<td>13</td>
<td>10</td>
</tr>
</tbody>
</table>
Data Characteristics

CDC Lines Group
As can be seen in Figure 19, mean shift had the highest average percent agreement in the CDC group. Participants averaged 61% agreement in baseline, 61% in intervention and 69% in the maintenance condition. As can be seen in Figure 20, overall average percent agreement for mean shift was 64%.

As can be seen in Figure 19, average percent agreement on variability was low. Participants averaged only 31% agreement in baseline, 51% in intervention and 47% in the maintenance condition. As shown in Figure 20, overall average percent agreement for variability was 44%.

As can be seen in Figure 19, overall average percent agreement for trend was also low. Participants averaged only 30% agreement in baseline, 41% in intervention but 72% in the maintenance condition. Overall average percent agreement for variability was 48% as shown in Figure 20.

Active Job Aid Group
As can be seen in Figure 21, mean shift also had the highest average percent agreement in the Active Job Aid group. Participants averaged 62% agreement in baseline, 59% in intervention and 55% in maintenance. Overall average percent agreement for mean shift was 59% as shown in Figure 22.

As can be seen in Figure 21, average percent agreement on variability was low. Participants averaged only 31% agreement in baseline, 51% in intervention and 47% in the maintenance condition. Overall average percent agreement for variability was 44%, as shown in Figure 22.

Average percent agreement for trend was also low. As seen in Figure 21, participants averaged 44% agreement in baseline, 45% in intervention and 52% in the maintenance condition. Overall average percent agreement for variability was 47% as shown in Figure 22.

Control Group
As can be seen in Figure 23, mean shift also had the highest average percent agreement in the Control group. Overall average percent agreement for mean shift was 55%. Overall average
percent agreement for variability was 51%. Average percent agreement for trend was also low. Overall average percent agreement for trend was 43%.

Summary

Both job aid groups and control group participants failed to accurately identify the characteristics evident to experts in the graphs that they viewed. Mean shift had the highest percent agreement followed by variability and, finally, trend. The job aids did not increase percent agreement with the experts on these characteristics. That is, the job aids did not effectively control the behavior of correctly identifying the graph characteristics that were the most important when making their decision about the effect of the intervention.

Intra-subject Agreement

CDC Lines Group

As seen in Table 6, based on comparisons of the repeated packet, intra-subject agreement ranged from 50% in Participant 101 to 100% agreement in Participant 102. Both Participants 103 and 104 had 83% intra-subject agreement. For graph 1, between-subject agreement was 67% for two participants and 100% for the other two. Graph 34 had three participants with 100% agreement and one participant with 67% agreement. Graph 44 had one participant with 0% agreement, two with 67% and one with 100% agreement. Graph 5 had three participants with 67% agreement and one participant with 100% agreement.

Active Job Aid Group

As seen in Table 7, based on comparisons of the repeated packet, intra-subject agreement ranged from 50% in Participant 204 to 92% agreement in Participant 201. Participant 202 had 55% intra-subject agreement and 203 had 75% intra-subject agreement. For graph 1, between-subject agreement was 0% for one participant, 67% for two participants and 100% for the remaining participant. Graph 34 had two participants with 100% agreement and two participants with 67% agreement. Graph 44 had one participant with 0% agreement, two with 67% and one with 100% agreement. Graph 5 had all participants with 67% agreement.

Control Group
Figure 19. Average percent agreement of CDC Lines Job Aid participants with experts on graph characteristics for each condition.

Figure 20. Overall average percent agreement with experts on graph characteristics by CDC Lines Job Aid participants,
Figure 21. Average percent agreement of Active Job Aid participants with experts on graph characteristics for each condition.

Figure 22. Overall average percent agreement with experts on graph characteristics by Active Job Aid participants.
Figure 23. Overall average percent agreement of Control participants with experts on graph characteristics.
As seen in Table 8, based on comparisons of the repeated packet, intra-subject agreement ranged from 83% in Participant 2 to 100% agreement in Participant 3. Participants 1 and 4 both had 92% intra-subject agreement. For graph 1, between-subject agreement was 67% for two participants and 100% for the other two participants. Graph 34 had all participants with 100% agreement and two participants with 67% agreement. Graph 44 and Graph 5 had three participants each with 100% agreement, and one participant with 67% agreement.

TABLE 6
*Intra-subject agreement for repeated package by CDC Lines group participants*

<table>
<thead>
<tr>
<th>CDC 101</th>
<th>Condition</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Graph #</td>
<td>Baseline</td>
<td>Intervention</td>
<td>Maintenance</td>
<td>Agreement</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>C</td>
<td>A</td>
<td>A</td>
<td>67%</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>C</td>
<td>A</td>
<td>A</td>
<td>67%</td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>C</td>
<td>B</td>
<td>A</td>
<td>0%</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>C</td>
<td>C</td>
<td>A</td>
<td>67%</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td>50%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CDC 102</th>
<th>Condition</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Graph #</td>
<td>Baseline</td>
<td>Intervention</td>
<td>Maintenance</td>
<td>Agreement</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td>100%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CDC 103</th>
<th>Condition</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Graph #</td>
<td>Baseline</td>
<td>Intervention</td>
<td>Maintenance</td>
<td>Agreement</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>C</td>
<td>A</td>
<td>C</td>
<td>67%</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>C</td>
<td>B</td>
<td>C</td>
<td>67%</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td>83%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CDC 104</th>
<th>Condition</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Graph #</td>
<td>Baseline</td>
<td>Intervention</td>
<td>Maintenance</td>
<td>Agreement</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>67%</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>A</td>
<td>C</td>
<td>A</td>
<td>67%</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td>83%</td>
<td></td>
</tr>
</tbody>
</table>
TABLE 7

*Intra-subject agreement for repeated package by Active Job Aid group participants*

<table>
<thead>
<tr>
<th>ACTIVE 201</th>
<th>Condition</th>
<th>Graph #</th>
<th>Baseline</th>
<th>Intervention</th>
<th>Maintenance</th>
<th>Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td></td>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>34</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td></td>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>44</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td></td>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>5</td>
<td>C</td>
<td>C</td>
<td>A</td>
<td></td>
<td></td>
<td>67%</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>92%</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ACTIVE 202</th>
<th>Condition</th>
<th>Graph #</th>
<th>Baseline</th>
<th>Intervention</th>
<th>Maintenance</th>
<th>Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>A</td>
<td>C</td>
<td>C</td>
<td></td>
<td></td>
<td>67%</td>
</tr>
<tr>
<td>34</td>
<td>A</td>
<td>C</td>
<td>A</td>
<td></td>
<td></td>
<td>67%</td>
</tr>
<tr>
<td>44</td>
<td>A</td>
<td>B</td>
<td></td>
<td></td>
<td></td>
<td>0%</td>
</tr>
<tr>
<td>5</td>
<td>A</td>
<td>C</td>
<td>A</td>
<td></td>
<td></td>
<td>67%</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>55%</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ACTIVE 203</th>
<th>Condition</th>
<th>Graph #</th>
<th>Baseline</th>
<th>Intervention</th>
<th>Maintenance</th>
<th>Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td></td>
<td></td>
<td>67%</td>
</tr>
<tr>
<td>34</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td></td>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>44</td>
<td>A</td>
<td>C</td>
<td>C</td>
<td></td>
<td></td>
<td>67%</td>
</tr>
<tr>
<td>5</td>
<td>A</td>
<td>C</td>
<td>C</td>
<td></td>
<td></td>
<td>67%</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>75%</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ACTIVE 204</th>
<th>Condition</th>
<th>Graph #</th>
<th>Baseline</th>
<th>Intervention</th>
<th>Maintenance</th>
<th>Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>B</td>
<td>C</td>
<td>A</td>
<td></td>
<td></td>
<td>0%</td>
</tr>
<tr>
<td>34</td>
<td>B</td>
<td>A</td>
<td>A</td>
<td></td>
<td></td>
<td>67%</td>
</tr>
<tr>
<td>44</td>
<td>A</td>
<td>C</td>
<td>C</td>
<td></td>
<td></td>
<td>67%</td>
</tr>
<tr>
<td>5</td>
<td>A</td>
<td>C</td>
<td>C</td>
<td></td>
<td></td>
<td>67%</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>50%</strong></td>
</tr>
</tbody>
</table>
### TABLE 8

Intra-subject agreement for repeated package Control group participants

<table>
<thead>
<tr>
<th>Control 1</th>
<th>Graph #</th>
<th>Baseline</th>
<th>Baseline</th>
<th>Baseline</th>
<th>Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>A</td>
<td>C</td>
<td>C</td>
<td>67%</td>
</tr>
<tr>
<td></td>
<td>34</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>44</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>100%</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>92%</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Control 2</th>
<th>Graph #</th>
<th>Baseline</th>
<th>Baseline</th>
<th>Baseline</th>
<th>Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>B</td>
<td>B</td>
<td>A</td>
<td>67%</td>
</tr>
<tr>
<td></td>
<td>34</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>44</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>B</td>
<td>B</td>
<td>C</td>
<td>67%</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>83%</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Control 3</th>
<th>Graph #</th>
<th>Baseline</th>
<th>Baseline</th>
<th>Baseline</th>
<th>Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>34</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>44</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>100%</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>100%</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Control 4</th>
<th>Graph #</th>
<th>Baseline</th>
<th>Baseline</th>
<th>Baseline</th>
<th>Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>34</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>44</td>
<td>C</td>
<td>A</td>
<td>A</td>
<td>67%</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>100%</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>92%</strong></td>
</tr>
</tbody>
</table>
DISCUSSION

Applied behavior analysts strive to change socially significant behavior for the benefit of individuals and society. Their interventions are designed and implemented to have robust effects on behavior. If the changes in behavior are small, interventions are not considered successful. The goal is to be able to predict and control the behavior of individuals by identifying the effects of variables on individual’s behavior. This requires the use of within-subject research.

In within-subject research, the effects of an intervention are observed at different points in time for the same organism. The effects are replicated (or reproduced) at different times so that a judgment can be made based on the overall pattern of data. The particular experimental arrangement employed dictates how baseline performance can be used to predict future performance. The technique of choice to analyze the results in applied behavior-analytic work has been visual analysis.

Visual analysis has a long history of being criticized because of the seemingly inherent subjectivity it. Previous research has demonstrated that even skilled behavioral researchers disagree about the effects depicted by graphed behavioral data. Many of the earlier studies in this area compared the reliability and validity of visual analysis with various statistical tests and found it lacking both. Many behavior analysts have indicated that the problem is the lack of standard decision rules to determine intervention effects. Recent research has focused on the graph characteristics that affect decision-making and the differences in decision-making by experts and novices in the field of behavior analysis. This has led to current research on the most effective way to train behavior analysts to reliably and validly interpret graphically displayed data.

This study demonstrates that the accuracy of Board Certified Behavior Analysts in deciding intervention effects can be improved at least on a short-term basis. Overall, participants unanimously agreed with the experts on 46 out of 56 graphs when the job aids were employed. Participants dichotomously agreed with the experts for 49 out of the 56 graphs when the job aids were employed. For a majority of the graphs, both experts and participants agreed on whether or not a graph showed an intervention effect.

Although training behavior analysts to unanimously agree on intervention effects is the ultimate goal, dichotomous agreement is an important first step. Dichotomous agreement is
achieved when all trainees agree with experts on which graphs show an intervention effect (choice A in this experiment) and which graphs do not (choices B and C in this experiment). After this distinction is being reliably made, then, making more subtle determinations of the no-effect graphs can be achieved. It is important to make these determinations because they lead to different courses of action. If it is decided that the intervention did not cause a change in behavior (choice B), then the intervention should not be continued. However, if there is not enough information to make a decision (choice C), more data collection is required before a making a determination (i.e., the treatment needs to continue.)

One of the main findings was that the CDC Lines Job Aid had an effect on unanimous agreement during the intervention phase. There was a definite increase in mean shift from the baseline condition when the CDC Lines Job Aid was introduced. In addition, there was also a reversal of the downtrend from baseline. For three out of four participants, variability in the unanimous agreement scores was decreased or remained unchanged, however, for the one participant, variability increased. When the CDC Lines Job Aid was removed, mean levels returned towards baseline. Trend remained stable or took a downward shift. While in some participants variability increased, it did not greatly overlap with the intervention data, except for one participant.

The Active Job Aid did not seem to have improved performance for three out of the four participants. While it may have produced some difference in means across intervention sessions, the variability of scores remained high and there was a consistent downtrend across conditions. This was similar to the performance of the Control group, who also had very variable data and two control participants had similar downtrends.

Examination of the data as dichotomous agreement led to an increase in the average percent agreement for all three groups. The CDC Lines Job Aid also had an effect on dichotomous agreement during the intervention phase. There was a definite increase in mean from the baseline condition when the CDC Lines Job Aid was introduced. In addition, there was also a reversal of the downtrend from baseline. For three out of four participants, variability in all dichotomous agreement scores was decreased or remained unchanged. In the dichotomous agreement scores, three out of the four participants had decreased variability. When the CDC Lines Job Aid was removed, mean levels returned towards baseline. Trend remained stable or
took a downward shift. While in some participants variability increased, it did not greatly overlap with the intervention data, except for one participant. It is important to note that the impact of the CDC Lines Job Aid would have been greater with no variability for all participants but Participant 104 made an error in choice making when using the CDC Lines Job Aid. Even though the job aid had provided the correct answer, the participant circled a different choice. This looks to be the participant’s error because even she indicated that she did not agree with the choice.

Visually analyzing the data for the Active Job Aid group, three out of the four participants did not increase their dichotomous agreement with the experts after the Active Job Aid was introduced. In these three participants, data remained variable and overlap was evident between data points in all conditions. In the control group, variability in and trending of scores were also not really affected.

The CDC Lines Job Aid was helpful to the participants. The amount of time spent training was less than 10 minutes. Beginners could be trained with the CDC Lines Job Aid, since this method increased accuracy. However, its effects were not long-lasting; as soon as the job aid was removed, performance decreased. Although the participants anecdotally reported liking the job aid, it did not teach them how to make correct interpretations on their own.

Unlike previous research, participants in this study were more conservative in their interpretations than other studies have found their participants to be. According to experts in the field, being more conservative (i.e., failing to detect significant effects) is preferable than detecting non-existent effects. However, a standard of acceptable proportion of Type II to Type I errors has not been determined. Nevertheless, in this study participants were as equally likely to detect non-existent effects (Type I errors) as to fail to detect effects (Type II errors). The control group was more conservative overall than either of the job aid groups.

The CDC Lines group made slightly more Type I errors than Type II errors. However, only two Type I errors and two Type II errors were made during the intervention. The CDC Lines group decreased the number of overall errors during intervention but did not make a difference in the type of error.

The Active Job Aid group was also more likely to make Type I errors than Type II errors. Only three Type I errors were made during the intervention condition, so most of the Type I
errors were also made during baseline and maintenance conditions. However, half of Type II errors were committed in the intervention condition. The Active Job Aid made the participants more conservative in their choices, maybe too conservative.

Even though the participants were instructed to take as much time as they wanted in analyzing the data, the participants completed the experiment quickly, perhaps too quickly. Control participants analyzed almost two graphs per minute. This lack of time devoted to evaluating the data may have led to the low percent agreement levels with the experts. Even though data was not collected from the experts on time taken to complete the analysis of the graphs, anecdotal reports suggest that evaluation took much longer for each graph.

Previous research that has evaluated time spent on each graph has found that participants devoted more time to analyzing the data (Austin & Mawhinney, 1999; Normand & Bailey, 2003.) In those studies participants were asked to speak aloud and their evaluations were recorded. The requirement of speaking out loud produced more time devoted to each graph. However, those studies did not find that participants’ verbal reports agreed with experts’ decisions. The verbal protocol analyses revealed that participants did attend to important characteristics, but, unlike experts, they appeared to lack the necessary repertoire to modify their decisions based on the observed characteristics. Although no specific times are provided, Fisher et al. (2003) mention in their study that participants did not spend any more than 15 minutes evaluating the graphs in either the baseline or intervention conditions. Since the experts anecdotally reported that they spent much more time analyzing the data in this study than did participants, this might be a variable worth testing in the future.

The job aids increased the time spent evaluating each graph as compared to baseline and maintenance conditions. The CDC Lines Job Aid doubled the time spent evaluating each graph from one graph per minute in baseline and maintenance, to one graph every two minutes in intervention. The Active Job Aid increased evaluation time from one-and-half graphs per minute in baseline to one graph every two minutes in intervention. Since both job aids asked questions about the characteristics of the data, it seems plausible that the extra time was spent evaluating the effects of these characteristics on the data. The CDC Lines Job Aid asked questions numerically while the Active Job Aid asked for descriptive data, which might account for the increased time in the Active Job Aid participants. Time spent evaluating data, however, did not
make a difference in correct interpretations from the participants. Again, participants did attend to important characteristics, but, unlike experts, they appeared to lack the necessary repertoire to modify their decisions based on the observed characteristics. Participants were more likely to evaluate each phase individually and not integrate their decisions while experts were more likely to look at both phases as a whole.

Participants in this study had trouble correctly identifying the graph characteristics that had the greatest impact on their decision about intervention effects. This was true even when the participants agreed with the experts on whether or not the intervention had an effect. It seems that participants underestimated the impact that trend and variability have on the interpretation of graphed behavioral data. This was apparent in both the training of the Active Job Aid and the results of all three groups. In the training, trend was the one characteristic that had to be redrawn by the participant after it was shown that it had been drawn incorrectly. Even though, the participants were able to accurately draw the direction of the trend, they were not able to accurately recognize the slope of the trend. Even with feedback, it took participants multiple trials to approximate the actual trend of the data. No participant drew the trend correctly on the first trial. The number of trials ranged from two trials to seven trials. In retrospect, the training may have swayed the participants to become more conservative in their choice making. The training was so rigid and participants were required to be extremely accurate in relation to actual mean shift, trend and variability, that multiple trials were needed for the participants to approximate them. This training might have made the participants doubt their confidence in previous decision-making, thus making them more likely to choose the option that they did not have enough information to make determinations of intervention effects. It might be that less rigorous training leads to the Active Job Aid having more of an effect on participant agreement with the experts.

Even though the job aids increased agreement with the experts on intervention effects, they still did not improve the ability to label those characteristics that had the greatest impact on their decisions. This was expected for the CDC Lines Job Aid since the CDC lines made the interpretations numerical in nature. However, the Active Job Aid asked the participants, specifically about the effects of these three characteristics, basing their answers on their decisions to these questions. Participants in this group made more errors on the variability of the
data. They erroneously overestimated the overlap of scores between baseline and intervention conditions. Though it made participants more conservative in their choice making, they were not able to correctly identify it as a determination factor in their decision-making.

Overall, participants were not reliable in their decision-making across conditions. This can be seen by the variable intra-subject agreement in all three groups. This is also corroborated by the fact that no participant mentioned having seen a graph previously. The Control group had the highest intra-subject agreement with the lowest agreement percentage above 80%. The CDC Lines had a large range from 50% to 100%, but three out of four participants had 83% or higher percent agreement. The lowest intra-subject agreement percentage was in the Active Job Aid group, in which no participant had 100% intra-subject agreement and three out of four participants had 75% or lower percent agreement.

Even though the CDC lines group had all participants with 100% agreement within the intervention condition, intra-subject agreement varied between baseline and maintenance conditions, supporting the finding that the CDC Lines Job Aid controlled the participants’ behavior. The Active Job Aid intra-subject data also corroborates previous findings that the job aid made them more conservative. There were more “no effect” choices in the intervention condition, as compared to baseline and maintenance. When the job aid was available, participants were more likely to choose a “no effect” option than when the job aid was not present, thus increasing their intra-subject variability. The Control group had the highest intra-subject agreement, presumably because nothing affected their choice-making, even practice.

In this study, single-subject methodology was utilized to answer questions about the effectiveness of the job aids in increasing reliability and validity of visual analysis. For the CDC Lines Job Aid, experimental control was shown both through the reversal design for each participant and through the multiple-baseline design across participants. However, for the Active Job Aid, the ability for it to control the participants’ behavior was not clearly demonstrated. This job aid seems to have made the participants make too many conservative choices, negating its effectiveness to decrease detection of non-existent effects. That is, most participants were more likely to decide that they needed more information in order to decide, while the experts believed that they had enough data to make a decision.
When compared to other studies in this area, I believe that I have extended the research in this area. Only a few previous studies have utilized actual JABA graphs. Jones et al. (1978) utilized them in one of the first studies to analyze visually presented data. More recently, research by Austin and Mawhinney (1999) also included actual data. As mentioned previously in this manuscript, the graphs in this study were a subset of the data set utilized in Normand and Bailey (2003). Though using actual data does constrain the research, it provides external validity to the results.

The same graph characteristics have been evaluated in many of the studies. However, prior to 1997, research by DeProspero and Cohen (1979), Wampold and Furlong (1981), Furlong and Wampold (1982), and Ottenbacher (1990) focused almost exclusively on the effects of manipulating stimulus characteristics of the data on the final judgment made by raters exposed to them. This led to experimenters having to infer to which dimensions of the data the participants had reacted. Austin and Mawhinney (1999) extended the research in this area by collecting information from the participants as to which characteristics they believed they were responding. As in Normand and Bailey (2003), this study has strengthened the finding that experts are responding to certain graph characteristics of the data, i.e., mean shift, variability and trend. However, neither study was able to train participants to respond in the same way.

Other studies have also found differences in performance between experts and novices. Furlong and Wampold (1982) were one on the first to try to establish a group of experts to judge intervention effects. They concluded that the variable manner in which experts view and interpret graphs was responsible for the lack of interjudge agreement reported in previous research. (DeProspero & Cohen, 1979; Jones et al., 1978) As in Normand and Bailey (2003), the experts in this study were behavior analysts with their doctoral degrees who had been practicing in this field for over 20-years.

The definition of a novice has also varied throughout the years from participants having only one course in single-subject design (Warpold & Furlong 1982), to “novice” behavior analysts, graduate students in behavior analysis (Normand & Bailey, 2003) and recently employed behavior analysts. (Fisher et al., 2003) This study extends the research by having “working” behavior analysts who are recently certified and are out in the field and are responsible for making intervention effect decisions as part of their job.
While proponents of visual analysis contend that it is conservative in nature (Baer, 1977; Parsonson & Baer, 1992; Michael; 1974), the data collected by Matyas and Greenwood (1990), along with Richard’s et al. (1997), demonstrated that novices and experts are more likely to detect non-existent effects. Normand and Bailey (2003) found that even though their participants did make more Type I errors, the experts did not. Fischer et al. (2003) claimed that previous research has inflated the number of Type I errors committed by participants; this study has collaborated Fisher’s finding. Participants were just as likely to detect non-existent effect, as to fail to detect existent effects. In addition, unlike most of the previous research already discussed above, Fisher et al. (2003) and this study found higher agreement rates with the experts than any other previous study.

This research study further replicates the findings of Fisher et al. (2003). While Fisher et al. (2003) did their study utilizing DC lines, they later discovered that changing the standard deviation of the DC lines produced new CDC lines which had a better ability to lower likelihood of detecting non-existent effects, while maintaining adequate power. In their discussion, the authors suggested the since the CDC method had identical target responses as the DC method, the results of their studies should be applicable to the CDC method. This claim has been validated, by replicating the results of the CDC method in changing participants’ agreement scores.

At this point, everyone agrees that novices and experts vary in their skills to analyze visual data. However, recent research suggests that this disparity is not as large as it was thought to be. In addition, my study demonstrates that a supplemental aid has improved the performance of novices to expert levels. While this aid has been mathematical in nature, further research can discover ways to improve the aid based on visual analysis methods. The characteristics to which the participants have to attend have been identified; a better way for the aid to control the participant’s behavior is now required.

**Limitations**

Different characteristics of the participants might have affected the results of this study. First, there seems to be limited motivation by the participants to perform at their optimum levels. This can be seen by the downtrend in the data in some of the control participants, as well, the mean performance levels in the maintenance conditions, which were lower than the original
baseline. So, although they anecdotally seemed nervous and wanted to perform well in front of the experimenter, this was not enough of an incentive to maintain optimal performance throughout the study.

In addition, some participants did better in baseline than other participants. It may be that the job aids are not necessary for those who already have the pertinent skills to visually analyze data. Especially, in the Active Job Aid group, those who performed at a higher average in baseline, decreased performance levels in intervention and maintenance phases. As stated previously, the job aid probably made already conservative participants, even more so, decreasing their ability to make concrete decisions (i.e., there was or was not an effect caused by the intervention), and instead opted for not having enough information to make a decision. If the participants had been screened, selecting only those who had low percentage agreement, it might have been easier to show an improvement with the job aids.

Another limitation of this study is the small number of graphs that were analyzed by participants. Since each participant acted as their own control, the Control participants could have been given one of the two job aids to increase the number of subjects in the multiple-baseline design, increasing the ability to analyze the effectiveness of the independent variable. The number of graphs also allowed the means of the packets and the conditions to be too sensitive to outliers. Median data for each packet may have given a better picture of the participants’ agreement levels with the experts.

Future Research

One area that has not yet been addressed is the lack of programmed consequences for either correct or incorrect answers regarding effects. It may even be necessary to provide consequences for attending to the data or the job aid. It is likely that without consequences some participants lacked the motivation to perform to the best of their ability, as can be seen in the downward trend of some of the participants in the Control group. As behavior analysts, we are aware that antecedent manipulations, like the use of job aids, without any consequences will improve performance on a short-term basis but rarely produce long-lasting effects. This is important because many behavior analysts work in environments where there is a lack of feedback by other professional behavior analysts to prevent behavior drift.
Since experts perform better than novices, we also need to identify the type of experiences that foster improved performances. It seems that a more thorough, fine-grained analysis of experts’ behaviors is still required. It may be possible to glean important information from videotaping experts while they analyze graphs, as well as have verbal reports in order to have a more complete picture of the chains of behavior involved in their decisions. Having job aids made from the results of these analyses might improve participants’ agreements. Such information could also allow for programmed exposure to these experiences for students.

Finally, the graphs in this study only use JABA graphs that had been accepted for publication, and, therefore, presumably, the editors agreed with the conclusions of the authors about effects. Students need to be exposed to data that is rejected by journals in order to form a more complete picture of the visual analysis of data. It may also be useful to systematically present the graphs for analysis. At first, students would analyze unambiguous data in which all experts agree that there is or is not an effect. As graphs become more ambiguous, the job aids seem to become more effective in their ability to help make decisions. Students, also, need to be presented graphs in which experts do not agree in order to learn the graph characteristics that make visual agreement difficult for everyone.

**Implications for Clinical Practice**

The data and interpretations that have been discussed have important implications for not only researchers, but also, clinicians, as well as educators. As noted earlier, research in this area has involved novices of all kinds. It is clear, based on their inability to reach the same conclusions as experts, that undergraduate and graduate students do not receive enough training in visual analysis and the interpretation of graphically displayed data. As they become professionals, delivering behavior analytic services to different populations, they are required to make decisions based on their analysis of the visually displayed data. Given the lack of thorough training and practice, is it surprising that these behavior analysts make mistakes when analyzing clinical data?

As professionals, behavior analysts are under pressure from other groups, teachers, parents, even corporations, to change the behavior of the individuals for whom they are working. Therefore, it would not be surprising if behavior analysts are probably biased in seeing effects in their data that may not be truly be there. Behavior analysts need to be encouraged to refrain
from isolating their data reviewing process. The importance of presenting data at conferences and at local review committees is an important responsibility for behavior analysts. Only by having other behavior analysts review our work can we truly be confident about the effects of our interventions.

In order to prevent situations like this, more extensive and empirically validated methods of instruction seem warranted. In order to develop such instructional methods and improve visual analysis in general, future research needs to be able to further identify and isolate those variables that influence visual data analysis. In addition, there needs to be an adoption by behavior analysts of one concrete method of data interpretation.

Despite its apparent shortcomings and limitations, visual analysis remains an important area of research and practice. Slowly, but surely, the field of psychology is starting: 1) to place more emphasis on within-subject data, 2) to increase the use of graphical display of data and 3) to abandon the sole reliance on statistical methods. (Morgan & Morgan, 2001; Smith, Best, Stubbs, Archibald, & Roberson-Nay, 2002; Wilkinson, 2002) By looking at the individual, psychology is beginning to understand that as Skinner believed, in order to predict and control behavior, it is necessary to understand the difference among each participant and not treat variability as unwanted “noise.”
APPENDIX A

JABA Source Information for the Experimental Graphs

<table>
<thead>
<tr>
<th>Graph</th>
<th>Volume</th>
<th>Page</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graph 1</td>
<td>26</td>
<td>188</td>
<td>top</td>
</tr>
<tr>
<td>Graph 2</td>
<td>33</td>
<td>538</td>
<td></td>
</tr>
<tr>
<td>Graph 3</td>
<td>34</td>
<td>537</td>
<td>mid</td>
</tr>
<tr>
<td>Graph 4</td>
<td>21</td>
<td>248</td>
<td>bottom</td>
</tr>
<tr>
<td>Graph 5</td>
<td>21</td>
<td>248</td>
<td>top</td>
</tr>
<tr>
<td>Graph 6</td>
<td>20</td>
<td>136</td>
<td>top</td>
</tr>
<tr>
<td>Graph 7</td>
<td>20</td>
<td>136</td>
<td>bottom</td>
</tr>
<tr>
<td>Graph 8</td>
<td>24</td>
<td>287</td>
<td>top</td>
</tr>
<tr>
<td>Graph 9</td>
<td>24</td>
<td>285</td>
<td>bottom</td>
</tr>
<tr>
<td>Graph 10</td>
<td>24</td>
<td>285</td>
<td>top</td>
</tr>
<tr>
<td>Graph 11</td>
<td>22</td>
<td>75</td>
<td>top</td>
</tr>
<tr>
<td>Graph 12</td>
<td>32</td>
<td>105</td>
<td>bottom</td>
</tr>
<tr>
<td>Graph 13</td>
<td>31</td>
<td>522</td>
<td></td>
</tr>
<tr>
<td>Graph 14</td>
<td>31</td>
<td>177</td>
<td>middle</td>
</tr>
<tr>
<td>Graph 15</td>
<td>31</td>
<td>111</td>
<td>bottom</td>
</tr>
<tr>
<td>Graph 16</td>
<td>33</td>
<td>261</td>
<td>bottom</td>
</tr>
<tr>
<td>Graph 17</td>
<td>33</td>
<td>261</td>
<td>top</td>
</tr>
<tr>
<td>Graph 18</td>
<td>32</td>
<td>383</td>
<td>bottom</td>
</tr>
<tr>
<td>Graph 19</td>
<td>32</td>
<td>383</td>
<td>top</td>
</tr>
<tr>
<td>Graph 20</td>
<td>32</td>
<td>383</td>
<td>top</td>
</tr>
<tr>
<td>Graph 21</td>
<td>32</td>
<td>369</td>
<td></td>
</tr>
<tr>
<td>Graph 22</td>
<td>32</td>
<td>213</td>
<td>top</td>
</tr>
<tr>
<td>Graph 23</td>
<td>27</td>
<td>512</td>
<td>top</td>
</tr>
<tr>
<td>Graph 24</td>
<td>27</td>
<td>442</td>
<td>bottom</td>
</tr>
<tr>
<td>Graph 25</td>
<td>27</td>
<td>442</td>
<td>top</td>
</tr>
<tr>
<td>Graph 26</td>
<td>27</td>
<td>512</td>
<td>bottom</td>
</tr>
<tr>
<td>Graph 27</td>
<td>26</td>
<td>256</td>
<td>bottom</td>
</tr>
<tr>
<td>Graph 28</td>
<td>26</td>
<td>188</td>
<td>middle</td>
</tr>
<tr>
<td>Graph 29</td>
<td>32</td>
<td>131</td>
<td></td>
</tr>
<tr>
<td>Graph 30</td>
<td>32</td>
<td>383</td>
<td>bottom</td>
</tr>
<tr>
<td>Graph 31</td>
<td>20</td>
<td>79</td>
<td></td>
</tr>
<tr>
<td>Graph 32</td>
<td>24</td>
<td>285</td>
<td>middle</td>
</tr>
<tr>
<td>Graph 33</td>
<td>24</td>
<td>384</td>
<td>top</td>
</tr>
<tr>
<td>Graph 34</td>
<td>24</td>
<td>699</td>
<td></td>
</tr>
<tr>
<td>Graph 35</td>
<td>126</td>
<td></td>
<td>bottom</td>
</tr>
<tr>
<td>Graph 36</td>
<td>21</td>
<td>70</td>
<td></td>
</tr>
<tr>
<td>Graph 37</td>
<td>19</td>
<td>170</td>
<td>top</td>
</tr>
<tr>
<td>Graph 38</td>
<td>19</td>
<td>170</td>
<td>middle</td>
</tr>
<tr>
<td>Graph 39</td>
<td>19</td>
<td>170</td>
<td>bottom</td>
</tr>
</tbody>
</table>
### JABA Source Information for the Experimental Graphs

<table>
<thead>
<tr>
<th>Graph</th>
<th>26</th>
<th>200</th>
<th>top</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graph 1</td>
<td>27</td>
<td>439</td>
<td>top</td>
</tr>
<tr>
<td>Graph 2</td>
<td>20</td>
<td>297</td>
<td>bottom</td>
</tr>
<tr>
<td>Graph 2</td>
<td>20</td>
<td>296</td>
<td>middle</td>
</tr>
<tr>
<td>Graph 3</td>
<td>20</td>
<td>296</td>
<td>bottom</td>
</tr>
<tr>
<td>Graph 4</td>
<td>20</td>
<td>296</td>
<td>bottom</td>
</tr>
<tr>
<td>Graph 5</td>
<td>34</td>
<td>227</td>
<td>bottom</td>
</tr>
<tr>
<td>Graph 6</td>
<td>34</td>
<td>162</td>
<td>middle</td>
</tr>
<tr>
<td>Graph 7</td>
<td>33</td>
<td>617</td>
<td>top</td>
</tr>
<tr>
<td>Graph 8</td>
<td>22</td>
<td>75</td>
<td>bottom</td>
</tr>
<tr>
<td>Graph 9</td>
<td>22</td>
<td>81</td>
<td>middle</td>
</tr>
<tr>
<td>Graph 10</td>
<td>22</td>
<td>81</td>
<td>bottom</td>
</tr>
<tr>
<td>Graph 11</td>
<td>24</td>
<td>56</td>
<td>altered</td>
</tr>
</tbody>
</table>
APPENDIX B
COMPLETE SET OF EXPERIMENTAL GRAPHS
1. Please circle a, b, or c based on your interpretation of the data above. Based on the data above, I _____. (Circle only one answer)

a. believe that the intervention caused a change in behavior.
b. believe that the intervention did not change the behavior.
c. can’t make a decision based on the available information.

2. Based on the data, which feature(s), if any, was/were the most important when making your decision about the effect of the intervention? Keep in mind that it may have been lack of a feature that influenced your decision. (Circle all that apply.)

a. Mean Shift
b. Variability
c. Trend
1. Please circle a, b, or c based on your interpretation of the data above. Based on the data above, I _____ (Circle only one answer)
   a. believe that the intervention caused a change in behavior.
   b. believe that the intervention did not change the behavior.
   c. can’t make a decision based on the available information.

2. Based on the data, which feature(s), if any, was/were the most important when making your decision about the effect of the intervention? Keep in mind that it may have been lack of a feature that influenced your decision. (Circle all that apply.)
   a. Mean Shift
   b. Variability
   c. Trend
1. Please circle a, b, or c based on your interpretation of the data above. Based on the data above, I ____ (Circle only one answer)

a. believe that the intervention caused a change in behavior.
b. believe that the intervention did not change the behavior.
c. can’t make a decision based on the available information.

2. Based on the data, which feature(s), if any, was/were the most important when making your decision about the effect of the intervention? Keep in mind that it may have been lack of a feature that influenced your decision. (Circle all that apply.)

a. Mean Shift
b. Variability
c. Trend
1. Please circle a, b, or c based on your interpretation of the data above. Based on the data above, I _____ (Circle only one answer)

a. believe that the intervention caused a change in behavior.
b. believe that the intervention did not change the behavior.
c. can’t make a decision based on the available information.

2. Based on the data, which feature(s), if any, was/were the most important when making your decision about the effect of the intervention? Keep in mind that it may have been lack of a feature that influenced your decision. (Circle all that apply.)

a. Mean Shift
b. Variability
c. Trend
1. Please circle a, b, or c based on your interpretation of the data above. Based on the data above, I _____ (Circle only one answer)
   a. believe that the intervention caused a change in behavior.
   b. believe that the intervention did not change the behavior.
   c. can’t make a decision based on the available information.

2. Based on the data, which feature(s), if any, was/were the most important when making your decision about the effect of the intervention? Keep in mind that it may have been lack of a feature that influenced your decision. (Circle all that apply.)
   a. Mean Shift
   b. Variability
   c. Trend
1. Please circle a, b, or c based on your interpretation of the data above.
Based on the data above, I _____ (Circle only one answer)

   a. believe that the intervention caused a change in behavior.
   b. believe that the intervention did not change the behavior.
   c. can’t make a decision based on the available information.

2. Based on the data, which feature(s), if any, was/were the most important when making your decision about the effect of the intervention? Keep in mind that it may have been lack of a feature that influenced your decision. (Circle all that apply.)

   a. Mean Shift
   b. Variability
   c. Trend
1. Please circle a, b, or c based on your interpretation of the data above.
   Based on the data above, I _____ (Circle only one answer)
   
   a. believe that the intervention caused a change in behavior.
   b. believe that the intervention did not change the behavior.
   c. can’t make a decision based on the available information.

2. Based on the data, which feature(s), if any, was/were the most important when making your decision about the effect of the intervention? Keep in mind that it may have been lack of a feature that influenced your decision. (Circle all that apply.)
   
   a. Mean Shift
   b. Variability
   c. Trend
1. Please circle a, b, or c based on your interpretation of the data above. Based on the data above, I _____ (Circle only one answer)
   a. believe that the intervention caused a change in behavior.
   b. believe that the intervention did not change the behavior.
   c. can’t make a decision based on the available information.

2. Based on the data, which feature(s), if any, was/were the most important when making your decision about the effect of the intervention? Keep in mind that it may have been lack of a feature that influenced your decision. (Circle all that apply.)
   a. Mean Shift
   b. Variability
   c. Trend
1. Please circle a, b, or c based on your interpretation of the data above. Based on the data above, I _____ (Circle only one answer)

a. believe that the intervention caused a change in behavior.
b. believe that the intervention did not change the behavior.
c. can’t make a decision based on the available information.

2. Based on the data, which feature(s), if any, was/were the most important when making your decision about the effect of the intervention? Keep in mind that it may have been lack of a feature that influenced your decision. (Circle all that apply.)

a. Mean Shift
b. Variability
c. Trend
1. Please circle a, b, or c based on your interpretation of the data above. Based on the data above, I _____ (Circle only one answer)

   a. believe that the intervention caused a change in behavior.
   b. believe that the intervention did not change the behavior.
   c. can’t make a decision based on the available information.

2. Based on the data, which feature(s), if any, was/were the most important when making your decision about the effect of the intervention? Keep in mind that it may have been lack of a feature that influenced your decision. (Circle all that apply.)

   a. Mean Shift
   b. Variability
   c. Trend
1. Please circle a, b, or c based on your interpretation of the data above. Based on the data above, I _____ (Circle only one answer)
   a. believe that the intervention caused a change in behavior.
   b. believe that the intervention did not change the behavior.
   c. can’t make a decision based on the available information.

2. Based on the data, which feature(s), if any, was/were the most important when making your decision about the effect of the intervention? Keep in mind that it may have been lack of a feature that influenced your decision. (Circle all that apply.)
   a. Mean Shift
   b. Variability
   c. Trend
1. Please circle a, b, or c based on your interpretation of the data above.
Based on the data above, I ____ (Circle only one answer)

a. believe that the intervention caused a change in behavior.
b. believe that the intervention did not change the behavior.
c. can’t make a decision based on the available information.

2. Based on the data, which feature(s), if any, was/were the most important when making your decision about the effect of the intervention? Keep in mind that it may have been lack of a feature that influenced your decision. (Circle all that apply.)

a. Mean Shift
b. Variability
c. Trend
1. Please circle a, b, or c based on your interpretation of the data above. Based on the data above, I _____ (Circle only one answer)

   a. believe that the intervention caused a change in behavior.
   b. believe that the intervention did not change the behavior.
   c. can’t make a decision based on the available information.

2. Based on the data, which feature(s), if any, was/were the most important when making your decision about the effect of the intervention? Keep in mind that it may have been lack of a feature that influenced your decision. (Circle all that apply.)

   a. Mean Shift
   b. Variability
   c. Trend
1. Please circle a, b, or c based on your interpretation of the data above. Based on the data above, I _____ (Circle only one answer)

a. believe that the intervention caused a change in behavior.
b. believe that the intervention did not change the behavior.
c. can’t make a decision based on the available information.

2. Based on the data, which feature(s), if any, was/were the most important when making your decision about the effect of the intervention? Keep in mind that it may have been lack of a feature that influenced your decision. (Circle all that apply.)

a. Mean Shift
b. Variability
c. Trend
1. Please circle a, b, or c based on your interpretation of the data above. Based on the data above, I _____ (Circle only one answer)

   a. believe that the intervention **caused a change in behavior**.
   b. believe that the intervention **did not change the behavior**.
   c. can’t make a decision based on the available information.

2. Based on the data, which feature(s), if any, was/were the most important when making your decision about the effect of the intervention? Keep in mind that it may have been lack of a feature that influenced your decision. (Circle all that apply.)

   a. Mean Shift
   b. Variability
   c. Trend
1. Please circle a, b, or c based on your interpretation of the data above. Based on the data above, I _____ (Circle only one answer)
   a. believe that the intervention caused a change in behavior.
   b. believe that the intervention did not change the behavior.
   c. can’t make a decision based on the available information.

2. Based on the data, which feature(s), if any, was/were the most important when making your decision about the effect of the intervention? Keep in mind that it may have been lack of a feature that influenced your decision. (Circle all that apply.)
   a. Mean Shift
   b. Variability
   c. Trend
1. Please circle a, b, or c based on your interpretation of the data above.
   Based on the data above, I _____ (Circle only one answer)
   a. believe that the intervention caused a change in behavior.
   b. believe that the intervention did not change the behavior.
   c. can’t make a decision based on the available information.

2. Based on the data, which feature(s), if any, was/were the most important when making your decision about the effect of the intervention? Keep in mind that it may have been lack of a feature that influenced your decision. (Circle all that apply.)
   a. Mean Shift
   b. Variability
   c. Trend
1. Please circle a, b, or c based on your interpretation of the data above. Based on the data above, I _____ (Circle only one answer)

   a. believe that the intervention caused a change in behavior.
   b. believe that the intervention did not change the behavior.
   c. can’t make a decision based on the available information.

2. Based on the data, which feature(s), if any, was/were the most important when making your decision about the effect of the intervention? Keep in mind that it may have been lack of a feature that influenced your decision. (Circle all that apply.)

   a. Mean Shift
   b. Variability
   c. Trend
1. Please circle a, b, or c based on your interpretation of the data above.
Based on the data above, I _____ (Circle only one answer)

a. believe that the intervention **caused a change in behavior**.
b. believe that the intervention **did not change the behavior**.
c. can’t make a decision based on the available information.

2. Based on the data, which feature(s), if any, was/were the most important when making your decision about the effect of the intervention? Keep in mind that it may have been lack of a feature that influenced your decision. (Circle all that apply.)

a. Mean Shift
b. Variability
c. Trend
Graph 17

1. Please circle a, b, o’r c based on your interpretation of the data above. Based on the data above, I _____ (Circle only one answer)

   a. believe that the intervention caused a change in behavior.
   b. believe that the intervention did not change the behavior.
   c. can’t make a decision based on the available information.

2. Based on the data, which feature(s), if any, was/were the most important when making your decision about the effect of the intervention? Keep in mind that it may have been lack of a feature that influenced your decision. (Circle all that apply.)

   a. Mean Shift
   b. Variability
   c. Trend
1. Please circle a, b, or c based on your interpretation of the data above. Based on the data above, I _____ (Circle only one answer)
   a. believe that the intervention caused a change in behavior.
   b. believe that the intervention did not change the behavior.
   c. can’t make a decision based on the available information.

2. Based on the data, which feature(s), if any, was/were the most important when making your decision about the effect of the intervention? Keep in mind that it may have been lack of a feature that influenced your decision. (Circle all that apply.)
   a. Mean Shift
   b. Variability
   c. Trend
1. Please circle a, b, or c based on your interpretation of the data above. 
   Based on the data above, I _____ (Circle only one answer)

   a. believe that the intervention caused a change in behavior.
   b. believe that the intervention did not change the behavior.
   c. can’t make a decision based on the available information.

2. Based on the data, which feature(s), if any, was/were the most important when making your decision about the effect of the intervention? Keep in mind that it may have been lack of a feature that influenced your decision. (Circle all that apply.)

   a. Mean Shift
   b. Variability
   c. Trend
1. Please circle a, b, or c based on your interpretation of the data above.
Based on the data above, I _____ (Circle only one answer)
   a. believe that the intervention caused a change in behavior.
   b. believe that the intervention did not change the behavior.
   c. can’t make a decision based on the available information.

2. Based on the data, which feature(s), if any, was/were the most important when making your decision about the effect of the intervention? Keep in mind that it may have been lack of a feature that influenced your decision. (Circle all that apply.)
   a. Mean Shift
   b. Variability
   c. Trend
1. Please circle a, b, or c based on your interpretation of the data above.
Based on the data above, I _____ (Circle only one answer)

a. believe that the intervention **caused a change in behavior**.
b. believe that the intervention **did not change the behavior**.
c. can’t make a decision based on the available information.

2. Based on the data, which feature(s), if any, was/were the most important when making your decision about the effect of the intervention? Keep in mind that it may have been lack of a feature that influenced your decision. (Circle all that apply.)

a. Mean Shift
b. Variability
c. Trend
1. Please circle a, b, or c based on your interpretation of the data above. Based on the data above, I _____ (Circle only one answer)

a. believe that the intervention caused a change in behavior.
b. believe that the intervention did not change the behavior.
c. can’t make a decision based on the available information.

2. Based on the data, which feature(s), if any, was/were the most important when making your decision about the effect of the intervention? Keep in mind that it may have been lack of a feature that influenced your decision. (Circle all that apply.)

a. Mean Shift
b. Variability
c. Trend
1. Please circle a, b, or c based on your interpretation of the data above.
Based on the data above, I _____ (Circle only one answer)

a. believe that the intervention caused a change in behavior.
b. believe that the intervention did not change the behavior.
c. can’t make a decision based on the available information.

2. Based on the data, which feature(s), if any, was/were the most important when making your decision about the effect of the intervention? Keep in mind that it may have been lack of a feature that influenced your decision. (Circle all that apply.)

a. Mean Shift
b. Variability
c. Trend
1. Please circle a, b, or c based on your interpretation of the data above.
   Based on the data above, I _____ (Circle only one answer)

   a. believe that the intervention **caused** a change in behavior.
   b. believe that the intervention **did not change** the behavior.
   c. can’t make a decision based on the available information.

2. Based on the data, which feature(s), if any, was/were the most important when making your decision about the effect of the intervention? Keep in mind that it may have been lack of a feature that influenced your decision. (Circle all that apply.)

   a. **Mean Shift**
   b. **Variability**
   c. **Trend**
1. Please circle a, b, or c based on your interpretation of the data above. 
Based on the data above, I _____. (Circle only one answer)

a. believe that the intervention caused a change in behavior.
b. believe that the intervention did not change the behavior.
c. can’t make a decision based on the available information.

2. Based on the data, which feature(s), if any, was/were the most important when making your decision about the effect of the intervention? Keep in mind that it may have been lack of a feature that influenced your decision. (Circle all that apply.)

a. Mean Shift
b. Variability
c. Trend
1. Please circle a, b, or c based on your interpretation of the data above.  
   Based on the data above, I _____ (Circle only one answer)  
   a. believe that the intervention caused a change in behavior.  
   b. believe that the intervention did not change the behavior.  
   c. can’t make a decision based on the available information.

2. Based on the data, which feature(s), if any, was/were the most important when making your decision about the effect of the intervention? Keep in mind that it may have been lack of a feature that influenced your decision. (Circle all that apply.)  
   a. Mean Shift  
   b. Variability  
   c. Trend
1. Please circle a, b, or c based on your interpretation of the data above.
Based on the data above, I _____ (Circle only one answer)
   a. believe that the intervention caused a change in behavior.
   b. believe that the intervention did not change the behavior.
   c. can’t make a decision based on the available information.

2. Based on the data, which feature(s), if any, was/were the most important when making your decision about the effect of the intervention? Keep in mind that it may have been lack of a feature that influenced your decision. (Circle all that apply.)
   a. Mean Shift
   b. Variability
   c. Trend
1. Please circle a, b, or c based on your interpretation of the data above.
Based on the data above, I _____ (Circle only one answer)

a. believe that the intervention caused a change in behavior.
b. believe that the intervention did not change the behavior.
c. can’t make a decision based on the available information.

2. Based on the data, which feature(s), if any, was/were the most important when making your decision about the effect of the intervention? Keep in mind that it may have been lack of a feature that influenced your decision. (Circle all that apply.)

a. Mean Shift
b. Variability
c. Trend
1. Please circle a, b, or c based on your interpretation of the data above. 
   Based on the data above, I _____ (Circle only one answer)
   a. believe that the intervention caused a change in behavior.
   b. believe that the intervention did not change the behavior.
   c. can’t make a decision based on the available information.

2. Based on the data, which feature(s), if any, was/were the most important when making your decision about the effect of the intervention? Keep in mind that it may have been lack of a feature that influenced your decision. (Circle all that apply.)
   a. Mean Shift
   b. Variability
   c. Trend
Graph 31

1. Please circle a, b, or c based on your interpretation of the data above. Based on the data above, I _____ (Circle only one answer)
   
a. believe that the intervention caused a change in behavior.
b. believe that the intervention did not change the behavior.
c. can’t make a decision based on the available information.

2. Based on the data, which feature(s), if any, was/were the most important when making your decision about the effect of the intervention? Keep in mind that it may have been lack of a feature that influenced your decision. (Circle all that apply.)
   
a. Mean Shift
b. Variability
c. Trend
1. Please circle a, b, or c based on your interpretation of the data above.
Based on the data above, I _____ (Circle only one answer)

a. believe that the intervention caused a change in behavior.
b. believe that the intervention did not change the behavior.
c. can’t make a decision based on the available information.

2. Based on the data, which feature(s), if any, was/were the most important when making your decision about the effect of the intervention? Keep in mind that it may have been lack of a feature that influenced your decision. (Circle all that apply.)

a. Mean Shift
b. Variability
c. Trend
1. Please circle a, b, or c based on your interpretation of the data above. Based on the data above, I _____ (Circle only one answer)

a. believe that the intervention caused a change in behavior.
b. believe that the intervention did not change the behavior.
c. can’t make a decision based on the available information.

2. Based on the data, which feature(s), if any, was/were the most important when making your decision about the effect of the intervention? Keep in mind that it may have been lack of a feature that influenced your decision. (Circle all that apply.)

a. Mean Shift
b. Variability
c. Trend
1. Please circle a, b, or c based on your interpretation of the data above.
   Based on the data above, I ______ (Circle only one answer)
   a. believe that the intervention caused a change in behavior.
   b. believe that the intervention did not change the behavior.
   c. can’t make a decision based on the available information.

2. Based on the data, which feature(s), if any, was/were the most important when making your decision about the effect of the intervention? Keep in mind that it may have been lack of a feature that influenced your decision. (Circle all that apply.)
   a. Mean Shift
   b. Variability
   c. Trend
1. Please circle a, b, or c based on your interpretation of the data above.
Based on the data above, I _____ (Circle only one answer)

a. believe that the intervention caused a change in behavior.
b. believe that the intervention did not change the behavior.
c. can’t make a decision based on the available information.

2. Based on the data, which feature(s), if any, was/were the most important when making your decision about the effect of the intervention? Keep in mind that it may have been lack of a feature that influenced your decision. (Circle all that apply.)

a. Mean Shift
b. Variability
c. Trend
Graph 45

1. Please circle a, b, or c based on your interpretation of the data above.
   Based on the data above, I _____ (Circle only one answer)
   a. believe that the intervention caused a change in behavior.
   b. believe that the intervention did not change the behavior.
   c. can’t make a decision based on the available information.

2. Based on the data, which feature(s), if any, was/were the most important when making your decision about the effect of the intervention? Keep in mind that it may have been lack of a feature that influenced your decision. (Circle all that apply.)
   a. Mean Shift
   b. Variability
   c. Trend
Graph 30

1. Please circle a, b, or c based on your interpretation of the data above. Based on the data above, I _____ (Circle only one answer)
   a. believe that the intervention **caused a change in behavior**.
   b. believe that the intervention **did not change the behavior**.
   c. can’t make a decision based on the available information.

2. Based on the data, which feature(s), if any, was/were the most important when making your decision about the effect of the intervention? Keep in mind that it may have been lack of a feature that influenced your decision. (Circle all that apply.)
   a. Mean Shift
   b. Variability
   c. Trend
1. Please circle a, b, or c based on your interpretation of the data above. 
Based on the data above, I _____ (Circle only one answer)

a. believe that the intervention caused a change in behavior.
b. believe that the intervention did not change the behavior.
c. can’t make a decision based on the available information.

2. Based on the data, which feature(s), if any, was/were the most important when making your decision about the effect of the intervention? Keep in mind that it may have been lack of a feature that influenced your decision. (Circle all that apply.)

a. Mean Shift
b. Variability
c. Trend
1. Please circle a, b, or c based on your interpretation of the data above.
   Based on the data above, I _____ (Circle only one answer)
   
   a. believe that the intervention caused a change in behavior.
   b. believe that the intervention did not change the behavior.
   c. can’t make a decision based on the available information.

2. Based on the data, which feature(s), if any, was/were the most important when making your decision about the effect of the intervention? Keep in mind that it may have been lack of a feature that influenced your decision. (Circle all that apply.)

   a. Mean Shift
   b. Variability
   c. Trend
1. Please circle a, b, or c based on your interpretation of the data above. 
Based on the data above, I _____ (Circle only one answer)

a. believe that the intervention caused a change in behavior. 
b. believe that the intervention did not change the behavior. 
c. can’t make a decision based on the available information.

2. Based on the data, which feature(s), if any, was/were the most important when making your decision about the effect of the intervention? Keep in mind that it may have been lack of a feature that influenced your decision. (Circle all that apply.)

a. Mean Shift 
b. Variability 
c. Trend
1. **Please circle a, b, or c based on your interpretation of the data above.**
   Based on the data above, I _____ (Circle only one answer)
   
   a. believe that the intervention *caused a change in behavior.*
   b. believe that the intervention *did not change the behavior.*
   c. can’t make a decision based on the available information.

2. **Based on the data, which feature(s), if any, was/were the most important when making your decision about the effect of the intervention? Keep in mind that it may have been lack of a feature that influenced your decision. (Circle all that apply.)

   a. Mean Shift
   b. Variability
   c. Trend
1. Please circle a, b, or c based on your interpretation of the data above. 
   Based on the data above, I _____ (Circle only one answer)
   
a. believe that the intervention **caused a change in behavior**.
b. believe that the intervention **did not change the behavior**.
c. can’t make a decision based on the available information.

2. Based on the data, which feature(s), if any, was/were the most important when making your decision about the effect of the intervention? Keep in mind that it may have been lack of a feature that influenced your decision. (Circle all that apply.)

   a. **Mean Shift**
b. **Variability**
c. **Trend**
1. Please circle a, b, or c based on your interpretation of the data above. Based on the data above, I _____ (Circle only one answer)

   a. believe that the intervention caused a change in behavior.
   b. believe that the intervention did not change the behavior.
   c. can’t make a decision based on the available information.

2. Based on the data, which feature(s), if any, was/were the most important when making your decision about the effect of the intervention? Keep in mind that it may have been lack of a feature that influenced your decision. (Circle all that apply.)

   a. Mean Shift
   b. Variability
   c. Trend
1. Please circle a, b, or c based on your interpretation of the data above. Based on the data above, I _____ (Circle only one answer)

a. believe that the intervention **caused a change in behavior**.
b. believe that the intervention **did not change the behavior**.
c. can’t make a decision based on the available information.

2. Based on the data, which feature(s), if any, was/were the most important when making your decision about the effect of the intervention? Keep in mind that it may have been lack of a feature that influenced your decision. (Circle all that apply.)

a. Mean Shift
b. Variability
c. Trend
<table>
<thead>
<tr>
<th>Packet</th>
<th>Graph A</th>
<th>Graph A</th>
<th>Graph B</th>
<th>Graph C</th>
<th>Graphs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Packet 1</td>
<td>Graph A</td>
<td>Graph A</td>
<td>Graph B</td>
<td>Graph C</td>
<td>19 25 4 12</td>
</tr>
<tr>
<td>Packet 2</td>
<td>Graph A</td>
<td>Graph A</td>
<td>Graph B</td>
<td>Graph C</td>
<td>3 28 15 6</td>
</tr>
<tr>
<td>Packet 3</td>
<td>Graph A</td>
<td>Graph A</td>
<td>Graph B</td>
<td>Graph C</td>
<td>9 29 16 8</td>
</tr>
<tr>
<td>Packet 4</td>
<td>Graph A</td>
<td>Graph A</td>
<td>Graph B</td>
<td>Graph C</td>
<td>1 34 44 5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Packet</th>
<th>Graph A</th>
<th>Graph A</th>
<th>Graph B</th>
<th>Graph C</th>
<th>Graphs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Packet 5</td>
<td>Graph A</td>
<td>Graph A</td>
<td>Graph B</td>
<td>Graph C</td>
<td>14 10 24 27</td>
</tr>
<tr>
<td>Packet 6</td>
<td>Graph A</td>
<td>Graph A</td>
<td>Graph B</td>
<td>Graph C</td>
<td>17 35 38 26</td>
</tr>
<tr>
<td>*Packet 7</td>
<td>Graph B</td>
<td>Graph A</td>
<td>Graph B</td>
<td>Graph C</td>
<td>32 40 39 33</td>
</tr>
<tr>
<td>Packet 8</td>
<td>Graph A</td>
<td>Graph A</td>
<td>Graph B</td>
<td>Graph C</td>
<td>1 34 44 5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Packet</th>
<th>Graph A</th>
<th>Graph A</th>
<th>Graph B</th>
<th>Graph C</th>
<th>Graphs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Packet 9</td>
<td>Graph A</td>
<td>Graph A</td>
<td>Graph B</td>
<td>Graph C</td>
<td>20 51 46 42</td>
</tr>
<tr>
<td>Packet 10</td>
<td>Graph A</td>
<td>Graph B</td>
<td>Graph B</td>
<td>Graph C</td>
<td>18 31 50 33</td>
</tr>
<tr>
<td>*Packet 11</td>
<td>Graph B</td>
<td>Graph C</td>
<td>Graph B</td>
<td>Graph C</td>
<td>11 37 23 45</td>
</tr>
<tr>
<td>Packet 12</td>
<td>Graph A</td>
<td>Graph A</td>
<td>Graph B</td>
<td>Graph C</td>
<td>1 34 44 5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Packet</th>
<th>Graph A</th>
<th>Graph A</th>
<th>Graph B</th>
<th>Graph C</th>
<th>Graphs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Packet 13</td>
<td>Graph A</td>
<td>Graph A</td>
<td>Graph B</td>
<td>Graph C</td>
<td>30 41 22 47</td>
</tr>
<tr>
<td>Packet 14</td>
<td>Graph A</td>
<td>Graph A</td>
<td>Graph B</td>
<td>Graph C</td>
<td>7 48 49 43</td>
</tr>
</tbody>
</table>

*Indicates the two packets that were added to the original set of 12 packets.
CDC INSTRUCTIONS

GENERAL INSTRUCTIONS

I would like you to look at a series of behavioral graphs and make some decisions about the change in behavior caused by the interventions. All graphs have been taken from research articles published in the *Journal of Applied Behavior Analysis*. Some of the graphs may have been altered slightly to conform to an ABA design and all the information about the subject, setting, specific behavior and intervention have been removed.

All of the graphs have been labeled with rate per minute as the dependent variable on the ordinate (y-axis) and with sessions as the measure of time on the abscissa (x-axis). Even though the labels may be different than the original graph, every attempt was made to keep the same scale. All of the graphs will be in ABA format that is, a baseline phase followed by an intervention phase followed by a second baseline phase.

I would like you to decide whether or not there was any change in trend, variability and/or level of the rate following the intervention that can be attributed to the intervention. That is, would you say that the intervention caused a behavior change based on the data available to you? If you determined that there was an increase or decrease in rate, I would like you to indicate that there was a change in behavior caused by the intervention. Please indicate your decision by circling an answer to Question 1 below each graph.

I would also like you to indicate which feature(s) of the data (or lack thereof) influenced your decision about the effect of the intervention. For Question 2, please circle as many of the answers that apply.

CDC INSTRUCTIONS

For some of the graphs, you will be provided with a form in which two criterion (CDC) lines will be superimposed on the treatment data to help you make your decision. When the criterion lines are provided, the original ABA graph will be divided into two graphs. The first graph will have the first baseline phase (A) and the treatment phase (B). The CDC criterion lines will be superimposed on the B phase. The second graph will consist of the second baseline phase (A) and the treatment phase (B). The CDC lines will be superimposed on the B phase. Each time that you use the CDC form, please be sure to fill out each step on the form (i.e., indicate your answer in writing) and use the answer from Step 5 on the form to answer Question 1 underneath the original ABA graph. (See attached sheet for details)

Each CDC form will be presented after the graph with which it is to be used and will be paper clipped to that graph. Additionally, a number indicating the graph with which it is to be used will be written in the bottom right hand corner of the CDC form.

When you reach the first graph with the CDC lines, there will be a note for you to signal the experimenter. At this time she will give you a short training on how to use the CDC form to arrive at your answer for Question 1. The training will include three graphs and the experimenter will go through each step on the CDC form for each graph. At the end of the training, you will be instructed to continue using the CDC form on your own.

NOTE:
During the session, I will answer any questions that you have about what you are supposed to do. However, I cannot answer any questions about visual analysis or within-subject research in general or about the graphed data or the contextual information provided for each graph.
ACTIVE JOB AID INSTRUCTIONS

GENERAL INSTRUCTIONS

I would like you to look at a series of behavioral graphs and make some decisions about the change in behavior caused by the interventions. All graphs have been taken from research articles published in the *Journal of Applied Behavior Analysis*. Some of the graphs may have been altered slightly to conform to an ABA design and all the information about the subject, setting, specific behavior and intervention have been removed.

All of the graphs have been labeled with rate per minute as the dependent variable on the ordinate (y-axis) and with sessions as the measure of time on the abscissa (x-axis). Even though the labels may be different than the original graph, every attempt was made to keep the same scale. All of the graphs will be in ABA format that is, a baseline phase followed by an intervention phase followed by a second baseline phase.

I would like you to decide whether or not there was any change in trend, variability and/or level of the rate following the intervention that can be attributed to the intervention. That is, would you say that the intervention caused a behavior change based on the data available to you? If you determined that there was an increase or decrease in rate, I would like you to indicate that there was a change in behavior caused by the intervention. Please indicate your decision by circling an answer to Question 1 below each graph.

I would also like you to indicate which feature(s) of the data (or lack thereof) influenced your decision about the effect of the intervention. For Question 2, please circle as many of the answers that apply.

JOB AID INSTRUCTIONS

For some of the graphs, you will be provided with a job aid to help you make your decision. For each job aid that you use, please be sure to fill out each step on the job aid (i.e., indicate your answer in writing) and use the answer from Step 10 on the job aid to answer Question 1 underneath the graph. (See attached sheet for details)

Each job aid will be presented before the graph with which it is to be used and will be paper clipped to that graph. Additionally, a number indicating the graph with which it is to be used will be written in the upper right hand corner of the job aid.

When you reach the first graph with a job aid, there will be a note for you to signal the experimenter. At this time she will give you a short training on how to use the job aid to arrive at your answer for Question 1. The training will include three graphs and the experimenter will go through each step on the job aid for each graph. At the end of the training, you will be instructed to continue using the job aid on your own.

NOTE:
During the session I will answer any questions that you have about what you are supposed to do. However, I cannot answer any questions about visual analysis or within-subject research in general or about the graphed data or the contextual information provided for each graph.
CONTROL INSTRUCTIONS

GENERAL INSTRUCTIONS
I would like you to look at a series of behavioral graphs and make some decisions about the change in behavior caused by the interventions. All graphs have been taken from research articles published in the Journal of Applied Behavior Analysis. Some of the graphs may have been altered slightly to conform to an ABA design and all the identifying information about the subject, setting, specific behavior and intervention have been removed.

All of the graphs have been labeled with rate per minute as the dependent variable along ordinate (y-axis) and with observation sessions across the abscissa (x-axis). Even though the labels may be different than the original graph, every attempt was made to keep the same scale. All of the graphs will be in ABA format that is, a baseline phase followed by an intervention phase followed by a second baseline phase.

I would like you to decide whether or not there was any change in trend, variability and/or level of the rate following the intervention that can be attributed to the intervention. That is, would you say that the intervention caused a behavior change based on the data available to you? If you determined that there was an increase or decrease in rate, I would like you to indicate that there was a change in behavior caused by the intervention. Please indicate your decision by circling an answer to Question 1 below each graph.

Additionally, I would like you to indicate which feature(s) of the data (or lack thereof) influenced your decision about the effect of the intervention. For Question 2, please circle as many of the answers that apply.

You may take as long as you need for each graph.

NOTE:
During the session I will answer any questions that you have about what you are supposed to do. However, I cannot answer any questions about visual analysis or within-subject research in general or about the graphed data.
**Step 1**
Count and write down the number of treatment sessions for the 1st graph  
Count and write down the number of treatment sessions for the 2nd graph

**Step 2**
*For 1st graph data:* Count the number of data points in the treatment phase that are above both of the dashed criterion lines  
*For 2nd graph data:* Count the number of data points in the treatment phase that are below both of the dashed criterion lines

**Step 3**
On the table provided, look up the number of data points that need to be above (or below) both criterion lines to conclude that the intervention caused a change in behavior

**Graph 1**  
**Graph 2**
(Use numbers from Step 1 under Length, look up Needed next to it)

**Step 4**
Is number for Graph 1 in Step 2 \(\geq\) the number for Graph 1 in Step 3? *Yes*  
Is number for Graph 2 in Step 2 \(\geq\) the number for Graph 2 in Step 3? *Yes*

**Step 5**
What conclusion(s) would you draw based on the data?

*Choice A (intervention caused a change in behavior)*
*IF:*  
“Yes” to Graph 1 in Step 4 AND “Yes” to Graph 2 in Step 4

*Choice B (intervention did not cause a change in behavior)*
*IF:*  
“No” to Graph 1 in Step 4 AND “No” to Graph 2 in Step 4

*Choice C (Can’t make a decision based on the available information)*
*IF:*  
“Yes” to Graph 1 in Step 4 AND “No” to Graph 2 in Step 4

OR

“No” to Graph 1 in Step 4 AND “Yes” to Graph 2 in Step 4
### CDC TABLE AND INSTRUCTIONS

<table>
<thead>
<tr>
<th>Length</th>
<th>Needed</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>11</td>
<td>9</td>
</tr>
<tr>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td>13</td>
<td>10</td>
</tr>
<tr>
<td>14</td>
<td>11</td>
</tr>
<tr>
<td>15</td>
<td>12</td>
</tr>
<tr>
<td>16</td>
<td>12</td>
</tr>
<tr>
<td>17</td>
<td>12</td>
</tr>
<tr>
<td>18</td>
<td>13</td>
</tr>
<tr>
<td>19</td>
<td>13</td>
</tr>
<tr>
<td>20</td>
<td>14</td>
</tr>
</tbody>
</table>

This table allows you to look up the number of data points in the treatment phase and the corresponding number of data points that must be above or below both criterion lines in order to conclude that there is a reliable treatment effect using the CDC method.

Look up the answer for Step 1 under the Length column above. This is the number of points in the treatment (intervention/B) phase. Directly across from this number is the number of points that are needed to be above or below both criterion lines. This is the answer to Step 3.
VISUAL ANALYSIS JOB AID

**Key Terms**

**Mean Shift:** A change in the average rate of behavior across two phases. Mean shift may be observed with or without trend and variability present in the data. (See especially Steps 8 and 9 below)

**Trend:** A relatively consistent pattern of data in a single direction (e.g., up, down). The presence of a trend makes drawing conclusions about the effects of the intervention difficult, unless the trend is reversed. (See especially Steps 5 and 6 below)

**Variability:** Fluctuation in behavior over time. Excessive variability in the data during baseline or other phases can interfere with drawing conclusions about the treatment. As a general rule, the greater the variability in the data, the more difficult it is to draw conclusions about the effects of the intervention. (See especially Steps 2 and 5 below)

___Step 1
Draw the mean (average) rate of behavior in baseline, in intervention and in return to baseline.

___Step 2
Determine if there is a mean shift across the first baseline and intervention.  
(See **Key Terms** for definition)

___Step 3
Determine if there is a mean shift across intervention and the second baseline.

___Step 4
Draw the trend of behavior in baseline, in intervention and in return to baseline.  
(See **Key Terms** for definition)

___Step 5
Determine if the trend across the first baseline is in the same direction as intervention.

___Step 6
Determine if the trend across intervention is the same as the second baseline.

___Step 7
Draw the variability of behavior in baseline, in intervention and in return to baseline by drawing the range of each phase. (See **Key Terms** for definition)

___Step 8
Determine if there is an overlap of more than 50% of the data between the range in the first baseline and intervention.

___Step 9
Determine if there is an overlap of more than 50% of the data between the range in intervention and the second baseline.

___Step 10
What conclusion(s) would you draw based on the data?

---

___Choice A (intervention caused a change in behavior)
**IF:**  
“Yes” to Steps 2 and 3  AND  “No” to Steps 8 and 9

---

___Choice B (intervention did not cause a change in behavior)
**IF:**  
“Yes” Step 2  AND  “Yes” to Step 8 and 9  
**OR**  
“No” to Step 2

---

___Choice C (Can’t make a decision based on the available information)
**IF:**  
“Yes” to Step 2  AND  “No” to Step 8 and “Yes” to Step 9

---

DO YOU AGREE WITH THIS CONCLUSION?

_____Yes  
_____No
APPENDIX F

INFORMED CONSENT FORM
I freely and voluntarily and without element of force or coercion, consent to be a participant in the research project entitled “Improving the Reliability and Validity of Visual Inspection of Data by Behavior Analysts: An Empirical Comparison of Two Training Methods to Improve Visual Inspection and Interpretation, the Active Job Aid and the Conservative Dual-Criteria (CDC).”

This research is being conducted Marilin Colón, MS who is a graduate student in psychology at Florida State University. I understand the purpose of his research project is to better understand some of the variables that influence the interpretation of single-subject behavioral graphs. Specifically, this study will examine the differences between analyses using job aids, analyses with criterion lines and analyses without any aids.

I understand I will be asked to participate in two sessions, each lasting approximately 1-hour. The total time commitment would be about 2-hours. The research assistant will answer my questions or I will be referred to a knowledgeable source.

I understand there is a possibility of a minimal level of risk involved if I agree to participate in this study. I might experience anxiety when analyzing the graphs or in deciding if there was a behavior change indicated by the data. The research assistant will be available to talk with me about any emotional discomfort I may experience while participating. I am also able to stop my participation at any time I wish.

I understand there are benefits for participating in this research project. First, my own awareness about my decision-making relevant to graphed behavioral data might be increased. Also, I will be providing the psychological community with valuable insight into the variables influencing the analysis and interpretation of behavioral graphs. This knowledge can assist professionals in developing guidelines for such visual analyses.

I understand my participation is voluntary and I may stop participation at anytime. All of my answers will be kept confidential to the extent allowed by law and identified by a subject code number. My name will not appear on any of the results.

I understand that this consent may be withdrawn at any time without prejudice, penalty or loss of benefits to which I am otherwise entitled. I have been given the right to ask and have answered any inquiry concerning the study. Questions, if any, have been answered to my satisfaction.

I understand that I may contact Marilin Colón, Florida State University, Department of Psychology, (321) 432-8378 or Dr. Jon S. Bailey, Florida State University, Department of Psychology, (850) 644-6443, for answers to questions about the research, my rights as a research subject, and potential research-related injuries. Results will be sent to me upon my request. I may also contact the Florida State University Human Subjects Institutional Review Board at 2035 E. Paul Dirac Drive, Box 15100 Sliger Bldg., Innovation Park, Tallahassee, FL 32310 or by phone at (850) 644-8633.

I have read and understand this consent form.

___________________________________________________ __________________
(Subject) (Date)
APPENDIX G

POST-STUDY PARTICIPANT SURVEY
Post-Study Survey

1. Do you currently work as a behavior analyst?  
    Yes  No

2. For how many years have you worked as a behavior analyst?  
   a. 0 years  
   b. 1-2 years  
   c. 3-5 years  
   d. 5-7 years  
   e. 7-10 years

3. Do you view graphed behavioral data as part of your job?  
   Yes  No

4. If you answered “yes” to question 3, how often do you view graphed behavioral data?  
   a. daily  
   b. weekly  
   c. monthly  
   d. seldom  
   e. never

5. Are you a Board Certified Behavior Analyst?  
   Yes  No

6. If you answered “yes” to question 5, for how many years have you been certified?  
   a. Less than 1 year  
   b. 1 year  
   c. 2 years  
   d. 3 years  
   e. More than 3 years

7. What is the highest diploma/degree that you have obtained?  
   a. High School Diploma  
   b. Bachelor  
   c. Masters  
   d. Doctorate

8. When did you receive this degree? __________ (year)

9. From where did you earn this degree? _____________________

10. If applicable, do you feel that the job aids provided in this study helped you to better analyze the graphs?  
    Yes  No

11. If applicable, do you feel that the CDC lines provided in this study helped you to better analyze the graphs?  
    Yes  No
DEBRIEFING STATEMENT

I understand that the purpose of the current research is to better understand how individuals view graphs and make decisions about their relevance. I further understand that my name will not be directly related to the data collected during my session and that if I would like a copy of the results from this study as well as a written discussion of its implications, I may do so any time after the final data is collected and analyzed.

If I have any questions regarding this research, I have been informed to contact Marilin Colón directly at (321) 432-8378.

I have read and understand this debriefing form.

Subject’s printed name

(Subject) (Date)

(Witness) (Date)
APPENDIX H

HSIRB APPROVAL LETTER
HSIRB APPROVAL LETTER

Office of the Vice President For Research
Human Subjects Committee
Tallahassee, Florida 32306-2763
(850) 644-8673  FAX (850) 644-4392

APPROVAL MEMORANDUM

Date: 11/3/2005

To:  
Marlin Colon  
355 Lofts Drive  
Melbourne, FL 32904

Dept.: PSYCHOLOGY DEPARTMENT

From: Thomas L. Jacobson, Chair

Re: Use of Human Subjects in Research

The forms that you submitted to this office in regard to the use of human subjects in the proposal referenced above have been reviewed by the Secretary, the Chair, and two members of the Human Subjects Committee. Your project is determined to be Exempt per 45 CFR § 46.101(b) 2 and has been approved by an accelerated review process.

The Human Subjects Committee has not evaluated your proposal for scientific merit, except to weigh the risk to the human participants and the aspects of the proposal related to potential risk and benefit. This approval does not replace any departmental or other approvals, which may be required.

If the project has not been completed by 11/1/2006 you must request renewed approval for continuation of the project.

You are advised that any change in protocol in this project must be approved by resubmission of the project to the Committee for approval. Also, the principle investigator must promptly report, in writing, any unexpected problems causing risks to research subjects or others.

By copy of this memorandum, the chairman of your department and/or your major professor is reminded that he/she is responsible for being informed concerning research projects involving human subjects in the department, and should review protocols of such investigations as often as needed to ensure that the project is being conducted in compliance with our institution and with DHHS regulations.

This institution has an Assurance on file with the Office for Protection from Research Risks. The Assurance Number is IRB00000469.

Cc: Jon S. Bailey  
HSC No. 2005.759
REFERENCES


Appleton-Century Crofts.


BIOGRAPHICAL SKETCH

Marilin Colón was born in Guadalajara, Mexico on December 15, 1970. At the age of two, she moved to Puerto Rico, where she lived until she moved to Baltimore, Maryland in 1981. Marilin attended Roland Park Country School from seventh grade until she graduated in 1988. Upon graduation, she entered the Johns Hopkins University where she earned a BA in Behavioral Biology in May 1992. In the summer of 1992, Marilin moved to Tallahassee, Florida. She started working for Dr. Maxin Reiss and became interested in behavior analysis. In 1994, she enrolled in the doctoral program at Florida State University under the supervision of Dr. Jon S. Bailey. She received her MS in psychology in 1998. On March 12, 2004, she married fellow behavior analyst, Dr. Matthew Normand. Marilin currently serves as Central Florida Area Coordinator for Behavior Management Consultants, Inc. and resides in Melbourne, Florida with her husband and two dogs, Maximus and Samantha.