An Evaluation of a Computer-Based Training on the Visual Analysis of Single-Subject Data

Katie Snyder
Utah State University

Follow this and additional works at: https://digitalcommons.usu.edu/etd

Part of the Psychology Commons, Special Education and Teaching Commons, and the Teacher Education and Professional Development Commons

Recommended Citation
AN EVALUATION OF A COMPUTER-BASED TRAINING ON THE VISUAL
ANALYSIS OF SINGLE-SUBJECT DATA

by

Katie Snyder

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

DOCTOR OF PHILOSOPHY

in

Disability Disciplines

Approved:

Timothy A. Slocum
Major Professor

Ron Gillam
Committee Member

Sarah Bloom
Committee Member

Bob Morgan
Committee Member

Michael Twohig
Committee Member

Mark R. McLellan
Vice President for Research and
Dean of the School of Graduate Studies

UTAH STATE UNIVERSITY
Logan, UT
2013
ABSTRACT

An Evaluation of a Computer-Based Training on the Visual Analysis of Single-Subject Data

by

Katie Snyder, Doctor of Philosophy

Utah State University, 2013

Major Professor: Dr. Timothy A. Slocum
Department: Special Education and Rehabilitation

Visual analysis is the primary method of analyzing data in single-subject methodology, which is the predominant research method used in the fields of applied behavior analysis and special education. Previous research on the reliability of visual analysis suggests that judges often disagree about what constitutes an intervention effect. Considering that visual analysis involves a complex set of discriminations and sometimes produces disagreement among experts, it is particularly important to examine methods to train individuals to visually analyze data. Only a handful of studies have investigated the effectiveness of trainings on visual analysis. Most have relied on mechanical methods and/or evaluated the effectiveness of the procedure using a narrow set of graphs (e.g., graphs without slope). The purpose of this study was to evaluate two training methods using graphs with various combinations of effect types. The computer-based training, which includes high rates of practice with feedback, was compared to a lecture condition and a control condition. Results indicate that both training methods (i.e., the computer-based training and the lecture) were more effective than a control condition, but were not
substantially different from one another. We discuss the implications of these results for training individuals in visual analysis as well as directions for future research.
PUBLIC ABSTRACT

An Evaluation of a Computer-Based Training on the Visual Analysis of Single-Subject Data

by

Katie Snyder, Doctor of Philosophy
Utah State University, 2013

Federal education policies, such as No Child Left Behind and the Individuals with Disabilities Education Improvement Act, mandate the use of scientifically-proven or research-based curricula and interventions. Presumably, interventions that have a large amount of scientific evidence documenting their success are more likely to be effective when implemented with students in school settings.

In special education, single-subject research is the predominant methodology used to evaluate the effectiveness of interventions. In single-subject research, a target behavior is measured under baseline conditions (i.e., before the intervention of interest is implemented) and intervention conditions. The data for each condition are graphed, and analyzed visually to evaluate whether the behavior appears to have changed as a result of the intervention. The conditions are replicated, and interventions that produce reliable changes in behavior are considered effective.

Although visual analysis seems straightforward, previous research suggests that experts often disagree about what constitutes an intervention effect using visual analysis. This disagreement has important implications for the identification of research-based curricula and interventions. If two experts disagree about whether a single-subject graph depicts an effective intervention, our confidence in any conclusions we might reach about
the intervention using that graph is diminished. As a result, that study cannot contribute meaningfully to the body of literature supporting the use of the intervention.

Given the importance identifying research-based interventions and the complexity of visual analysis, it is particularly important to examine methods to train individuals to visually analyze data. Few studies to date have evaluated approaches to training visual analysis. In the present study, we developed an assessment tool to measure participants’ visual analysis skills. This assessment was used to compare the effects of a lecture to the effects of an interactive, computer-based training on visual analysis. We also included a no-treatment condition in which participants received no instruction on visual analysis. One hundred and twenty three undergraduate participants were randomly assigned to one of the three groups. Results suggested that the lecture and computer-based training both significantly improved participants’ visual analysis skills compared to no-treatment condition. These results suggest that structured approaches to teaching visual analysis improve novices’ accuracy. More systematic training approaches, such as those investigated here, could also increase agreement among experts and result in more valid identification of research-based interventions.
ACKNOWLEDGMENTS

The Utah Department of Education provided the funding that made this research possible.

Many thanks to Dr. Timothy A. Slocum, who has been extremely generous with his time and energy on this project and throughout my time at Utah State University. I especially appreciate your willingness to spend countless hours examining and discussing a seemingly infinite number of graphs. Your patience, humor, enthusiasm, and continued efforts to push me beyond my comfort zone have made this experience enjoyable and rewarding.

I am grateful to my committee members, Drs. Sarah Bloom, Ron Gillam, Bob Morgan, and Mike Twohig, for their input on this study and their contributions to my growth throughout the program.

Special thanks to Casey Clay, Shanun Kunnavatana, and Dr. Andrew Samaha for lending their eyes and expertise to our efforts to identify valid graphs, and to Dr. Benjamin Lignugaris/Kraft for helping me field-test the graphs with undergraduates.

I am thankful for the assistance of the instructors and students of SPED 4000 and 5530, whose support and participation made this study possible.

Brady Wolfe has been a constant source of strength and support over the past four years and beyond. Thank you for coming on this adventure with me, and for eagerly anticipating the next one.

Katie Snyder
## CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>iii</td>
</tr>
<tr>
<td>PUBLIC ABSTRACT</td>
<td>v</td>
</tr>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>vii</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>x</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>xi</td>
</tr>
<tr>
<td>CHAPTER</td>
<td></td>
</tr>
<tr>
<td>I. INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>II. LITERATURE REVIEW</td>
<td>4</td>
</tr>
<tr>
<td>Agreement Between Raters</td>
<td>4</td>
</tr>
<tr>
<td>Statistical Analysis and Visual Analysis</td>
<td>12</td>
</tr>
<tr>
<td>Agreement Between Raters with Visual Aids</td>
<td>19</td>
</tr>
<tr>
<td>Training Visual Analysis</td>
<td>22</td>
</tr>
<tr>
<td>Purpose Statement and Research Questions</td>
<td>38</td>
</tr>
<tr>
<td>III. METHODS</td>
<td>41</td>
</tr>
<tr>
<td>Participants</td>
<td>41</td>
</tr>
<tr>
<td>Setting</td>
<td>41</td>
</tr>
<tr>
<td>Materials</td>
<td>42</td>
</tr>
<tr>
<td>Dependent Measures</td>
<td>51</td>
</tr>
<tr>
<td>Interobserver Agreement</td>
<td>56</td>
</tr>
<tr>
<td>Experimental Design</td>
<td>56</td>
</tr>
<tr>
<td>Procedures</td>
<td>56</td>
</tr>
<tr>
<td>Treatment Integrity</td>
<td>59</td>
</tr>
<tr>
<td>IV. RESULTS</td>
<td>60</td>
</tr>
<tr>
<td>Missing Scores</td>
<td>60</td>
</tr>
<tr>
<td>Relative Effectiveness of the Interventions</td>
<td>61</td>
</tr>
<tr>
<td>Effectiveness and Efficiency of the Computer-Based Training</td>
<td>85</td>
</tr>
<tr>
<td>Treatment Acceptability</td>
<td>90</td>
</tr>
<tr>
<td>V. DISCUSSION</td>
<td>93</td>
</tr>
</tbody>
</table>
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Description of Analytical Approaches Related to the Split-Middle Line</td>
<td>15</td>
</tr>
<tr>
<td>2</td>
<td>Approaches to Visual Analysis Evaluated in Training Studies</td>
<td>24</td>
</tr>
<tr>
<td>3</td>
<td>Descriptive Statistics for Demographic Variables</td>
<td>42</td>
</tr>
<tr>
<td>4</td>
<td>ANOVAs of Demographic Variables</td>
<td>42</td>
</tr>
<tr>
<td>5</td>
<td>Mean Scores for the Overall Assessment (Standard Deviations in Parentheses)</td>
<td>62</td>
</tr>
<tr>
<td>6</td>
<td>ANOVAs/ANCOVAs of the Overall Assessment</td>
<td>62</td>
</tr>
<tr>
<td>7</td>
<td>Mean Scores for Acquisition and Reduction Graphs (Standard Deviations in</td>
<td>69</td>
</tr>
<tr>
<td></td>
<td>Parentheses)</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Mean Scores for Slope and Level Questions (Standard Deviations in</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>Parentheses)</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>ANOVAs of the Acquisition Graphs</td>
<td>70</td>
</tr>
<tr>
<td>10</td>
<td>ANOVAs of the Reduction Graphs</td>
<td>71</td>
</tr>
<tr>
<td>11</td>
<td>ANOVAs of the Level Questions</td>
<td>71</td>
</tr>
<tr>
<td>12</td>
<td>ANOVAs of Slope Questions</td>
<td>71</td>
</tr>
<tr>
<td>13</td>
<td>Proportion of Participants Meeting 80% Accuracy Criterion on Posttest</td>
<td>78</td>
</tr>
<tr>
<td>14</td>
<td>Number of Items in Each Category of Change from Pretest to Posttest</td>
<td>79</td>
</tr>
<tr>
<td>15</td>
<td>Descriptive Statistics of Secondary Dependent Variables</td>
<td>85</td>
</tr>
<tr>
<td>16</td>
<td>Effectiveness and Efficiency Measures of Modules and Instructional Segments</td>
<td>88</td>
</tr>
<tr>
<td>17</td>
<td>Descriptive Statistics and ANOVAs of Treatment Acceptability Ratings</td>
<td>91</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Process for generating and evaluating graphs for inclusion in assessment</td>
<td>45</td>
</tr>
<tr>
<td>2</td>
<td>The eight types of graphs included in the assessment</td>
<td>46</td>
</tr>
<tr>
<td>3</td>
<td>Mean, minimum, and maximum of percent of undergraduates responding correctly about the level and slope changes of the graphs in the assessment packets.</td>
<td>47</td>
</tr>
<tr>
<td>4</td>
<td>Descriptive statistics for the parameters of the assessment graphs.</td>
<td>48</td>
</tr>
<tr>
<td>5</td>
<td>Instructional design for the Level Change module.</td>
<td>53</td>
</tr>
<tr>
<td>6</td>
<td>Instructional design for the Slope Change module.</td>
<td>54</td>
</tr>
<tr>
<td>7</td>
<td>Instructional design for the Slope and Level Change module.</td>
<td>55</td>
</tr>
<tr>
<td>8</td>
<td>Distributions of overall pretest scores.</td>
<td>63</td>
</tr>
<tr>
<td>9</td>
<td>Means and 95% confidence intervals of posttest scores as a function of pretest score and group.</td>
<td>65</td>
</tr>
<tr>
<td>10</td>
<td>Distributions of overall posttest scores.</td>
<td>65</td>
</tr>
<tr>
<td>11</td>
<td>Means and 95% confidence intervals of change scores as a function of pretest score and group.</td>
<td>67</td>
</tr>
<tr>
<td>12</td>
<td>Distributions of overall change scores.</td>
<td>67</td>
</tr>
<tr>
<td>13</td>
<td>Means and 95% confidence intervals of change scores on acquisition graphs as a function of pretest score and group.</td>
<td>73</td>
</tr>
<tr>
<td>14</td>
<td>Means and confidence intervals of change scores on reduction graphs as a function of pretest score and group.</td>
<td>74</td>
</tr>
<tr>
<td>15</td>
<td>Means and confidence intervals of change scores on level questions as a function of pretest score and group.</td>
<td>76</td>
</tr>
<tr>
<td>16</td>
<td>Means and confidence intervals of change scores on slope questions as a function of pretest score and group.</td>
<td>77</td>
</tr>
<tr>
<td>17</td>
<td>Distribution of proportional changes on acquisition and reduction graphs.</td>
<td>81</td>
</tr>
<tr>
<td>Page</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Distribution of proportional changes on slope and level questions. ........................</td>
<td>82</td>
</tr>
<tr>
<td>19</td>
<td>Three graphs on which both intervention groups improved by at least 25% ..................</td>
<td>83</td>
</tr>
<tr>
<td>20</td>
<td>The acquisition graph on which both intervention groups worsened by 25% ...............</td>
<td>84</td>
</tr>
<tr>
<td>21</td>
<td>Distributions of change scores for participants within each pretest-score group and intervention group.</td>
<td>100</td>
</tr>
</tbody>
</table>
CHAPTER I
INTRODUCTION

Single-subject research designs are widely used in the fields of applied behavior analysis and special education. These designs allow researchers to determine precisely when and to what magnitude an independent variable affects a dependent variable through repeated measurement of the behavior over time under different experimental conditions (Kazdin, 1982; Kratochwill, 1978; Kratochwill & Levin, 1992; Parsonson & Baer, 1978). Commonly, the dependent variable is measured repeatedly before the introduction of the independent variable (i.e., baseline or “A” phase) and during and/or after the implementation of the independent variable (i.e., treatment or “B” phase). These phases are then replicated within subject (i.e., ABAB or multiple baseline across behaviors or settings) or between subjects (i.e., multiple baseline across subjects).

Inferences about a causal relationship between the independent and dependent variable in single-subject research are made on the basis of visual analysis (Kazdin, 1982; Kratochwill, 1978; Kratochwill & Levin, 1992; Parsonson & Baer, 1978). That is, data from each phase are graphed serially and judgments about intervention effects are made based on visually apparent differences between phases. Specifically, researchers look for a) changes in the slope, level, and/or variability of the data that would not be predicted to occur without the active manipulation of the independent variable and b) multiple replications of the effect over time (Kratochwill et al., 2010).

Although visual analysis is the cornerstone of single-subject research, some investigators have found that experts in visual analysis often disagree about what constitutes an intervention effect (e.g., DeProspero & Cohen, 1979). Further, few studies
have examined the variables that influence expert judgments. The potential lack of reliability across visual analysts raises several concerns. Although human judgment is an inevitable element of any scientific process, visual analysis may be particularly vulnerable to individual bias (Kazdin, 1982). Some have argued that visual analysis is likely to be influenced by idiosyncratic factors and individual variability related to “history, training, experience, and vigilance” (Fisch, 1998, p. 112) in the absence of formal guidelines to operationalize the decision-making process (Furlong & Wampold, 1981; Kazdin, 1982). If the analysis of data is susceptible to individual variability (i.e., if it is subjective), judgments about experimental control may differ based on who is interpreting the data. As a result, the determination of the “truth” becomes complicated. The field has long emphasized the importance of objective measurement in the direct observation of behavior; one primary purpose of assessing interobserver agreement is to rule out observer differences or biases as potential sources of variability in the data. A parallel can be drawn between data collection and data analysis; unless the analysis of results is also objective, our science becomes difficult to interpret.

This problem may manifest in the peer-review process. It is possible that two different sets of reviewers could come to different conclusions about whether the same graph depicts a pattern that demonstrates experimental control. If reviewers disagree with one another, the decision is essentially put solely in the hands of the editor. This undermines the purpose and value of peer reviews.

In terms of translating research to practice, this potential subjectivity raises concerns about the empirical support of our “research-based interventions.” If the interpretation of results varies depending on who is analyzing the data, we may be
recommending that practitioners implement interventions that lack strong empirical support. This may become more apparent as educational institutions (e.g., What Works Clearinghouse) emphasize the use of systematic reviews of single-subject literature to make clinical recommendations (Kratochwill et al., 2010). The quality and rigor of these reviews, and resulting practice recommendations, are dependent on attaining consensus among visual analysts with regard to the presence or absence of an intervention effect.

In sum, lack of agreement among visual analysts has important implications for both research and practice. Inconsistent interpretation across experienced judges suggests that, in some cases, visual analysis of graphical data is far from intuitive. In turn, the critical role of developing and evaluating systematic procedures for training individuals to conduct valid and reliable visual analysis becomes apparent.
CHAPTER II

LITERATURE REVIEW

Agreement Between Raters

Although the process of visually examining data seems relatively straightforward, research suggests that experienced visual analysts often disagree about whether a given graph demonstrates an experimental effect. Ottenbacher (1993) conducted a review of research on agreement among visual analysts; the mean interrater agreement across 14 reviewed studies was 58% ($SD=12\%$). The studies varied along a number of characteristics, including participants’ level of expertise (e.g., experts or students/clinicians), source of the graphs (e.g., real or hypothetical data), form of the graphs (e.g., AB or ABAB designs), and form of the response (e.g., dichotomous or scaled); Ottenbacher (1993) coded studies for these various characteristics and computed average agreement scores for studies with common characteristics. None of the agreement scores were significantly different from one another and all were between 50% and 65%; in other words, different features of the graphs (e.g., the presence or absence of a trend line) and participants (e.g., experts or students/clinicians) did not appear to produce different agreement scores. Interpretation of these results is complicated by the small number of studies contributing to some characteristics (e.g., real data, n=3) and the degree of overlap between certain characteristics (e.g., presence of a trend line and student participants; Bailey, 1984; Hojem & Ottenbacher, 1988); however, they do suggest that agreement among visual analysts is poor across a number of study and participant characteristics. A closer examination of select studies reviewed by
Ottenbacher (1993) and recent research on visual analysis reveals nuances within and across studies that may influence interrater agreement.

DeProspero and Cohen (1979) asked 114 editors and reviewers of the *Journal of Applied Behavior Analysis* (JABA) and the *Journal of the Experimental Analysis of Behavior* (JEAB) to rate the demonstration of experimental control evidenced by each of nine graphs on a scale of 0 (low experimental control) to 100 (high experimental control). The authors generated a pool of 36 graphs, each depicting a reversal design (i.e., ABAB) with programmed level changes, slopes, and variability. Interrater agreement was determined by calculating the correlation between pairs of raters who received the same graph packets. The average Pearson correlation between raters was 0.61, and was poor for all but the most obvious graphs. This level of agreement would be considered grossly inadequate for direct observation of a dependent variable.

Kahng et al. (2010) replicated and extended the study by DeProspero and Cohen (1979), speculating that improvements in training in the 30 years since the original study might result in better agreement among experts. Kahng et al. (2010) used the same graph generation techniques and 100-point scale used by DeProspero and Cohen (1979), sent 45 editors of *JABA* all 36 graphs, and asked them to make a dichotomous “yes” or “no” determination of experimental control for each graph in addition to rating it on the 100-point scale. In contrast to the findings of DeProspero and Cohen (1979), Kahng et al. (2010) reported an average Pearson correlation of 0.93 between pairs of raters on the 100-point measure across all types of result patterns. One potential explanation for these contradictory findings is that the growth of the field and creation of standards such as those imposed by the Behavior Analyst Certification Board have increased formal
training opportunities in visual analysis. The potential role of training in the divergent results of Kahng et al. (2010) and DeProspero and Cohen (1979) suggests a need to empirically evaluate methods of training individuals in visual analysis. The difference may also be attributed to the inclusion of the dichotomous response by Kahng et al. (2010); however it is not clear how this might have affected the scale ratings assigned by participants. In addition, Kahng et al. (2010) defined experimental control in their instructions for participants; DeProspero and Cohen (1979) did not. Finally, participant differences may have contributed to the divergent results: Kahng et al. (2010) included only members of the editorial board and associate editors of JABA, whereas DeProspero and Cohen (1979) included editors and guest reviewers of both JABA and JEAB.

Kahng et al. (2010) is an outlier in this literature; other studies on the reliability of visual analysis across raters have generally bolstered the findings of DeProspero and Cohen (1979). Ottenbacher (1986) asked 46 occupational therapists to determine whether or not each of five AB graphs depicted a clinically significant change in behavior. Fewer than 80% of the participants agreed about the presence or absence of a clinically significant effect for three of the five graphs, and the average agreement across the graphs was 53%. However, it is difficult to generalize the results of this study given that only five graphs were included.

Gibson and Ottenbacher (1988) addressed this limitation by using 24 AB graphs that were generated with programmed mean shifts, level changes, slope, variability, overlap, and autocorrelation across phases. Twenty rehabilitation therapists rated their level of agreement with the statement, “There is a significant change in performance across the two phases,” on a 6-point Likert-type scale. To analyze agreement among
participants, the authors divided scale responses into two categories: ratings of 0, 1, and 2 indicate disagreement with the statement (i.e., no change) and ratings of 3, 4, and 5 indicate agreement with the statement (i.e., significant change). Even after collapsing specific ratings into two broad categories, agreement was above 80% for only 11 of the 24 graphs. The authors also analyzed which features of graphs (e.g., slope, variability, etc.) were associated with high agreement and disagreement among raters. Slope was strongly correlated with disagreement; that is, raters tended to disagree when there was a change in slope across the phases. On the other hand, large mean shifts and immediate level changes were strongly associated with agreement among raters. Other graphic features, such as variability, autocorrelation, and overlap, were not correlated with interrater agreement.

Continuing to evaluate agreement among visual analysts, Ottenbacher (1990a) generated graphs and analyzed data using procedures similar to those used by Gibson and Ottenbacher (1988). Instead of using a response scale, Ottenbacher (1990a) asked participants to respond “yes,” “no,” or “uncertain” about the demonstration of a clinically significant change across phases. Sixty-one practitioners from various fields (e.g., special education, speech and language pathology, occupational therapy) viewed six AB graphs; agreement among participants was above 80% for only two of these. Like Gibson and Ottenbacher (1988), Ottenbacher (1990a) analyzed which graphic characteristics were associated with interrater agreement. Large level changes were correlated with high agreement; in other words, when a graph depicted a large level change, raters tended to agree about the demonstration of an effect. High variability and
slope changes, on the other hand, were correlated with disagreement, suggesting that these features were associated with inconsistent ratings among the participants.

In a more recent study, Brossart, Parker, Olson, and Mahadevan (2006) generated 35 AB datasets and asked faculty and doctoral students in Educational Psychology to rate their level of certainty that the behavior improved due to the intervention on a 5-point scale. The authors provided a scenario describing the target behavior and intervention depicted in the graphs. Like previous studies, agreement was poor: the average individual rater-to-group correlation across the graphs was 0.58.

Lieberman, Yoder, Reichow, and Wolery (2010) evaluated the effects of two graphic characteristics on experts’ ratings of functional relationships evidenced by multiple baseline across participants. The authors manipulated slope and the consistency in the latency of change across participants in 16 generated multiple baseline graphs, providing contextual information about the dependent variable, setting, participants, and intervention. For each graph, review board members from three behavioral journals judged whether they were a) confident that the graph depicted a functional relation, b) confident that the graph did NOT depict a functional relation, or c) not confident either way. The average percent agreement across all pairs of raters and graphs was 40%; only 4% of pairs had 80% or higher agreement on all graphs. Interestingly, the level of agreement was essentially the same across all graph types; that is, the average agreement for graphs with steep slopes that were thought to be more obvious was about the same as those with shallow slopes. Given this poor agreement, the authors concluded that research is needed in order to evaluate “…whether direct instruction with performance feedback affects agreement among visual analysts when making functional relation
judgments....” (Lieberman et al., 2010, p. 42).

All of the preceding studies used hypothetical data that was generated by the authors to feature specific combinations of graphic characteristics thought to influence visual analysis. Several authors have investigated interrater agreement with real, published data. Jones, Weinrott, and Vaught (1978) selected 24 JABA graphs with small treatment effects and few data points per phase to compare the agreement between visual and statistical analysis (see “Statistical Analysis”); they also reported agreement between visual analysts. Eleven researchers, professors, and graduate students evaluated whether or not there was a level change for each AB contrast in 24 graphs: agreement ranged from 0.04 to 0.79, with a median agreement score of 0.39. Similarly, Park, Marascuilo, and Gaylord-Ross (1990) randomly selected AB phases from JABA graphs and asked five participants (1 Ph.D, 4 Master’s level) to determine whether each of 44 graphs depicted “significant,” “nonsignificant,” or “unclear” intervention effects. As in Brossart et al. (2006), contextual information was provided for each graph. The authors reported 60% mean agreement between pairs of raters.

As part of a larger study on training visual analysis, Normand (2003) asked three Ph.D.-level experts to evaluate 60 ABA graphs from JABA, and provided contextual information about the participant, setting, behavior and intervention. Normand (2003) used a multiple-choice format with the following response options: (a) the intervention caused an increase in behavior, (b) the intervention caused a decrease in behavior, (c) there was no change in behavior, (d) there was a change that was not caused by the intervention or (e) can’t make a decision based on the graph. The three experts agreed unanimously on the multiple-choice answers for only 40% of the graphs. When the
multiple-choice responses were dichotomized into “change” and “no change” categories, the experts agreed on 67% of the graphs.

All previously reviewed studies examined reliability on graphs depicting baseline and intervention phases; Danov and Symons (2008) evaluated interrater agreement on 26 multielement graphs depicting functional analyses selected randomly from articles published in *JABA*. Forty-three participants (29 students, 7 faculty, and 7 unidentified) selected one of 12 response options, which included all possible functions and combinations of functions (e.g., undifferentiated, maintained by escape, maintained by attention and tangible reinforcement). The overall mean interrater agreement was 0.63; for faculty raters only, agreement was virtually equivalent at 0.65. It should be noted that substantially more response options (12) were available than the other reviewed studies, and this may have contributed to the low agreement observed here.

The results of Kahng et al. (2010) notwithstanding, these studies suggest that agreement among visual analysts is often poor. However, significant limitations of these studies must be acknowledged. Participants’ experience with visual analysis prior to the study was often poorly described (e.g., “All subjects…had experience in the graphing and charting of student progress…”; Ottenbacher, 1990a, p. 438) or absent (e.g., Gibson & Ottenbacher, 1988; Ottenbacher, 1986). Based on the descriptions available, it appears that there was wide variation in expertise across the studies. Aside from Brossart et al. (2006), Kahng et al. (2010), and Lieberman et al. (2010), authors rarely defined the judgment task operationally (e.g., participants were asked to judge whether graphs showed a “significant change,” Ottenbacher, 1990a, p. 438); as a result, participants within a study may have interpreted the task differently. Across studies, comparisons are
difficult based on the nature of the change participants were asked to judge (functional vs. clinical) and the type of response participants were asked to give (dichotomous vs. scaled). Some studies contained relatively few graphs (Ottenbacher, 1986; Ottenbacher, 1990a), which raises questions about whether sufficient combinations of data patterns were sampled. Several studies used only graphs depicting acquisition targets (e.g., Brossart et al., 2006; DeProspero & Cohen, 1979; Kahng et al., 2010) and many did not specify whether acquisition targets, reduction targets, or both were included (Gibson & Ottenbacher, 1988; Ottenbacher, 1986; Ottenbacher, 1990a). Importantly, studies differed in the magnitude and combinations of level change, trend change, and variability depicted in the graphs, and none described changes relative to the variation in the data. For example, DeProspero and Cohen (1979) and Kahng et al. (2010) described “mean shift” as the percent of change in means from baseline to treatment (e.g., a baseline mean of 10 and a treatment mean of 15 would result in a mean shift of 0.50), without accounting for within-phase variability (e.g., by reporting mean shift as an effect size).

Finally, few of these studies provided contextual information, such as hypothetical participants, settings, interventions, or behaviors; however, those that did provide this information reported very poor agreement (Brossart et al., 2006; Lieberman et al., 2010; Normand, 2003; Park et al., 1990). Many in the field have criticized this body of literature for omitting this contextual information that is available in the clinical and research settings and thought to influence visual analysis (Parsonson & Baer, 1992). The results of studies that did include this information, however, suggest that its inclusion may not substantially impact agreement. Although the limitations of these studies make it difficult to draw firm conclusions about agreement among visual analysts, the fact that
10 out of 11 studies found low levels of agreement supports the general contention that reliability of interpretation of single-subject research results is problematic under a variety of circumstances. These results have led researchers to investigate alternative methods of analyzing single-subject data, including various methods of statistical analysis.

**Statistical Analysis and Visual Analysis**

Compared with statistical analysis, which involves precise methods and a relatively objective (though not necessarily valid) criterion for determining the presence of an intervention effect, visual analysis has been criticized for its apparent subjectivity (Jones, Vaught, & Weinrott, 1977). However, common features of single-subject data, such as autocorrelation and trend, have posed significant challenges for those attempting to analyze these data statistically (Campbell, 2004). The error (i.e., variation from a regression line) in autocorrelated data is not independent; that is, the error associated with one data point influences the error associated with the subsequent data point. Because autocorrelation is defined in terms of deviation from a regression line, it cannot be attributed to simple trend across a dataset. This property violates the fundamental assumption of independent observations required by most statistical tests, and also influences effect size measures to different degrees (Brossart et al., 2006; Manolov & Solanas, 2008; Shadish, Rindskopf, & Hedges, 2008).

It has also been argued that visual analysis is relatively conservative given the idea that “a difference has to be seen to be affirmed” (italics added; Baer, 1977, p. 169). In other words, visual analysis may result in the identification of fewer, but more potent,
behavior-change variables than statistical analysis. Michael (1974) and Baer (1977) have argued that visual analysis is relatively conservative and that statistical methods are inappropriate for single-subject research because they might detect intervention effects that are much smaller than traditionally have been accepted in the field. Despite mathematical and philosophical challenges to statistical analysis of single-subject data, researchers have investigated the utility of various methods and compared the accuracy of these methods to visual analysis.

**Inferential Statistics**

Jones et al. (1977) suggested that interrupted time-series analysis (ITSA) be used as a supplement to visual analysis. In this method, autocorrelation is removed from each data point by estimating the autocorrelation parameter ($\alpha$), multiplying the previous data point ($y_{n-1}$) by $\alpha$, and subtracting the product from the current data point (i.e., $y_n-\alpha*y_{n-1}$; Crosbie, 1993). The adjusted data are then subjected to linear regression to test for statistically significant changes in level and trend between adjacent phases. The authors applied this analysis to several published datasets, and found that the results of ITSA sometimes confirmed the original authors’ conclusions, sometimes contradicted them, and sometimes detected additional effects not mentioned by the original authors. In addition to evaluating interrater agreement, Jones et al. (1978) conducted ITSA on each of the AB contrasts within the 24 graphs and compared the results to the decisions made by visual analysts. Agreement between ITSA and visual analysis ranged from 50% to 65%, with lower agreement for graphs with (a) higher autocorrelation and (b) a statistically significant effect as determined by the ITSA.
Some authors have also compared the results of visual analysis with the results of a split-middle analysis (Kazdin, 1982; White, 1974). The *split-middle line* is a quick estimation of the least-squares linear regression line that is considered more accurate than visually estimating a line of best fit (Cooper, Heron, & Heward, 2007). Typically, the split-middle line is used to aid visual analysis in a fairly informal way (Kazdin, 1982). The *split-middle analysis* uses the split-middle line to evaluate the statistical significance of treatment effects. The analyst projects the split-middle line of the baseline phase into the treatment phase, and counts the proportion of data points falling above (or below) the line. The binomial formula is applied to determine the probability of that proportion of points falling above (or below) the line by chance (i.e., the p-value). Table 1 provides an overview of analytical methods based on the split-middle line that have been used in various studies.

Studies comparing the results of visual analysis with results of the split-middle analysis have found poor agreement between these two methods. For example, Ottenbacher (1986) found that a majority of visual analysts (i.e., >50%) agreed with the results of the split-middle analysis on only two of five graphs (40%). In Ottenbacher (1990b), 30 health and education professionals decided whether there was a significant change in performance across the phases of 24 AB graphs; the split-middle analysis was also conducted on the graphs. In this study, the majority of raters agreed with the split-middle analysis on 10 of 24 graphs (46%).
### Table 1

**Description of Analytical Approaches Related to the Split-Middle Line**

<table>
<thead>
<tr>
<th>Approach</th>
<th>Studies</th>
<th>Procedures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Split-middle line only</td>
<td>Bailey (1984)</td>
<td>1. Split-middle lines superimposed on baseline and/or treatment phase</td>
</tr>
<tr>
<td></td>
<td>Harbst et al. (1991)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Normand &amp; Bailey (2006)</td>
<td></td>
</tr>
<tr>
<td>Split-middle analysis</td>
<td>Hojem &amp; Ottenbacher (1988)</td>
<td>1. Draw split-middle line for baseline phase</td>
</tr>
<tr>
<td></td>
<td>Ottenbacher (1986)</td>
<td>2. Project into treatment phase</td>
</tr>
<tr>
<td></td>
<td>Ottenbacher (1990b)</td>
<td>3. Count proportion of data points above or below line</td>
</tr>
<tr>
<td></td>
<td>Richards et al. (1998)</td>
<td>4. Consult table based on binomial test OR use binomial test to identify p-value</td>
</tr>
<tr>
<td></td>
<td>Fisher et al. (2003)</td>
<td>5. If using table, target proportion of data points must be above split-middle line</td>
</tr>
<tr>
<td>Refined split-middle analysis</td>
<td>Fisher et al. (2003)</td>
<td><strong>Dual Criterion (DC)</strong></td>
</tr>
<tr>
<td></td>
<td>Colon (2006)</td>
<td>1. Steps 1-4 from split-middle analysis</td>
</tr>
<tr>
<td></td>
<td>Stewart et al. (2007)</td>
<td>2. Draw mean line for baseline phase</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3. Project into treatment phase</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4. Count proportion of data points above or below line</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5. Consult table based on binomial test OR use binomial test to identify p-value</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6. If using table, target proportion of data points must be met for both criterion lines</td>
</tr>
<tr>
<td></td>
<td>Colon (2006)</td>
<td><strong>Conservative Dual Criterion (CDC)</strong></td>
</tr>
<tr>
<td></td>
<td>Stewart et al. (2007)</td>
<td>1. Steps 1-3 from DC Method</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2. Adjust both criterion lines by .25 standard deviations</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3. Steps 4-6 from DC Method</td>
</tr>
</tbody>
</table>
Richards, Taylor, and Ramasamy (1997) extended the findings of Ottenbacher (1990b) by evaluating agreement between the split-middle analysis and the visual analysis of 62 undergraduate and graduate students in special education on the same graphs used by Ottenbacher (1990b). The two methods of analysis were in agreement on 11 graphs (46%), and there were no significant differences between the undergraduate and graduate students in terms of agreement with the split-middle analysis. As a group, visual analysts reported an effect when the split-middle analysis did not on 25% of the graphs; visual analysts reported a non-effect when the split-middle analysis did on 29% of the graphs.

These studies raise an important point about comparing statistical analysis and visual analysis: when agreement between the two methods is low, it is not clear which method should be considered more valid (Jones et al., 1978). Given the reliability of statistical analysis and the apparent inconsistency of visual analysis, Jones et al. (1978) suggested that the results of statistical analysis are more valid. However, reliability is necessary but not sufficient for validity; that is, a procedure may be replicable and produce the same results every time it is implemented, but those results may not be accurate (Matyas & Greenwood, 1990). In addition, researchers have found that the accuracy of ITSA is compromised by phases with small numbers of data points (Huteima, 1985) and the split-middle analysis is negatively impacted by autocorrelation (Crosbie, 1987).

Matyas and Greenwood (1990) attempted to circumvent this question by comparing visual analysts’ judgments about generated AB graphs to the statistical parameters used to create the graphs. Graduate students who were enrolled in a course on
single-subject designs and had just completed a series of lectures on visual analysis served as participants in the study. The authors manipulated three parameters – level change, random error, and autocorrelation – each at three different levels, to produce 27 graphs. Participants judged whether each AB graph depicted (a) no effect, (b) a level change, (c) a trend change, (d) level and trend change, or (e) other type of change. The authors used the programmed parameters as the criterion for determining the accuracy of participants’ responses; that is, the programmed parameters served as the “correct” answer. (Because random error was included in the formula used for generating the graphs, the obtained parameters differed from the programmed parameters.) The authors analyzed participants’ rates of Type I and Type II errors as a function of different levels of the programmed parameters (i.e., autocorrelation, variability, and level change). Participants committed the highest percentage (80%) of Type I errors on the graph with no programmed level change, high variability, and a medium amount of autocorrelation; Type I errors were also high for the graphs with medium variability and autocorrelation and high variability and high autocorrelation. Type II error rates were below 20% for all graphs with a programmed level change, regardless of the presence of variability and autocorrelation. These results suggest that visual analysis may not be as conservative as some have claimed (e.g., Baer, 1977), particularly when data are variable and autocorrelated. However, the authors generated only one graph representing each combination of the three parameters. It is possible that this graph was not representative of the programmed parameters (Fisher, Kelley, & Lomas, 2003). In other words, even though certain parameters were programmed, the obtained parameters may have differed substantially given the inclusion of the random error variable. Random error adds
variation to the data path, and this coupled with the autocorrelation parameter may result in a visually apparent treatment effect that was not programmed. For example, one graph in Matyas and Greenwood had a programmed effect size of 0 (i.e., no effect), but an approximate obtained effect size of 1.67.

**Effect Sizes**

Like inferential statistics, the appropriate use of effect sizes to summarize the results of single-subject research has been a point of contention (Brossart et al., 2006; Kratochwill et al., 2010; Olive & Smith, 2005; Parker et al., 2005). Effect sizes may provide a useful quantitative summary of behavior change across phases. Although such metrics do not, in and of themselves, address questions of causality, Parker et al. (2005) argued that studies with strong internal validity and large effect sizes may suggest that the intervention caused the observed change in behavior. However, most effect size calculations that have been applied to single-subject data are negatively affected by autocorrelation and trend (Brossart et al., 2006), the same challenges that confront statistical significance tests.

Researchers have investigated fairly straightforward effect size metrics that are simple to calculate and interpret (e.g., percent of nonoverlapping data points, PND; Scruggs & Mastropieri, 1998) as well as more complex statistical analyses such as multilevel modeling (van den Noortgate & Onghena, 2003) and Tau-U (Parker, Vannest, Davis, & Sauber, 2011). Although work continues on identifying an effect size metric that is conceptually and statistically appropriate for single-subject data, two important limitations of the research to this point preclude a reliance on effect sizes. First, the degree to which different effect sizes correlate with one another when applied to the same
dataset varies (Brossart et al., 2006; Campbell, 2004; Olive & Smith, 2005). Campbell (2004) applied three effect sizes to published datasets and found significant correlations between the PND, percent of zero data points (PZD), and mean baseline reduction (MBLR); correspondence between effect size measures suggests that they may be measuring the same phenomenon. In contrast, Brossart et al. (2006) calculated five different effect sizes for 35 AB graphs and found that $R^2$ values between statistics ranged from 0.034 to 0.895 for the subset of graphs that depicted a “very effective” intervention. Second, some effect sizes have poor correlations with the results of visual analysis (e.g., $r=.39$; Brossart et al., 2006; Maggin, Johnson, Chafoulea, Ruberto, & Berggren, 2012). As previously discussed with regard to tests of statistical significance, it is impossible to know whether the resulting effect size or visual analysis is more valid in the case of disagreement. Although the continued exploration of effect sizes for single-subject data is worthwhile, the identification of an appropriate metric has been elusive so far.

### Agreement Between Raters with Visual Aids

Several researchers have examined whether trend lines, also referred to as lines of progress (Bailey, 1984) or celeration lines (Harbst, Ottenbacher, & Harris, 1991), improve the agreement among visual analysts. This line summarizes the direction and magnitude of change (or slope) across a phase, and may be estimated using least-squares regression or the split-middle line (White, 1974). In this group of studies, the authors superimposed trend lines on graphs but they did not provide training on the interpretation of the line.
Bailey (1984) examined the separate and combined effects of trend lines and semilogarithmic scales on the interrater agreement of 13 Masters students in Special Education. On a semilogarithmic scale, the same distance on the graph represents the same relative change. In other words, the distance between 1 and 2 is equal to the distance between 5 and 10 because both distances represent the same proportional change (i.e., doubling). Bailey (1984) presented five JABA graphs in four different ways: (a) as presented by the authors, on equal-interval scales without trend lines; (b) on equal-interval scales with trend lines; (c) on semilogarithmic scales without trend lines; and (d) on semilogarithmic scales with trend lines. The 20 graphs were presented in random order, and participants were asked to evaluate whether there was a change in level and/or a change in slope for each phase change. Bailey conducted a three-factor analysis of variance to evaluate the effects of scale type, presence of trend line, and type of change on interrater agreement; he found a significant main effect \( (p<.01) \) for the presence of a trend line and a significant interaction \( (p<.05) \) between the presence of a trend line and the type of change. In other words, the trend line improved interrater agreement significantly for judgments of both level change and trend change; however, the trend line had a more marked improvement on judgment of trend than on judgment of level. On equal interval scales without a trend line, agreement on trend change and level change was 66% and 73%, respectively. When trend lines were added, agreement increased to 84% and 85% for trend and level change. The type of scale did not have a significant effect on agreement among raters.

Given these promising results, it is unsurprising that other researchers continued investigating the effects of trend lines on interrater agreement. In a replication and
extension of the study by Gibson and Ottenbacher (1988), Harbst et al. (1991) used the same graphs, scale and data analysis techniques, but included trend lines to evaluate the effect of this visual aid on interrater agreement among 30 practicing occupational therapists and physical therapists. However, more than 80% of participants agreed on only 7 of the 24 graphs; in contrast, using the same graphs without trend lines, Gibson and Ottenbacher (1988) obtained higher than 80% agreement on 11 of the same 24 graphs. Thus, the trendlines appeared to reduce interrater agreement in this study.

One potential explanation for the low agreement among participants in Harbst et al. (1991) is their level of training and experience in visual analysis; although the authors detail their clinical experience and educational background, specific information about experience and/or coursework in visual analysis is not provided. It might be argued that Board Certified Behavior Analysts (BCBA) would be better prepared through training and coursework experiences to conduct accurate visual analysis. Normand and Bailey (2006) investigated the effects of trend lines on the interrater agreement of this population using actual and manipulated AB and ABA graphs. Trend lines were superimposed onto some of the graphs, while others were presented without trend lines. Five BCBAs judged whether each of 24 graphs depicted an increase, decrease, or no change in behavior from baseline to treatment. In addition, the participants were asked to “think aloud” as they judged each graph; verbal reports were audiotaped, transcribed, and analyzed. Although participants made more comments overall, and more statements about level change, trend, and variability specifically in response to the graphs with trend lines, they were less accurate when trend lines were present than when they were not (67% and 78%, respectively), and more than 80% agreed on only 12 of the 24 graphs.
The varied results across these three studies suggest that simply providing trend lines is unlikely to improve an individual’s visual analysis of single-subject data. Numerous factors, including a conceptual understanding of single-subject research methodology (Tindal & Deno, 1983) and previous experience with visual and statistical analysis (Harbst et al., 1991), may influence how trend lines are interpreted. Given that visual aids alone do not appear to be sufficient for increasing the accuracy or agreement of visual analysts, systematic training procedures or explicit rules for interpreting graphed time series data and/or trend lines might enhance the effects of visual aids or make them unnecessary altogether. A handful of studies have investigated the impact of training, alone and in combination with visual aids, on the visual analysis of single-subject data.

**Training Visual Analysis**

Hojem and Ottenbacher (1988) conducted the first study on training visual analysis with 39 upper-level undergraduates. Participants were randomly assigned to one of two groups that each received different training content. The visual group participated in a lecture on graphing conventions and the components of visual analysis, including level, trend, and variability. Definitions and examples of each component were provided; this approach could be described as a *global visual* analysis. Participants received instruction on what features to examine, but were not taught specific heuristics for how to estimate slope, level, or variability (see Table 2 for definitions of approaches to visual analysis and related studies). The quantitative group received instruction and practice in drawing split-middle lines and using the split-middle analysis to decide whether or not a treatment effect had occurred. This is a *computational-mechanical*
approach to visual analysis, as participants learned to calculate the split-middle line and consult the binomial distribution to obtain a \( p \)-value. After each group received their respective instruction, they determined whether a significant change had occurred from baseline to treatment in five AB graphs using a six-point Likert-type scale. For four out of five graphs, the two groups ratings were significantly different (\( p < .05 \)). Within groups, 80% or more of the participants using visual analysis agreed on only one graph; however, 75% or more agreed on four of the five graphs. In the group of participants using the split-middle analysis, 80% or more agreed on three of the five graphs.

Although this suggests that the split-middle analysis produced a higher reliability across raters, the authors did not evaluate accuracy of ratings by comparing participants’ responses to expert opinion or programmed changes. As previously discussed, the validity of the split-middle analysis has been empirically evaluated and found to be negatively influenced by autocorrelation (Crosbie, 1987), which is often present in single-subject data (Matyas & Greenwood, 1990). In addition, pretests were not conducted to evaluate within-group agreement prior to training; therefore, other participant characteristics may have influenced levels of agreement. Despite these limitations, it is possible that the instruction provided in the global visual approach was not explicit enough about how to estimate and compare level, slope, and variability.

Fisher et al. (2003) used a Monte Carlo simulation to evaluate the limitations of the split-middle analysis and to develop a refined procedure for visual analysis incorporating this technique. Fisher et al. (2003) generated graphs using an autoregressive model with and without programmed level changes and degrees of autocorrelation to evaluate the types of errors produced by the split-middle
Table 2

*Approaches to Visual Analysis Evaluated in Training Studies*

<table>
<thead>
<tr>
<th>Approach</th>
<th>Definition</th>
<th>Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global visual</td>
<td>Visual only; no rules or guidelines</td>
<td>Hojem &amp; Ottenbacher (1988)</td>
</tr>
<tr>
<td>Computational</td>
<td>Calculating slope and/or level lines</td>
<td>Jostad (2011)</td>
</tr>
<tr>
<td>Mechanical</td>
<td>Counting data points above/below given criterion lines</td>
<td>Fisher et al. (2003)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Colon (2006)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Stewart et al. (2007)</td>
</tr>
<tr>
<td>Computational-mechanical</td>
<td>Calculating slope and/or level lines and counting data points above/below lines</td>
<td>Hojem &amp; Ottenbacher (1988)</td>
</tr>
<tr>
<td>Structured visual</td>
<td>Following explicit, step-by-step instructions</td>
<td>Hagopian et al. (1997)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Colon (2006)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Jostad (2011)</td>
</tr>
</tbody>
</table>

This analysis produced excessive Type I errors when sampling error created an apparent, but unprogrammed, trend in baseline. To correct for this, Fisher et al. (2003) projected a second criterion line, the baseline mean line, into the treatment phase. The resulting dual-criterion (DC) method has two steps: (1) the split-middle analysis (i.e., projecting the baseline split-middle trend line into the treatment phase and counting the number of data points above or below the line) and (2) a similar analysis using the baseline mean line (i.e., projecting the baseline mean line into the treatment phase and counting the number of data points above or below the line). As in the split-middle analysis, a certain number of data points must fall above (or below) both lines for the determination of an effect to be made. Initial tests of the DC method resulted in Type I error rates that were considered too high; Fisher et al. (2003) then developed a more
conservative version of the DC method, called the conservative dual-criterion method (CDC), by raising both criterion lines by 0.25 baseline standard deviations.

After developing the refined methods, Fisher et al. (2003) generated 30,000 graphs from each of 20 different combinations of programmed level change and autocorrelation parameters. A Monte Carlo simulation evaluated each graph according to the criteria established by the SM, DC, and CDC methods along with general linear modeling (GLM) and interrupted time series (ITSE). The authors evaluated the rates of Type I and Type II errors for each method using the programmed parameters to specify the “correct” interpretation. The split-middle analysis committed the lowest proportion of Type II errors, but also the highest proportion of Type I errors (i.e., it was overly ‘liberal’ in declaring that there was a difference). The DC and CDC methods both produced very low Type I errors, but were still sensitive enough to detect most programmed effects; that is, both methods had sufficient power to avoid high rates of Type II errors.

Based on these outcomes, Fisher et al. (2003) trained behavior therapists to use the DC method to analyze AB graphs generated with the same parameters employed in the Monte Carlo simulation. In each baseline session, participants received a packet of 20 graphs and circled yes or no in response to the question, “Is there a reliable treatment effect?” In training, participants were provided with sample graphs containing the two criterion lines, a table specifying the number of data points that have to be above both lines to demonstrate a treatment effect, and rules for using the criterion lines. In the one-on-one training, the experimenter modeled how to apply the DC method to the sample graphs. These approaches to visual analysis are mechanical in nature and do not require
knowledge of level, trend, or variability; the analyst simply counts the number of data points above each criterion line and consults the table to determine whether that number of data points corresponds to an effect or not. Following training, assessment sessions resumed in a manner very similar to baseline; however, after training had been completed, all assessment graphs contained the two criterion lines. In baseline, participants’ accuracy compared to programmed parameters was generally stable between 50-60%; after training and with the criterion lines, accuracy increased to between 80-100%. Fisher et al. (2003) then extended these findings by training a large group of behavior analysts to use the DC method, with similar outcomes. Given the paucity of research on training individuals to analyze single-subject data, the study by Fisher et al. (2003) is an important contribution that has motivated further examination of these methods. However, it is difficult to argue that the DC and CDC methods are truly visual; in essence, Fisher et al. (2003) trained participants to conduct the same data analysis procedures that they had designed a computer program to run in an earlier phase of the study.

Stewart, Carr, Brandt, and McHenry (2007) extended the findings of Fisher et al. (2003) by training six university students to apply the CDC method in the analysis of AB graphs. In their evaluation of various visual inspection methods, Fisher et al. (2003) found the CDC method to be slightly more accurate than the DC method; however, the authors conducted the evaluation and training studies concurrently and had already begun training participants to use the DC method. Stewart et al. (2007) trained undergraduates to use the CDC method, which is similar to the DC method in that the individual analyzing the graph counts the number of data points above or below the criterion line
and consults a table to determine whether or not there is an intervention effect. Stewart et al. (2007) included a “traditional” visual analysis lecture following baseline to evaluate the effectiveness of didactic instruction on visual analysis. Otherwise, their procedures were very similar to Fisher et al. (2003) in terms of graph generation and assessment procedures. In baseline, most participants correctly judged approximately 50% of the graphs as demonstrating an effect or not. The traditional lecture consisted of a 12-minute videotaped lecture on the components of visual analysis as described by Cooper, Heron, and Heward (1987); continued assessment demonstrated no improvement in accuracy following this lecture. Participants were then trained to use the CDC method via a brief, videotaped lecture, and provided with the binomial table and explicit rules for interpreting the data. Like Fisher et al. (2003), following training in the CDC method, all assessment graphs contained the necessary criterion lines. After receiving training in the CDC method and given the criterion lines, participants’ accuracy increased to between 80-100%. Stewart et al. (2007) returned to baseline following the CDC phase; during this baseline condition, the criterion lines were not present on assessment graphs and all participants’ accuracy decreased to around 50%. Stewart et al. (2007) acknowledge this important limitation of the DC and CDC visual inspection methods: if accurate analysis does not maintain in the absence of the supplemental criterion lines, these methods may not be practical for use with published literature. Further, the CDC method requires that the split-middle line and mean line be drawn and then adjusted by 0.25 standard deviation – this adjustment necessitates statistical analysis and, for precise results, a data extraction program.
Researchers have also evaluated the effect of training raters to use a *structured visual* approach to the analysis of single-subject data. *Structured visual* approaches provide the rater with explicit, step-by-step criteria to follow in their analysis. Hagopian et al. (1997) developed structured criteria to aid in the inspection of functional analysis graphs. The criteria were designed to operationalize the process of evaluating a multielement graph displaying rates of behavior across several conditions of a functional analysis. After developing preliminary criteria, the authors discussed and came to a consensus on the function depicted in each of 64 graphs, consisting of actual data from individuals treated for severe behavior. The criteria were then refined and applied to the graphs; an agreement of 0.94 was obtained between the structured criteria and expert consensus. Hagopian et al. (1997) trained three doctoral students to apply the structured criteria; in baseline prior to training, they identified the correct function of about 60% of the graphs. Following training and with the structured criteria, accuracy increased to 80-100%. Although the materials in this study were limited to assessment graphs with multielement designs, it does suggest that specific guidelines may assist novices in more accurately evaluating behavioral data – even in the absence of experimenter-provided trend lines.

To investigate the use of structured criteria with ABA graphs, Normand (2003) developed a job aid designed to operationalize the steps involved in visual analysis. Job aids consisted of specific steps to guide the participant through graphic analysis (e.g., “Determine the mean [average] rate of behavior in baseline”); the active job aid required the participant to write down an answer to each step, and the passive job aid consisted of the same steps but without a prompt to record an answer. A total of 60 graphs were
selected from JABA, and Normand (2003) included contextual information about the participant, setting, behavior and intervention. Undergraduate participants were selected who had completed at least one course on single-subject research and passed a screening tool designed to test for a basic understanding of level, trend, and variability. Participants were randomly assigned to one of three groups – a control group that did not receive job aids or instruction, a passive job aid group, and an active job aid group. Each assessment consisted of 15 randomly-selected graphs about which the participant selected one of the following multiple-choice responses: a) the intervention caused an increase in behavior, b) the intervention caused a decrease in behavior, c) there was no change in behavior, d) there was a change that was not caused by the intervention or e) can’t make a decision based on the graph. As previously described, Normand (2003) asked three Ph.D.-level experts to judge the graphs to provide a basis for evaluating participant accuracy. The experts discussed and came to a consensus on the graphs they disagreed on.

In baseline, all participants scored graphs without job aids; agreement with the experts on the dichotomous scoring was around 70% for all three groups. In intervention, with the introduction of the job aids for the active and passive job aid groups and no job aid for the control group, the agreement of the all three groups decreased to around 50%. Accuracy further deteriorated in a fading phase, in which the job aid groups used the aids for every other graph. Returning to baseline, all three groups performed at about 50% accuracy. The author suggests several potential explanations for the results of the study. The level of disagreement among the experts raises questions about the validity of the criterion ratings of the graphs. The similarity of performance across groups suggests that neither intervention was successful, and the changes in performance in all groups’
performance after the training suggests that these changes may have been a result of
differential difficulty of graphs rather than the treatment. The job aids may not have
effectively controlled participants’ behavior because they lacked the conceptual
foundation for understanding the steps in the job aid. Systematic training procedures,
with the opportunity to practice using the job aid and receive feedback on specific steps,
may have increased the effectiveness of the job aids.

In a related study examining some of the questions generated by Normand (2003),
Colon (2006) compared the effectiveness, efficiency, and maintenance of training
individuals to use a structured visual approach (i.e., job aids) and a mechanical approach
(i.e., the CDC method; Fisher et al., 2003). A subset of the graphs used by Normand
(2003) were employed in this study – specifically, 36 graphs on which three Ph.D-level
experts agreed and eight graphs on which two of three experts agreed. Twelve
participants were randomly assigned to one of three groups: control, active job aid, and
CDC. A multiple baseline design was used to evaluate the effects of each intervention;
each intervention group (i.e., active job aid and CDC), comprised a set of multiple
baselines. In each assessment, participants evaluated four graphs using the same multiple-
choice format used by Normand (2003). In terms of dichotomous agreement with the
experts (i.e., change or no change), in baseline all participants averaged about 70%; only
one participant averaged below 50% agreement in baseline. In a staggered fashion,
participants in the active job aid and CDC groups received one-on-one training on their
respective methods. In the active job aid group, the experimenter explained the job aid
and provided feedback to participants as they answered each step of the job aid for each
of three sample graphs. In the CDC group, participants were provided with sample
graphs that contained the requisite criterion lines, instructions, and the binomial table specifying the number of data points required to be above or below the criterion lines to demonstrate an effect. As in the active job aid group, the experimenter explained the procedures and gave feedback as the participants used the CDC method to analyze the sample graphs. After training was complete, participants in these groups scored assessment packets using the job aid or CDC instructions. During this phase, the graphs in the assessment packets provided to the CDC group had the necessary criterion lines. All participants in the CDC group immediately improved their accuracy to near 100% following training and with the criterion lines; participants in the active job aid group also improved to an average of about 80% accuracy. The next phase was a return to baseline; participants in the active job aid group scored graphs without the job aid, and participants in the CDC group scored graphs without criterion lines and CDC instructions. The accuracy of participants in the active job aid group remained fairly stable above 80% in this phase, whereas the accuracy of the CDC group decreased to around 50%. The average training time for the active job aid participants was 28 minutes; for the CDC participants, training was only an average of 8 minutes in duration. Although training of the CDC method was more efficient and produced slightly higher accuracy than the active job aid in the intervention phase, the active job aid group maintained their performance in the absence of the supports while the CDC participants’ performance deteriorated.

Jostad (2011) continued investigating the effectiveness of job aids in a two-part experiment. In Study 1, the effects of training and a job aid were evaluated in a single-subject design; in Study 2, the job-aid and training were compared to traditional lecture
using a group design. A total of 500 AB graphs were produced, each representing one of
10 different combinations of data characteristics (e.g., flat slope in baseline and
increasing trend in treatment). Half of the graphs displayed effects and half displayed
noneffects. Each assessment packet contained 10 graphs, one of each type, and
participants judged whether the graphs depicted an effect or not. Unlike the previous
studies, Jostad (2011) did include slope change as a feature of some of the graphs. To
determine whether a particular graph depicted an effect, Jostad (2011) consulted the
programmed parameters and applied the CDC method to each graph.

The job aid was developed by consulting common recommendations for visual
analysis found in textbooks on single-subject research. Along with instructions to
evaluate level, trend, and variability, it contained graphs to illustrate particular steps (e.g.,
trend estimation), exceptions to be noted (e.g., if baseline and treatment have an
increasing trend, a level change does not indicate an effect), and a step combining all of
the information to make a final decision.

Seven undergraduates participated in Study 1. In baseline, they evaluated graphs
without instruction or feedback; accuracy was generally between 40-60%. With the
introduction of the job aid, three participants demonstrated a gain in accuracy to about
70-80%, while four participants remained around baseline levels. The researcher taught
participants to use the job aid with the assistance of a PowerPoint presentation illustrating
the use of the aid with sample graphs; however, feedback was not provided during this
instruction. After learning how to use the job aid, four participants improved to about
90% accuracy; attempts to fade the job aid for two of these participants resulted in
deteriorated performance. Two participants did not improve after receiving instruction
on the job aid. The author analyzed the errors that were being made and found that both participants were incorrectly estimating trend in the data paths; she provided individualized instruction on this skill. However, neither participant improved following this instruction; in a subsequent phase, both were given immediate, detailed feedback following each graph judgment. With these modifications, one participant judged graphs with an average of about 85% accuracy; the other remained at around 65% accuracy.

Forty undergraduates participated in Study 2, which served as an evaluation of the job aid plus training package in a larger group setting. The graphs, job aid, and computer presentation from Study 1 were used in Study 2; the researcher created a videotaped lecture on visual analysis (similar to Stewart et al., 2007) for the comparison condition. This lecture included conceptual instruction on visual analysis but also taught participants to draw split-middle lines to estimate slope; therefore, we categorize this approach as computational. Participants were randomly assigned to the job aid group or the lecture group. Prior to intervention, both groups completed two assessment packets containing a total of 20 graphs. The job aid plus training and the lecture were both delivered in a group setting. Following intervention, both groups completed two additional assessment packets. Those who did not score at or above 80% accuracy on posttest 1 then received the other intervention (e.g., participants in the job aid group who did not meet criterion watched the lecture) and completed posttest 2. Participants who still did not meet criterion received feedback as in Study 1.

The job aid group performed slightly better on the pretest than the lecture group (65% vs. 61%), but this difference was not statistically significant. The accuracy of both groups improved significantly from pre- to posttest; mean accuracy was 74% and 73%
for the job aid and lecture groups, respectively. Although the between-group difference
was not statistically significant, 55% of participants in the job aid group met the 80%
accuracy criterion on posttest 1 compared with 35% in the lecture group. Following
exposure to both interventions, only 12 participants failed to meet criterion on posttest 2;
after receiving feedback, all but 1 were judging graphs with at least 80% accuracy.

The results of Colon (2006) and Jostad (2011) suggest that instruction, practice,
and feedback on the use of a job aid may be an effective method for quickly teaching
some visual analysis skills. The importance of instruction, practice, and feedback is
highlighted by the results of Colon (2006) and Jostad (2011) compared with Normand
(2003), who did not deliver instruction on using the job aid. However, there are several
concerns with relying on these highly structured approaches to visual analysis. Currently,
it seems as though the lack of consensus and empirical data on the individual factors that
play a part in visual analysis and how those factors combine to influence visual analysis
would present a challenge to producing a valid job aid. The job aid might be validated by
comparing the results to expert opinion: Colon (2006) used expert opinion as the criterion
for participant accuracy, so we might extrapolate that the results of the job aid generally
agreed with expert opinion most of the time. In Jostad (2011), it was assumed that the
aids produce accurate judgments when used correctly by novices and judged against
programmed parameters and the CDC method; it is unknown whether experts would
agree with the decisions produced by the job aid. Finally, the structured nature of the job
aid is designed to produce inflexible analysis; for example, directing an individual to
estimate the overall trend in a phase does not allow the analyst to look more closely at
more subtle considerations like an ascending or descending trend at the end of the
baseline phase. An experienced visual analyst’s decision about a graph would likely be influenced by this data pattern, but a job aid focused only on the overall patterns of a phase would be insensitive to it.

The results of these studies do suggest that novice visual analysts can be trained to follow structured criteria and accurately determine whether an intervention effect is present (cf. Normand, 2003), but there are several important limitations of these studies related to a) the graphs that were used to assess the effectiveness of training and b) the approaches that participants were taught to use to evaluate graphs. In terms of the assessments, the graphs used by Fisher et al. (2003) and Stewart et al. (2007) warrant concerns. The equation used by these authors did not include programmed slope changes, resulting in graphs that did not have this important feature. Slope changes are common in actual single-subject data and previous research suggests that these changes in slope present challenges for novice and experienced visual analysts alike (Gibson & Ottenbacher, 1988; Normand & Bailey, 2006). Two of the three studies investigating the DC and CDC methods were conducted using graphs without slope; this raises questions about the generality of these approaches.

In addition, the use of programmed parameters (Fisher et al., 2003; Jostad, 2011; Stewart et al., 2007) and the DC/CDC methods (Jostad, 2011) as the criterion for accuracy may be problematic. Although mathematical parameters provide a concrete cut-off point for determining the presence of an effect (i.e., the parameter does not equal zero), very small changes may not be detected visually, or may be negated by other data patterns (e.g., a therapeutic trend at the end of baseline). For example, some very small programmed changes ($d=0.5$) were accurately detected by DC and CDC methods in
Fisher et al. (2003) and Stewart et al. (2007), but changes of this magnitude may not be considered valid within the tradition of visual analysis. In other words, without visual aids and structured criteria, experts may not agree that a change with $d=0.5$ constitutes an intervention effect. In Jostad (2011), a sample graph was included with an overall flat slope in baseline but an increasing trend at the end of baseline; this was said to depict a “true effect” because the level change was not predicted by the overall baseline trend. However, experts might disagree that this graph suggests an effect of the intervention due to the slope of the last three baseline data points. As an alternative to using mathematical parameters, researchers could rely on expert opinion as a potentially more valid accuracy criterion. However, those that have attempted use expert opinion as the accuracy criterion (e.g., Colon, 2006; Normand, 2003) were not able to obtain consensus among experts on a large number of graphs.

Finally, Colon (2006), Fisher et al. (2003), and Stewart et al. (2007) did not include criterion lines in the baseline assessment graphs. Although studies by Harbst et al. (1991) and Normand and Bailey (2006) suggest that the mere provision of trend lines does not improve accuracy or agreement of visual analysis, the criterion lines alone could have influenced participants’ accuracy in these studies.

The approaches that participants were taught to use to analyze graphs also raise concerns. Colon (2006) and Stewart et al. (2007) described some of the limitations related to the practicality of using mechanical approaches and criterion lines to guide visual analysis. Criterion lines are not typically printed on published graphs, nor did the authors teach participants themselves to draw criterion lines. The split-middle line and baseline mean line (for the DC method) are relatively simple to draw; adjusting those
lines by 0.25 standard deviation for the CDC method requires more involved mathematical computation and is likely cumbersome to conduct by hand. Although the authors demonstrate that, when provided criterion lines and explicit rules, participants can accurately judge graphs, the functional relevance of this skill remains unclear.

The majority of the reviewed training studies have generally focused on producing mechanical determinations of intervention effects; only Hojem and Ottenbacher (1988), Colon (2006), and Jostad (2011) provided instruction on the types of changes that might evidence an intervention effect (i.e., level, trend, and variability). This conceptual background may be necessary for accurate interpretation of the results of single-subject studies, and allows individuals not only to identify the presence or absence of an effect, but also to describe the data patterns that support that contention. A sufficient conceptual understanding may also allow the individual to analyze data more flexibly, taking into account nuances and subtleties (e.g., an overall flat slope but increasing slope of the last four baseline data points) of the data that may not be accommodated by the computational, mechanical, or structured visual approaches. An approach that teaches individuals the basic concepts of visual analysis and how to make visual discriminations, rather than providing explicit rules or instruction on computing criterion lines, may be an appropriate compromise. This visual discrimination approach may be more adaptable than computational, mechanical, or structured visual approaches, but also more systematic than the global visual approach.
Purpose Statement and Research Questions

The proposed study was designed to address the limitations of the literature on training individuals to visually analyze graphed single-subject data. Graphs were generated with an autoregressive model (Fisher et al., 2003; Matyas & Greenwood, 1990). Unlike previous studies (c.f., Jostad, 2011), slope change was included in the model so that this important feature of single-subject data was captured in the assessment items. In addition, previous research using algorithmically generated graphs used either programmed or obtained parameters of the graphs as “correct answer” and the criterion for participant accuracy, which raise concerns about validity. Instead, we used expert opinion as the standard against which participant responses were compared. No training study to date had evaluated determinations of slope and level change separately; all have asked participants more globally whether the graph demonstrates an effect or not. To analyze these separate but related judgments, we asked participants to make a judgment about slope and a judgment about level for each graph.

With regard to the training, previous research in this area has relied heavily on the use of criterion lines, which may not be practical for real-world applications and may not provide the conceptual background for accurate interpretation of results. Two trainings were developed and evaluated in the present study, both teaching the same, guided visual approach to analyzing the graphs. In both trainings, the same guidelines were provided for how to estimate slope and level and compare these features across two phases. The computer-based training (CBT) focused on providing participants with numerous opportunities to practice making visual discriminations about the presence or absence of level change and slope change across AB graphs. This training allowed participants to
move through the content at their own pace as they met mastery criteria both within and at the end of each module. Participants were randomly assigned to a) the computer-based training (CBT), b) a traditional lecture that covered the concepts of visual analysis but did not provide mandatory practice opportunities, or c) a control condition in which participants received no intervention.

**Primary Research Questions**

1. What is the relative effectiveness of the CBT on visual analysis, compared with a traditional lecture and no intervention, for teaching undergraduate to accurately identify changes in generated AB graphs of acquisition and reduction targets?

2. What is the relative effectiveness of the CBT on visual analysis, compared with a traditional lecture and no intervention, for teaching undergraduates to accurately identify changes in level in generated AB graphs of acquisition and reduction targets?

3. What is the relative effectiveness of the CBT on visual analysis, compared with a traditional lecture and no intervention, for teaching undergraduates to accurately identify changes in slope in generated AB graphs?

4. Does the CBT or traditional lecture result in a greater proportion of students meeting a 80% accuracy criterion with regard to level change, slope change, and/or overall accuracy?
Secondary Research Questions

1. How effective and efficient is each module of the CBT in terms of (a) errorless teaching, (b) producing gains on in-module pre/post tests, (c) producing gains in minimal training time?

2. How effective is each instructional segment in the CBT in terms of errorless teaching?

3. How effective is the CBT and traditional lecture on assessment items with various characteristics (e.g., those with trend in baseline and level change across phases)?
CHAPTER III

METHODS

Participants

One hundred and twenty three undergraduates without previous experience in visual analysis were recruited from Special Education 4000 and Special Education 5530. Consistent with the makeup of students in most special education courses at Utah State University, the majority of participants were female (88%). Across all groups, the mean age was 21.67 years ($SD=3.62$), the mean grade point average (GPA) was 3.54 ($SD=0.35$), and the mean number of years of completed undergraduate coursework was 2.14 ($SD=1.03$). Most participants (79%) were education majors (elementary, secondary, or special education), and another 14% were speech-language pathology majors. Table 3 provides descriptive statistics for these demographic variables for each group. The means and standard deviations of each variable are fairly similar across groups, and analyses of variance (ANOVAs) of age, GPA, and years of undergraduate coursework did not reach statistical significance (see Table 4). This indicates that any differences among the three groups with respect to these demographic variables were likely due to chance.

Setting

Experimental sessions took place in computer labs in the library or the education building on the Utah State University campus.
### Table 3

**Descriptive Statistics for Demographic Variables**

<table>
<thead>
<tr>
<th><strong>Group</strong></th>
<th><strong>n</strong></th>
<th><strong>Age</strong></th>
<th><strong>GPA</strong></th>
<th><strong>Years of Coursework</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>41</td>
<td>21.12 (2.96)</td>
<td>3.57 (0.33)</td>
<td>2.22 (1.07)</td>
</tr>
<tr>
<td>Lecture</td>
<td>41</td>
<td>21.85 (3.85)</td>
<td>3.59 (0.35)</td>
<td>2.18 (1.12)</td>
</tr>
<tr>
<td>CBT</td>
<td>41</td>
<td>22.05 (3.98)</td>
<td>3.44 (0.34)</td>
<td>2.02 (0.89)</td>
</tr>
</tbody>
</table>

### Table 4

**ANOVA of Demographic Variables**

<table>
<thead>
<tr>
<th><strong>Variable</strong></th>
<th><strong>F</strong></th>
<th><strong>df</strong></th>
<th><strong>p</strong></th>
<th><strong>η²</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>1.02</td>
<td>2, 120</td>
<td>.37</td>
<td>.017</td>
</tr>
<tr>
<td>GPA</td>
<td>2.38</td>
<td>2, 120</td>
<td>.10</td>
<td>.038</td>
</tr>
<tr>
<td>Years of Coursework</td>
<td>0.56</td>
<td>2, 120</td>
<td>.58</td>
<td>.009</td>
</tr>
</tbody>
</table>

### Materials

#### Graphs

Multiple AB graphs depicting acquisition targets and reduction targets, were generated in Microsoft Excel® using a linear equation. These assessment graphs were developed systematically through a multi-step process to yield graphs that present a challenge for novice visual analysts, but are agreed upon by experts. First, numerous graphs were generated using the equation $y_i = B_0 + B_1L + B_2S + (1-a)E_i + aE_{i-1}$, where $B_0$ is the baseline mean, $B_1$ is the immediate level change parameter, $L$ is a dummy variable coded 0 for baseline and 1 for treatment, $B_2$ is the slope change parameter, $S$ is a dummy variable coded 0 for baseline and n-1 for treatment (where n is the day of treatment), $a$ is the autocorrelation parameter, and $E_i$ is the random error at time $we$ (generated using the random number function in Excel®). All graphs had 20 total data points (10 in baseline...
and 10 in treatment), a baseline mean of 15, an autocorrelation parameter of 0.20, and random error produced from a normally distributed set of numbers with a mean of 0 and a standard deviation of 4.0 (i.e., random error term could be positive or negative). For acquisition graphs, the level change parameter was set at 0, 4, 8 and 10 and the slope parameter at 0, 0.5, 0.75, 1, and 1.5; for reduction graphs, the same absolute values were used with negative changes programmed. Various combinations of these parameters were programmed to produce numerous graphs. The obtained parameters differed from the programmed parameters because of the random error; therefore, we calculated the obtained effect size (treatment mean-baseline mean/pooled standard deviation) and obtained standardized slope change (treatment slope-baseline slope/pooled standard error of the slope) for each graph.

To identify specific graphs (and features of graphs) that experts agree upon but are challenging for individuals without formal training in visual analysis, we used the five-step process depicted in Figure 1. First, we field-tested 80 graphs (40 acquisition and 40 reduction) with 51 undergraduates who had not received training in visual analysis; approximately 20-25 students evaluated the slope and level change in each graph. The effect sizes in these graphs ranged from 0 to 2.75 and the slope changes ranged from 0 to 0.70. The results of this field-test were used to narrow down the mathematical parameters and data patterns that were challenging for novice visual analysts so that additional graphs could be generated: effect sizes around 1 and -1 produced divided responding (i.e., about 50% of the undergraduates said “yes” and 50% said “no” for graphs with effect sizes of this magnitude) and slope changes around 0.30 produced divided responding. Second, we used these parameters to inform subsequent
graph generation to develop assessment items that were sufficiently challenging to require instruction. We wanted to produce graphs with all possible combinations of changes for level and slope – this resulted in four types of graphs: (1) slope and level changes, (2) a slope change with no level change, (3) a level change with no slope change, and (4) no slope or level change (see Figure 2). Using the autoregressive equation and the target parameters from the field test, we produced approximately 60 additional acquisition and reduction graphs for each of the above graph types.

Third, the first and second authors independently inspected all additional generated graphs and answered “yes,” “no,” or “maybe” about whether each graph depicted (1) a change in level that was attributable to the intervention and (2) a change in slope that was attributable to the intervention. We kept the graphs on which the two experts agreed on both dimensions (i.e., regarding whether there was a change in level and a change in slope). We manipulated graphs that had one or more “maybe” responses and those that resulted in disagreements between the two experts to make the changes more obvious by increasing the magnitude of the slope and/or level change. The first and second author then rescored the manipulated graphs, and we compared our answers. Again, only graphs with independent expert consensus on both dimensions of the graph (based on independent ratings) were retained. This resulted in approximately 240 total graphs (120 acquisition and 120 reduction).
Figure 1. Process for generating and evaluating graphs for inclusion in assessment.
Fourth, three other experts evaluated the graphs to identify graphs with high levels of agreement. One expert judged all of the graphs, and the other two experts each judged half of the graphs. Graphs with 100% expert agreement on both dimensions, and some with 75% agreement on both dimensions, were identified and used in the next step of the development process.

The five experts who evaluated the graphs had a range of training backgrounds and research interests, although all were currently affiliated with the Department of Special Education and Rehabilitation at Utah State University. Two of the experts were faculty with doctoral degrees in Special Education or Psychology, and three were doctoral students in the Applied Behavior Analysis specialization of the Disability Disciplines doctoral program.

<table>
<thead>
<tr>
<th>Level Change</th>
<th>Slope Change</th>
<th>Reduction Graphs</th>
<th>Level Change</th>
<th>Slope Change</th>
<th>Acquisition Graphs</th>
</tr>
</thead>
<tbody>
<tr>
<td>YES</td>
<td>YES</td>
<td>Slope change</td>
<td>Level change</td>
<td>YES</td>
<td>Slope change</td>
</tr>
<tr>
<td></td>
<td>NO</td>
<td>No slope change</td>
<td>Level change</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NO</td>
<td>YES</td>
<td>Slope change</td>
<td>No level change</td>
<td>NO</td>
<td>No slope change</td>
</tr>
<tr>
<td></td>
<td>NO</td>
<td>No slope change</td>
<td>No level change</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Figure 2.* The eight types of graphs included in the assessment.
Four of the experts (including all doctoral student experts) are Board Certified Behavior Analysts. Broadly, experts’ research interests range from the assessment and treatment of severe problem behavior to the acquisition of language and reading skills. This relatively wide range of training backgrounds, professional experience, and research interests suggests that the experts may have been fairly representative of the population who uses single-subject research designs.

Fifth, we conducted another field-test with undergraduates to evaluate whether the effects and noneffects in the graphs with high expert agreement were also very salient to untrained raters. This field test included only graphs that were thought to be relatively difficult given the results of the first field-test. Fifteen graphs with high expert agreement were selected from each graph type for both acquisition and reduction targets, for a total of 120 graphs (8 types * 15 graphs per type). Each graph was judged by 16 to 17 undergraduates, and the results were summarized and compared to expert consensus.

<table>
<thead>
<tr>
<th>Correct Answer</th>
<th>Level Change</th>
<th>Slope Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>YES</td>
<td>n=10</td>
<td>n=10</td>
</tr>
<tr>
<td></td>
<td>M=69%</td>
<td>M=54%</td>
</tr>
<tr>
<td></td>
<td>Min=44%</td>
<td>Min=29%</td>
</tr>
<tr>
<td></td>
<td>Max=88%</td>
<td>Max=71%</td>
</tr>
<tr>
<td>NO</td>
<td>n=10</td>
<td>n=10</td>
</tr>
<tr>
<td></td>
<td>M=66%</td>
<td>M=57%</td>
</tr>
<tr>
<td></td>
<td>Min=53%</td>
<td>Min=31%</td>
</tr>
<tr>
<td></td>
<td>Max=81%</td>
<td>Max=75%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Correct Answer</th>
<th>Level Change</th>
<th>Slope Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>YES</td>
<td>n=10</td>
<td>n=10</td>
</tr>
<tr>
<td></td>
<td>M=49%</td>
<td>M=33%</td>
</tr>
<tr>
<td></td>
<td>Min=41%</td>
<td>Min=18%</td>
</tr>
<tr>
<td></td>
<td>Max=63%</td>
<td>Max=44%</td>
</tr>
<tr>
<td>NO</td>
<td>n=10</td>
<td>n=10</td>
</tr>
<tr>
<td></td>
<td>M=67%</td>
<td>M=72%</td>
</tr>
<tr>
<td></td>
<td>Min=53%</td>
<td>Min=50%</td>
</tr>
<tr>
<td></td>
<td>Max=75%</td>
<td>Max=88%</td>
</tr>
</tbody>
</table>

*Figure 3.* Mean, minimum, and maximum of percent of undergraduates responding correctly about the level and slope changes of the graphs in the assessment packets.
We then selected 5 graphs from each of the eight graph types with (a) 100% expert consensus and (b) low undergraduate accuracy (compared to expert opinion) to comprise the assessment packet. Figure 3 displays the average percent of undergraduates responding accurately about slope and level changes for the 20 acquisition and 20 reduction graphs that are included in the assessment; Figure 4 shows the mean and standard deviation of the effect sizes and standardized slope changes for each type of graph in the assessment.

**Computer-Based Training**

The CBT was developed using Adobe® Captivate®, an e-learning software program that enables the programmer to record audio to narrate presentations, upload videos, embed quizzes within each module, and include opportunities for the user to respond. The training was designed to provide instruction on evaluating the level and trend changes across AB comparisons and consisted of four modules: Introduction to Single-Subject Research, Level Change, Slope Change, and Level and Slope Change.

<table>
<thead>
<tr>
<th>Acquisition Graphs</th>
<th>Reduction Graphs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Level Change</strong></td>
<td><strong>Slope Change</strong></td>
</tr>
<tr>
<td><strong>YES</strong></td>
<td>Effect Size</td>
</tr>
<tr>
<td></td>
<td>SD=0</td>
</tr>
<tr>
<td></td>
<td>Slope Change</td>
</tr>
<tr>
<td></td>
<td>SD=0.04</td>
</tr>
<tr>
<td><strong>NO</strong></td>
<td>Effect Size</td>
</tr>
<tr>
<td></td>
<td>SD=0</td>
</tr>
<tr>
<td></td>
<td>Slope Change</td>
</tr>
<tr>
<td></td>
<td>SD=0.05</td>
</tr>
</tbody>
</table>

*Figure 4. Descriptive statistics for the parameters of the assessment graphs.*
Each module contained (a) a pretest evaluating the specific skill targeted in the module prior to instruction, (b) instruction featuring brief explanations of the effect type and examples and non-examples of that effect type, (c) opportunities to practice identifying examples and non-examples of the effect type with feedback, (d) remedial loops providing extra practice and feedback, and (e) a posttest evaluating acquisition of the skill.

Within a module, the pre- and posttest contained the same 10 graphs in the same sequence. These graphs and related questions addressed only the skill targeted within that module. For example, the in-module pre/posttest for Level Change consisted of 10 graphs, all with flat slopes in baseline and treatment because the Level Change module does not contain instruction on slope. Graphs were generated using the same general autoregressive equation used to produce the assessment graphs, but were created specifically for the in-module pre- and posttest.

The content for the instructional slides within the module (i.e., those teaching the participant about a skill through modeling that skill) was based partly on the chapters of textbooks on Applied Behavior Analysis describing visual analysis (e.g., Cooper et al., 2007). Opportunities for participants to practice identifying examples and nonexamples become progressively more difficult throughout each module (e.g., a graph with low variability and no overlap to a graph with variability and overlap). *Instructional segments* are units of practice that differ in the type of question asked (e.g., “Enter the level” vs. “Is there a change in level?”) or the characteristics of the graphs (e.g., large change with low variability vs. small change with high variability). Participants moved through each module at their own pace; when a participant accurately answered questions
about a discrimination in one instructional segment, she moved to the next instructional segment containing more difficult discriminations. All responses, correct and incorrect, were followed by feedback in the form of reinforcement (e.g., “That’s right.”) or correction (e.g., “The level of the treatment phase is higher than the level of the baseline phase. Here’s the approximate level of both phases. Let’s try another one.”). When a participant responded incorrectly, a remediation loop began that provided extra practice at the current discrimination level. Figures 5, 6, and 7 provide an overview of the instructional design of the modules.

Participants were required to score 90% or better on the posttest to complete the module and move to the next module. When participants did not meet the 90% criterion on the posttest, they were required to complete additional practice before attempting the posttest a second time. If participants did not meet criterion on their second attempt, they progressed to the next module.

**Traditional Lecture**

The traditional lecture intervention consisted of reading, a lecture, and self-study materials. We selected sections of a chapter on visual analysis from a widely used textbook on applied behavior analysis (Cooper et al., 2007). The lecture was created using Sony® Camtasia® and distributed to participants via CD-ROM. The lecture consisted of the instructional slides and corresponding script from the CBT (i.e., slides that describe or model a skill but do not require an active participant response) and was 36 min in duration. The self-study materials were 20 printed practice slides from the last module of the CBT (i.e., the most complex graphs with a question about slope and a question about level). These materials were printed so that the corresponding feedback
slide from the CBT, with the correct answers and slope and level lines for each phase, was on the back. Participants could use these flashcards to apply what they learned in the reading and lecture.

By equating the conditions on instructional content, we isolated the following set of variables that distinguish traditional lecture from CBT: (1) number and type of practice opportunities including feedback and remediation, (2) self-paced and individualized quality of CBT, (3) other features inherent in use of the computer and CBT.

**Dependent Measures**

**Primary Dependent Measure**

The number of correct judgments about effects and noneffects (using expert consensus as the criterion for accuracy) was the primary measure of the effectiveness of the CBT and traditional lecture. The assessment contained the 40 AB graphs described above (Materials section), with 20 depicting acquisition targets and 20 depicting reduction targets. Within each subset (i.e., the acquisition subset and the reduction subset), five graphs from each of the eight cells in Figure 2 were included for a total of 20 graphs. The graphs within each subset were presented in a random sequence and were accompanied by a hypothetical scenario to provide context for judging the intervention effects. We provided a basic definition of level and slope, and participants responded to two questions about each graph by circling “yes” or “no”:

1. Does the graph suggest that the treatment caused an improvement in the **LEVEL** of the behavior?   
   Yes   No
2. Does the graph suggest that the treatment caused an improvement in the *slope* of the behavior?  Yes  No

We analyzed overall accuracy (i.e., number of correct responses out of total number of response opportunities), level accuracy (i.e., number of correct responses to the question about level out of total number of level response opportunities), and slope accuracy (i.e., number of correct responses to the question about slope out of total number of slope response opportunities). The same graphs were used for the pretest and posttest, allowing us to evaluate changes in performance on specific items. Graphs were presented in the same order on the pretest and posttest.

**Secondary Dependent Measures**

To evaluate the effectiveness and efficiency of the CBT, we collected data on several secondary dependent variables. The Adobe Captivate® software collects data on every user response (both in-module pre/posttests and practice opportunities). These data were captured and analyzed to address the secondary research questions. Specifically, we analyzed data on (1) the number of correct responses on the in-module pretest, (2) the number of correct responses on the in-module posttest, (3) number of attempts to meet criterion on the in-module posttest, (4) number of errors on practice opportunities to meet criterion to advance to the in-module posttest, and (5) number of errors on practice opportunities within each instructional segment. The duration of each participant’s training time for each module was recorded.
Instructional design for the Level Change module. Arrows indicate participant movement through the module; at the Practice level, if one question is answered incorrectly, participants are automatically routed to the Remediation slides. The numbers in each box represent the number of graphs contained in that instructional segment.

**Figure 5.**
**Figure 6.** Instructional design for the Slope Change module.

<table>
<thead>
<tr>
<th>Skill Practiced Across</th>
<th>Estimating Slope</th>
<th>Discriminating Slope Change</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Baseline Direction (increasing, decreasing, flat)</td>
</tr>
<tr>
<td>Direction (increasing, decreasing, flat)</td>
<td></td>
<td>Treatment Direction (increasing, decreasing, flat)</td>
</tr>
<tr>
<td>Magnitude (steep, shallow)</td>
<td></td>
<td>Magnitude (steep, shallow)</td>
</tr>
<tr>
<td>Variability (low, medium, high)</td>
<td></td>
<td>Variability (medium, high)</td>
</tr>
<tr>
<td>Presence of Immediate Level Change</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 7. Instructional design for the Slope Change and Level Change Module.
Interobserver Agreement

A second, independent rater scored 41 randomly-selected pretests (33%) and 41 randomly-selected posttests (33%). Percent agreement was calculated on a point-by-point basis; the number of agreements were divided by the number of agreements plus disagreements. The two raters disagreed on two items out of 6,560 total items, for an agreement level of 99.9%.

Experimental Design

A randomized three-group comparison design was used to evaluate the visual analysis performance of a) participants trained via traditional lecture, b) participants trained via CBT, and b) participants who receive no training. Participants were randomly assigned to one of the three groups prior to taking the pretest.

Procedures

Pretest

Participants were given a paper copy of the assessment packet containing the 40 AB graphs and asked to answer both questions following each graph by circling “yes” or “no.” The researcher did not answer any questions or provide any feedback while administering the pretest. Immediately following the pretest, participants completed the training associated with their assigned group.
Traditional Lecture

The duration of time between pre- and posttest for participants in the traditional lecture group was matched to the mean duration of the first ten participants in the CBT (105 min). We recorded the clock time that each participant finished the pretest, added 105 min, and recorded the clock time that she was due to take the posttest.

Participants in the traditional lecture group were first asked to complete the reading on visual analysis. We explained that they could write on and/or highlight the reading if they wished to do so, and that we could answer any questions they had during their training. When they finished the reading, the experimenter answered any questions and gave the participant the lecture CD and self-study materials. We explained that the self-study materials could be used to practice what they learned in the reading and lecture, and that they could write on them if they wished to do so. In addition, at this point each participant was told how much time she had before her posttest (e.g., if she completed the reading in 30 min, she had 75 min to view the lecture and use the self-study materials), and was asked to view the lecture on a computer in a specified area of the lab. This step was taken to ensure that there was no crossover between groups. Participants were permitted to pause and rewind the lecture as they wished. Immediately prior to taking the posttest, we again asked participants in this group if they had any questions.

Computer-Based Training

Participants in the CBT group were given instructions for how to access and log in to the training, and directed to a specific area of the computer lab to complete their
training. Given the mastery-based nature of the training, the duration varied across participants.

**Control Group**

Participants in the control group did not receive any instruction on visual analysis between taking the pre- and posttests. Like the lecture group, the duration of time between pre- and posttest for participants in the control group was matched to the mean duration of the first ten participants in the CBT (105 min). Participants in the control group were required to remain in a specific area of the computer lab during the time between pre- and posttest, but were allowed to use computers, complete homework, and/or engage in leisure activities.

**Posttest**

Participants in the traditional lecture and control groups took their posttest 105 min after completing their pretest; participants in the CBT group took their posttest immediately after completing the training or 150 min after completing the pretest if they had not completed the training at this time. There were four participants (10%) in the CBT group who, after 150 min in the CBT, had not completed it. In these two cases, we terminated the training early so that participants had time to take the posttest.

Regardless of condition, participants were given a paper copy of the assessment packet containing the 40 AB graphs and asked to answer both questions following each graph by circling “yes” or “no.” The researcher did not answer any questions or provide any feedback while administering the posttest.
Treatment Acceptability

Following completion of the posttest, we gave participants in the intervention groups (i.e., the lecture and the CBT) a brief questionnaire to complete about the training (see Appendix A).

Treatment Integrity

Treatment integrity for participants in the traditional lecture group was measured by recording the duration of time participants spent reading the assigned chapter and viewing the video. Individual reading times were expected to vary, but based on a small sample of graduate students reading the assigned chapter, we estimated that 20 min was the minimum amount of time one could spend reading for comprehension. The reading time of only one participant (2%) was briefer than 20 min. The videotaped lecture was 36 min; all participants viewed the lecture for at least 36 min (some participants may have paused the lecture to take notes and/or use the provided self-study materials).

Treatment integrity for participants in the CBT group was measured by obtaining data from the questions requiring responses throughout the modules. The training was designed such that participants have to first complete the pretest, then accurately respond to numerous questions throughout the module to proceed to the posttest. Therefore, data captured by the program on participant responses to practice opportunities within the instructional module served as a measure of fidelity. Based on this measure of treatment integrity, 5 out of 41 participants (12%) in the CBT group did not complete the training as designed (this includes the 2 participants whose training was terminated prematurely by the experimenter).
CHAPTER IV

RESULTS

Missing Scores

A small number of participants did not answer some questions on the pretest and/or posttest; each experimental group had at least one participant who omitted questions on one of the assessments. Only a very small proportion (0.07%) of the total number of questions across all participants and both assessments were omitted. We imputed each missing score using a multi-step process that accounted for group membership, performance on other items in the assessment, and item difficulty. We calculated a modified score for all participants in the group by subtracting the missing item. For example, if a participant in the control group did not answer #18 in the set of acquisition graphs on the pretest, we obtained a modified score for all participants in the control group by subtracting their score on #18 on the pretest. For the participant who omitted the item, the modified score was the same as their total score. We then created a scatterplot depicting the relationship of participants’ modified score to their score on the omitted item and used SPSS to generate a regression equation of this relationship (y=mx+b, where y is the score on the missing item, m is the slope of the regression line, x is the modified score, and b is the y-intercept). We entered the modified score of the participant who omitted the item into the regression equation for their group to obtain a predicted score on that item for that participant. The predicted score was always a fraction between 0 and 1, and was entered into the participant’s datasheet exactly as obtained by the regression equation.
Relative Effectiveness of the Interventions

We conducted various analyses of each group’s performance on the pretest, posttest, and change from pretest to posttest with regard to (a) overall accuracy, (b) accuracy on acquisition graphs, (c) accuracy on reduction graphs, (d) accuracy on slope questions, and (e) accuracy on level questions. These analyses included computing descriptive statistics and effect sizes, examining distributions of scores using box-and-whisker plots, and conducting analyses of variance (ANOVAs) to evaluate the effect of the interventions on each dependent variable.

Overall Accuracy

Pretest. Table 5 summarizes the means and standard deviations of each group’s overall performance on the pretest. Although the CBT group’s mean pretest score is somewhat lower than the lecture group ($d = -0.33$) and control group ($d = -0.43$), an ANOVA of the overall pretest scores did not reveal statistically significant differences between the three groups. (Table 6 presents the results of three-group ANOVAs, pairwise comparisons, and effect sizes for the overall assessment.) The box-and-whisker plot of pretest scores in Figure 8 clearly shows that the distributions are overlapped, but potentially important differences are also apparent. First, the interquartile range and the total range of the control group are both larger than these measures of dispersion for the other groups; in contrast, the pretest scores in the lecture group were less variant, as evidenced by a smaller interquartile range and a smaller total range. In other words, the participants in the lecture group had more similar pretest scores than participants in the other groups. Second, the entire distribution of pretest scores for the CBT group is
shifted downward relative to the other groups, evidenced by a lower median score and a lower box. The plot suggests that the mean CBT pretest score was not excessively affected by outliers, and is fairly representative of the entire group of scores.

**Posttest.** In terms of overall accuracy on the posttest, the means of the CBT and lecture groups were very similar \((d = -0.24)\) and substantially higher than the mean of the control group \((d = 1.03 \text{ and } 1.43 \text{ respectively})\).

To evaluate the statistical significance of these differences, we planned to conduct ANCOVAs using pretest scores as the covariate. One assumption of ANCOVA is the homogeneity of slopes between the covariate and dependent variable across groups; that is, it is assumed that there is not a significant interaction between an individual’s pretest score and group membership.

Table 5

<table>
<thead>
<tr>
<th>Group</th>
<th>Pretest (Mean, SD)</th>
<th>Posttest (Mean, SD)</th>
<th>Change (Mean, SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>51.20 (9.14)</td>
<td>52.29 (8.90)</td>
<td>1.10 (6.78)</td>
</tr>
<tr>
<td>Lecture</td>
<td>50.02 (7.39)</td>
<td>63.05 (6.14)</td>
<td>13.02 (7.60)</td>
</tr>
<tr>
<td>CBT</td>
<td>47.34 (8.73)</td>
<td>61.27 (8.54)</td>
<td>13.93 (8.37)</td>
</tr>
</tbody>
</table>

Table 6

**ANOVAs/ANCOVAs of the Overall Assessment**

<table>
<thead>
<tr>
<th>Measure</th>
<th>THREE-GROUP COMPARISON</th>
<th>LECTURE vs. CONTROL</th>
<th>CBT vs. CONTROL</th>
<th>CBT vs. LECTURE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(F)</td>
<td>(df)</td>
<td>(p)</td>
<td>(\eta^2)</td>
</tr>
<tr>
<td>Pretest</td>
<td>(2.24)</td>
<td>2, 120</td>
<td>.11</td>
<td>.04</td>
</tr>
<tr>
<td>Posttest</td>
<td>(39.27)</td>
<td>2, 119</td>
<td>&lt;.001</td>
<td>.40</td>
</tr>
<tr>
<td>Change</td>
<td>(39.27)</td>
<td>2, 119</td>
<td>&lt;.001</td>
<td>.40</td>
</tr>
</tbody>
</table>

\(^a\)ANOVA. \(^b\)ANCOVA with pretest as covariate.
To test this assumption, we visually analyzed plots of the relationship of between pretest score and change score as a function of group membership, and statistically tested the interaction between pretest score and group. When the interaction between pretest score and group was significant (i.e., when the homogeneity of slopes assumption was violated), we conducted tests of simple main effects. These tests evaluate the significance of differences in mean change scores between groups at low, medium, and high levels of pretest scores. The three levels of pretest scores necessitated a \( p \) value of .017 for statistical significance. Any significant differences were further evaluated with pairwise comparisons at \( p=.017 \). When there was no significant interaction between pretest score and change score, we conducted the ANCOVAs as planned.

Figure 9 shows the means and 95% confidence intervals of posttest scores as a function of group and pretest score. Results of the statistical analysis indicated that

![Figure 8](image.png)  
\( \text{Figure 8. Distributions of overall pretest scores.} \)
posttest scores did not differ significantly as a function of the interaction between pretest score and group, $F(2, 117)=2.33, p=.10, \eta^2_p=.04$. We proceeded with the ANCOVA, which indicated that there were statistically significant differences at posttest ($\eta^2_p=.40$; see Table 6. Pairwise comparisons revealed that participants in both the lecture and the CBT groups outperformed participants in the control group. However, the comparison between the lecture group and the CBT group was not significant. A box-and-whisker plot of posttest scores (Figure 10) depicts an upward shift in the distributions of the lecture and CBT groups relative to the control group, and nearly identically sized and situated interquartile ranges for the two intervention groups (the low outlier in the CBT group is the posttest score for a participant who did not complete the training as intended). The distribution of posttest scores of the control group is very similar to its distribution of pretest scores. The small total range of the lecture group indicates that the scores of this group remained somewhat less dispersed than the scores of the other groups. The ANCOVA and distributions indicate that the lecture and CBT interventions were similarly effective for increasing participants’ overall posttest scores.

**Change from pretest to posttest.** The pattern of results for each group’s change from pretest to posttest largely mirror that of the posttest; the mean change for the CBT and lecture groups were similar ($d = 0.11$), and both were substantially larger than the mean change of the control group ($d = 1.69$ and $1.66$, respectively). The mean change score of the control group indicates that, on average, simply taking the test a second time resulted in only very small improvement ($d = 0.12$).
Figure 9. Means and 95% confidence intervals of posttest scores as a function of pretest score and group.

Figure 10. Distributions of overall posttest scores.
Like the plot of means and confidence intervals of posttest scores, the plot depicting the means and confidence intervals of change scores (Figure 11) shows that the slope of the regression line for the control group is slightly shallower than the slopes of the regression lines for the intervention groups. The test of the homogeneity of slopes of regression lines was not significant, $F(2, 117)=2.33$, $p=.10$, $\eta^2_p=.04$. An ANCOVA revealed significant differences in the mean change scores ($\eta^2_p=.40$; see Table 6); follow-up analyses indicated that the change scores for both the CBT group and the lecture group were significantly higher than the change score for the control group (CBT group, $d=1.69$ and lecture group, $d=1.66$), and the two treatment groups were not significantly different from one another ($d=0.11$). A box-and-whisker plot of change scores (Figure 12) shows that the upper 50% of change scores for the intervention groups do not overlap with the distribution of change scores for the control group, with the exception of one outlier. Although the interquartile range of the CBT group is substantially larger than that of the lecture group, the total ranges of the two groups are similar. This suggests that the middle half of change scores for the lecture group were less variable around the median than the change scores of the CBT group, but the two groups had equally disparate upper and lower quarters. It is important to note the range of changes produced by the lecture and CBT: in each case, there were participants who performed more poorly on the posttest than the pretest (i.e., they had negative change scores), and participants who improved by 30 or more questions on the posttest. The control group has the most condensed distribution of change scores, as would be expected given that these participants were not exposed to instruction.
Figure 11. Means and 95% confidence intervals of change scores as a function of pretest score and group.

Figure 12. Distributions of overall change scores.
**Interaction between group and pretest score.** It is possible that the differences in the distribution of pretest scores across the three groups could have benefitted the CBT condition because this group was farther from the measurement ceiling; the disproportionate number of participants with lower pretest scores in the CBT group might favor that intervention because those participants have more room for improvement before they reach the ceiling of the assessment. Figure 11 shows that across all intervention groups, the mean change score for participants with low pretest scores was highest, followed by participants with medium pretest scores and high pretest scores. At each pretest score level, the lecture and CBT groups were similar, and both scored substantially higher than the control group. In other words, on average, participants who scored low on the pretest gained a similar number of questions on the posttest regardless of whether they were in the lecture or the CBT group. Based on the visual analysis of this interaction, neither intervention was more effective than the other for participants with particular ranges of pretest scores.

The test evaluating the homogeneity of slopes assumption for overall change score supports this conclusion. That test examines the interaction between pretest score and group, and evaluates whether the change score varies as a function of an interaction between these two factors. As noted, the interaction of pretest score and group was not significant for overall change score, suggesting that differences in change scores cannot be attributed to an interaction between group membership and pretest score. Regardless of group, change scores were inversely related to pretest scores: in general, the higher participants scored on the pretest, the smaller their change score.
Subsets of the Assessment

Similar analyses were conducted on various subsets of the assessment – acquisition and reduction graphs, and slope and level questions – to evaluate whether one intervention was more effective than the other for specific types of graphs and/or discriminations. Tables 7 and 8 provide the means and standard deviations for each group on the pretest, posttest, and change for each of the subsets. For the most part, the data for the subsets closely resemble the data for the overall assessment. With the exception of level questions, on the pretest, the control group performed slightly better than the lecture group, who performed slightly better than the CBT group. Based on the descriptive statistics, the lecture group performed slightly better than the CBT group, and both intervention groups performed better than the control group on all subsets of the posttest. With the exception of slope questions, the CBT group’s mean change scores for all subsets were slightly higher than the lecture group’s mean change scores, and both intervention groups’ change scores were substantially larger than the control group’s change scores.

Table 7
Mean Scores for Acquisition and Reduction Graphs (Standard Deviations in Parentheses)

<table>
<thead>
<tr>
<th>Group</th>
<th>Acquisition Graphs</th>
<th>Reduction Graphs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pretest</td>
<td>Posttest</td>
</tr>
<tr>
<td>Control</td>
<td>25.10 (3.91)</td>
<td>25.66 (4.43)</td>
</tr>
<tr>
<td>Lecture</td>
<td>24.34 (3.37)</td>
<td>30.10 (3.81)</td>
</tr>
<tr>
<td>CBT</td>
<td>22.66 (4.03)</td>
<td>28.93 (4.03)</td>
</tr>
</tbody>
</table>
Table 8

*Mean Scores for Slope and Level Questions (Standard Deviations in Parentheses)*

<table>
<thead>
<tr>
<th>Group</th>
<th>Slope</th>
<th>Level</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pretest</td>
<td>Posttest</td>
<td>Change</td>
<td>Pretest</td>
<td>Posttest</td>
<td>Change</td>
<td>Pretest</td>
<td>Posttest</td>
<td>Change</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>24.93 (5.36)</td>
<td>25.59 (5.56)</td>
<td>0.66 (4.44)</td>
<td>26.27 (5.63)</td>
<td>26.71 (4.95)</td>
<td>0.44 (4.69)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lecture</td>
<td>22.85 (4.93)</td>
<td>31.42 (3.81)</td>
<td>8.56 (5.59)</td>
<td>27.17 (4.78)</td>
<td>31.63 (3.56)</td>
<td>4.46 (5.53)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CBT</td>
<td>22.14 (4.93)</td>
<td>30.05 (5.41)</td>
<td>7.73 (5.45)</td>
<td>25.02 (5.07)</td>
<td>31.22 (4.22)</td>
<td>6.20 (5.40)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 9

*ANOVAs of the Acquisition Graphs*

<table>
<thead>
<tr>
<th>Measure</th>
<th>THREE-GROUP COMPARISON</th>
<th>LECTURE vs. CONTROL</th>
<th>CBT vs. CONTROL</th>
<th>CBT vs. LECTURE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$F$</td>
<td>$df$</td>
<td>$p$</td>
<td>$\eta^2_p$</td>
</tr>
<tr>
<td>Pretest</td>
<td>4.47</td>
<td>2, 120</td>
<td>.01</td>
<td>.07</td>
</tr>
<tr>
<td>Change</td>
<td>25.03</td>
<td>2, 119</td>
<td>&lt;.001</td>
<td>.30</td>
</tr>
</tbody>
</table>

*a*ANOVA. *b*ANCOVA with pretest as covariate.

Tables 9-12 contain $F$ statistics, $p$-values, and effect sizes for ANOVAs, ANCOVAs, and tests of simple main effects of pretest scores and change scores for each subset of the assessment. The three-group ANOVAs of pretest scores for acquisition graphs and slope questions reached statistical significance. On pretests, the CBT group answered significantly fewer questions correctly about acquisition graphs than the control group and the lecture group ($d = -0.61$ and $d = -0.45$, respectively), and significantly fewer questions correctly about slope than the control group ($d = -0.54$).
Table 10

**ANOVA of the Reduction Graphs**

<table>
<thead>
<tr>
<th>Measure</th>
<th>Pretest Level</th>
<th>THREE-GROUP COMPARISON</th>
<th>LECTURE vs. CONTROL</th>
<th>CBT vs. CONTROL</th>
<th>CBT vs. LECTURE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F</td>
<td>df</td>
<td>p</td>
<td>(\eta_p^2)</td>
<td>p</td>
</tr>
<tr>
<td>Pretest</td>
<td>NA</td>
<td>0.65</td>
<td>2, 120</td>
<td>.52</td>
<td>.01</td>
</tr>
<tr>
<td>Change Score</td>
<td>Low</td>
<td>23.94</td>
<td>2, 117</td>
<td>&lt;.001</td>
<td>.29</td>
</tr>
<tr>
<td></td>
<td>Medium</td>
<td>30.19</td>
<td>2, 117</td>
<td>&lt;.001</td>
<td>.34</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>7.15</td>
<td>2, 117</td>
<td>.001</td>
<td>.11</td>
</tr>
</tbody>
</table>

Table 11

**ANOVA of Level Questions**

<table>
<thead>
<tr>
<th>Measure</th>
<th>Pretest Level</th>
<th>THREE-GROUP COMPARISON</th>
<th>LECTURE vs. CONTROL</th>
<th>CBT vs. CONTROL</th>
<th>CBT vs. LECTURE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F</td>
<td>df</td>
<td>p</td>
<td>(\eta_p^2)</td>
<td>p</td>
</tr>
<tr>
<td>Pretest</td>
<td>NA</td>
<td>1.78</td>
<td>2, 120</td>
<td>.17</td>
<td>.03</td>
</tr>
<tr>
<td>Change Score</td>
<td>Low</td>
<td>18.72</td>
<td>2, 117</td>
<td>&lt;.001</td>
<td>.24</td>
</tr>
<tr>
<td></td>
<td>Medium</td>
<td>18.18</td>
<td>2, 117</td>
<td>&lt;.001</td>
<td>.24</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>3.91</td>
<td>2, 117</td>
<td>.02</td>
<td>.06</td>
</tr>
</tbody>
</table>

Table 12

**ANOVA of Slope Questions**

<table>
<thead>
<tr>
<th>Measure</th>
<th>Pretest Level</th>
<th>THREE-GROUP COMPARISON</th>
<th>LECTURE vs. CONTROL</th>
<th>CBT vs. CONTROL</th>
<th>CBT vs. LECTURE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F</td>
<td>df</td>
<td>p</td>
<td>(\eta_p^2)</td>
<td>p</td>
</tr>
<tr>
<td>Pretest</td>
<td>NA</td>
<td>3.02</td>
<td>2, 120</td>
<td>.05</td>
<td>.05</td>
</tr>
<tr>
<td>Change Score</td>
<td>Low</td>
<td>24.35</td>
<td>2, 117</td>
<td>&lt;.001</td>
<td>.29</td>
</tr>
<tr>
<td></td>
<td>Medium</td>
<td>30.62</td>
<td>2, 117</td>
<td>&lt;.001</td>
<td>.34</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>5.03</td>
<td>2, 117</td>
<td>.008</td>
<td>.08</td>
</tr>
</tbody>
</table>
We planned to conduct ANCOVAs on each subset to evaluate any differences in change scores between groups for each subset of the assessment, and followed the same procedures used for the overall assessment: first testing the homogeneity of slopes assumption visually and statistically, and testing simple main effects when this assumption was violated.

**Acquisition graphs.** Figure 13 depicts the mean change score on acquisition graphs as a function of pretest score on acquisition graphs and group membership with error bars representing 95% confidence intervals. The slopes of the three lines are similar, suggesting that there is not an interaction between pretest score and group for acquisition graphs. The ANOVA of this interaction was not significant, $F(2, 117)=1.87$, $p=.16$, $\eta^2_p=.03$, and an ANCOVA was conducted as planned. The results follow the same pattern that was seen in the overall analysis: The lecture group and CBT group performed significantly better than the control group ($d=1.34$ and $d=1.60$ respectively), but were not significantly different from one another ($d=0.12$; see Table 9).

**Reduction graphs.** For reduction graphs, the slopes of the intervention groups appear steeper than the slope of the control group in Figure 14. An ANOVA of the interaction between pretest score and group membership was significant, $F(2, 117)=3.03$, $p=.05$, $\eta^2_p=.05$. This violates the homogeneity of slopes assumption, so we tested simple main effects as previously described. This test was significant for participants with low pretest scores ($p<.001$, $\eta^2_p=.29$), medium pretest scores, ($p<.001$, $\eta^2_p=.34$), and high pretest scores, ($p=.001$, $\eta^2_p=.11$). Pairwise comparisons indicated that, for all levels of pretest scores, the intervention groups outperformed the control group but did not differ significantly from one another (see Table 10). Effect sizes were calculated for these and
subsequent comparisons using the means and standard deviations of low, medium, and high pretest groupings within each treatment group. On the reduction graphs, effect sizes ranged from $d=1.34$ to $d=2.16$ for participants with low and medium pretest scores in the intervention groups compared to the control group, and were slightly smaller but still in the moderate to large range for participants with high pretest scores ($d=0.98$ and $d=0.66$). This pattern of results is similar to the results of the analysis of the overall assessment.

Figure 13. Means and 95% confidence intervals of change scores on acquisition graphs as a function of pretest score and group.
**Level questions.** Figure 15 shows change scores on level questions as a function of pretest score on level questions and group membership. Like the plot for reduction graphs, the regression line for the control group appears shallower than the regression lines for the intervention groups. The interaction between pretest score and group was significant for level questions, $F(2, 117) = 3.47, p = .03, \eta^2_p = .06$, suggesting that the slopes of the regression lines are different. The simple main effects tests were significant for low pretest scores ($p < .001, \eta^2_p = .24$) and medium pretest scores, ($p < .001, \eta^2_p = .24$), but not for high pretest scores ($p = .02, \eta^2_p = .06$). This is corroborated by the high degree of overlap between all three groups at the high pretest level. Pairwise comparisons indicate that, for participants with low and medium pretest level scores, the lecture group and CBT group improved significantly on level questions compared to the

![Figure 14](image-url)

*Figure 14.* Means and confidence intervals of change scores on reduction graphs as a function of pretest score and group.
control group (see Table 11). Effect sizes for these comparisons were large: for the lecture group, $d=1.52$ and $0.84$ for low and medium pretest scores respectively compared to the control group, and for the CBT group $d=1.43$ and $1.58$ for low and medium pretest. However, change scores for participants with high pretest level scores did not differ significantly across the three groups, and appear quite similar in the plot. The effect sizes for these comparisons were slightly smaller, but still in the moderate to large range: for the lecture group compared to the control group, $d=0.77$, and for the CBT group compared to the control group, $d=0.32$.

**Slope questions.** The plot of the interaction between pretest score on slope questions and group membership is displayed in Figure 16. Again, the line for the control group appears shallower than the line for the intervention groups, and the ANOVA of the interaction between was significant, $F(2, 117) = 4.18, p = .02$, $\eta_p^2 = .07$. The simple main effects tests were significant at low ($p < .001$, $\eta_p^2 = .29$), medium, ($p < .001$, $\eta_p^2 = .34$), and high pretest scores, ($p = .008$, $\eta_p^2 = .08$). Compared to the control group, participants with low and medium pretest scores in the lecture group ($d=2.43$ and $d=1.68$, respectively) and CBT group ($d=1.43$ and $d=1.45$, respectively) improved significantly, but were not significantly different from one another ($d =-0.64$ and $d=-0.15$ for low and medium pretest scores, see Table 12). Of the participants with high pretest scores, the comparison between the CBT and the control group was the only difference that reached statistical significance. The comparison between participants with high pretest scores in the lecture group and control group was not significant, but Figure 8 suggests this may be due to a lack of power. The intervention groups appear quite similar to one another with the exception of a very small degree of overlap between the lecture group and the control
group confidence intervals. In addition, the comparison between change scores of participants with high pretest slope scores in the lecture and CBT group compared with the control group yielded large effect sizes of $d=1.18$ and $d=1.39$, respectively.

In general, this analysis of the subsets suggests that one intervention was not more or less effective than the other with regard to either type of question (i.e., slope or level) or either type of graph (i.e., acquisition or reduction), and the role of pretest score remains unclear.

Figure 15. Means and confidence intervals of change scores on level questions as a function of pretest score and group.
Additional research is required to evaluate differences in how participants with high pretest scores on slope or level questions respond to each intervention, and the extent to which their performance is limited by ceiling effects of the assessment.

**Proportion meeting 80% accuracy criterion.** We analyzed the proportion of participants in each group meeting an 80% accuracy criterion on the posttest (a) overall, (b) on acquisition graphs, (c) on reduction graphs, (d) on slope questions, and (e) on level questions; these data are presented in Table 13. These results are comparable to the other results of the preceding analyses: a similar proportion of participants in the lecture

![Figure 16. Means and confidence intervals of change scores on slope questions as a function of pretest score and group.](image-url)
Table 13

Proportion of Participants Meeting 80% Accuracy Criterion on Posttest

<table>
<thead>
<tr>
<th>Group</th>
<th>Overall</th>
<th>Acquisition</th>
<th>Reduction</th>
<th>Slope</th>
<th>Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>12%</td>
<td>9%</td>
<td>21%</td>
<td>16%</td>
<td>14%</td>
</tr>
<tr>
<td>Lecture</td>
<td>49%</td>
<td>42%</td>
<td>63%</td>
<td>53%</td>
<td>51%</td>
</tr>
<tr>
<td>CBT</td>
<td>51%</td>
<td>26%</td>
<td>60%</td>
<td>49%</td>
<td>53%</td>
</tr>
</tbody>
</table>

group and CBT group met this criterion on the overall assessment and three of the four subsets, and substantially more participants from both the intervention groups met this criterion relative to the control group. The acquisition graphs are the exception: the proportion of participants from each intervention group meeting criterion was well below 50%, and 16% fewer participants (n=7) in the CBT group met criterion than the lecture group. On the overall assessment and all other subsets, the groups differed by no more than 3% (n=2).

**Item analysis.** To compare the proportion of participants who improved on individual items of the assessment across the three groups, we first calculated the proportion of participants within each group who answered each question correctly on the pretest and posttest, and calculated the percent change from pretest to posttest. Given very different pretest performance across groups on some items (e.g., 44% of participants in the control group answered item #14 [slope] correctly, compared with 59% of lecture participants and 24% of CBT participants), it is not appropriate to compare the proportion of participants from each group who answered particular items correctly at posttest. Likewise, comparing only the change in the proportion of participants answering each item correctly from pretest to posttest does not effectively accommodate items on which a particular group had a high proportion of participants answering correctly on the
pretest. In these cases, the ceiling affects how large the change can be; for example, 83% of participants in the lecture group answered item #4 (level) correctly on the pretest.

To accommodate these pretest differences, we first sorted the items by the change in the proportion of participants in the group who answered that item correctly. Items were categorized as improved if at least 10% more participants answered correctly on the posttest compared to the pretest; no change was defined as a change between 9% and -9%, and worsened was defined as a change of -10% or less. We examined the proportion of participants answering correctly on the pretest for items with no change; if at least 80% of participants answered that item correctly on the pretest, it was categorized as no possible improvement. The proportion of items in each of these categories for each group is presented in Table 14. The intervention groups were similar in the number of items on which a substantial proportion of participants improved, worsened, and did not change from pretest to posttest; the control group did not change on the majority of items.

The box-and-whisker plots in figures 17 and 18 show the distributions of the percent change in accuracy on items within each subset. Based on a comparison of the interquartile ranges and total ranges, it appears that the CBT produced more consistent

<table>
<thead>
<tr>
<th>Group</th>
<th>Improved$^a$</th>
<th>Worsened$^b$</th>
<th>Stable$^c$</th>
<th>No Possible Improvement$^d$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lecture</td>
<td>52</td>
<td>9</td>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td>CBT</td>
<td>58</td>
<td>7</td>
<td>11</td>
<td>3</td>
</tr>
<tr>
<td>Control</td>
<td>12</td>
<td>12</td>
<td>37</td>
<td>19</td>
</tr>
</tbody>
</table>

$^a$ more than +10% change. $^b$ more than -10% change. $^c$ between -9% and +9% change. $^d$ 80% or more participants in group answered correctly on pretest.
changes (i.e., smaller distributions) than the lecture on questions about reduction graphs. The median percent change on items is similar across the two interventions (approximately 20%), but the middle 50% of items about reduction graphs for the CBT group had improvements of approximately 10% to 25% compared with 2% to 30% for the lecture group. The distributions of percent change for items on acquisition graphs are similar across the two groups. For level questions, the median percent change on items is noticeably different: approximately 20% improvement for the CBT group compared with approximately 10% improvement for the lecture group. The middle 50% of level questions in the CBT group ranged from 10% to 25% improvement, compared to 2% to 25% in the lecture group. In contrast, it appears that the lecture may have produced more uniform changes than the CBT on slope questions (see Figure 17).

The items on which a large proportion (i.e., more than 25%) of the intervention groups improved from pretest to posttest were of particular interest because these items likely have characteristics that the instruction addressed effectively. There were 16 items on which both groups improved substantially, 8 on which only the lecture group improved substantially, and 6 on which only the CBT group improved substantially. Figure 19 contains examples of graphs on which both intervention groups improved by more than 25%. The eight items on which both groups had the largest improvement were asked about graphs with a therapeutic trend in baseline that continued into the intervention phase. In the didactic instruction, we included a rule that if the actual treatment data were not better than what would be predicted given the baseline data, the answer to both questions for that graph (i.e., slope and level) was “no.” Based on high proportions of improvement on the graphs to which this rule applied, it appears that this
instruction was effective. Two other features were prominent in the items that both groups improved substantially on: a) level questions about graphs with a therapeutic trend in treatment but no overall level change (4 items; correct answer was “No”) and b) slope questions about graphs with a change from a counter-therapeutic trend in baseline to a flat slope in the intervention phase (4 items; correct answer was “Yes”).

Figure 17. Distribution of proportional changes on acquisition and reduction graphs.
Similarly, the items on which a large proportion (i.e., more than 25%) of the intervention groups worsened from pretest to posttest were examined because these items might highlight aspects of the instruction that were misleading to participants. There were two items (about one graph) on which both groups worsened substantially (see Figure 20), one on which only the lecture group worsened, and one on which only the CBT group worsened. Interestingly, the four items on which intervention participants worsened were all questions from the acquisition set; participants in the two intervention groups did not worsen on any graphs in the reduction set. Three of these questions were about level and one was about slope.

![Figure 18](image_url)

*Figure 18.* Distribution of proportional changes on slope and level questions.
**Figure 19.** Three graphs on which both intervention groups improved by at least 25%.
These items on which the intervention groups’ performance worsened were examined and compared to identify any characteristics that the graphs might share. All three graphs had flat slopes in baseline; however, some had multiple trends within the phase and/or one data point that may have overly influenced participants’ estimation of the slope. If participants incorrectly estimated the slope in baseline, they may have applied the previously described rule about how the baseline slope influences judgment of level. For example, participants may have incorrectly estimated the baseline slope as increasing in the graph in Figure 20. Projecting an increasing trend line into the treatment phase, the actual data in treatment are not better than what would be predicted based on baseline. When the actual data were not better than predicted, participants were taught to answer “no” to the question about level even if the level appears different in the treatment phase, because the therapeutic baseline trend confounds any conclusion about a

*Figure 20.* The acquisition graph on which both intervention groups worsened by more than 25%.
treatment effect. Errors in estimating slope, coupled with application of this rule about
the effect of slope on level, might explain why level questions constituted 3 of the 4 items
on which the accuracy of the intervention groups declined.

**Effectiveness and Efficiency of the Computer-Based Training**

The secondary dependent variables were analyzed by computing descriptive
statistics to evaluate the effectiveness and efficiency of the modules and instructional
segments within the CBT (see Table 15).

**Training Durations**

The average duration of training for the entire CBT package was 105 min; as
anticipated, there was wide individual variation, with the minimum training time at 87
min and the maximum training time at 150 min. In general, training durations for each
module increased as participants progressed through the training; participants spent an

<table>
<thead>
<tr>
<th>Variable</th>
<th>Intro</th>
<th>Level</th>
<th>Slope</th>
<th>Slope and Level</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training Duration (min)</td>
<td>$M$</td>
<td>12.42</td>
<td>21.05</td>
<td>35.45</td>
<td>105</td>
</tr>
<tr>
<td></td>
<td>$SD$</td>
<td>1.75</td>
<td>4.78</td>
<td>6.53</td>
<td>13.90</td>
</tr>
<tr>
<td>In-Module Pretest</td>
<td>$M$</td>
<td>67.8%</td>
<td>81.5%</td>
<td>73.3%</td>
<td>70.8%</td>
</tr>
<tr>
<td></td>
<td>$SD$</td>
<td>15.1%</td>
<td>16.5%</td>
<td>15.4%</td>
<td>12.7%</td>
</tr>
<tr>
<td>In-Module Posttest</td>
<td>$M$</td>
<td>90.0%</td>
<td>97.3%</td>
<td>90.3%</td>
<td>82.2%</td>
</tr>
<tr>
<td></td>
<td>$SD$</td>
<td>6.4%</td>
<td>7.0%</td>
<td>9.5%</td>
<td>15%</td>
</tr>
<tr>
<td>In-Module Change</td>
<td>$M$</td>
<td>22.3%</td>
<td>15.9%</td>
<td>17.0%</td>
<td>11.4%</td>
</tr>
<tr>
<td></td>
<td>$SD$</td>
<td>13.9%</td>
<td>16.6%</td>
<td>16.8%</td>
<td>11.5%</td>
</tr>
<tr>
<td>Percent at 90% on 1st Attempt</td>
<td></td>
<td>98%</td>
<td>93%</td>
<td>63%</td>
<td>50%</td>
</tr>
<tr>
<td>Percent at 90% on 1st or 2nd Attempt</td>
<td></td>
<td>100%</td>
<td>98%</td>
<td>78%</td>
<td>68%</td>
</tr>
</tbody>
</table>
average of 12.42 min ($SD=1.75$) in the Introductory module, 21.05 min in the Level Change module ($SD=4.78$), 35.45 min in the Slope Change module ($SD=5.93$), and 37.03 min in the Slope and Level Change module ($SD=5.93$). The correlation between overall training duration and change score on the primary assessment was very small, $r(38)=.008, p=.96$.

**In-Module Pre and Posttest Performance**

Across the modules, in-module pretest performance ranged from 68% (Introduction) to 82% (Level Change); with the exception of the Slope and Level Change module, standard deviations of pretest scores were around 15%. Posttest scores ranged from 82% correct (Slope and Level Change) to 97% correct (Level Change). The proportion of participants meeting the 90% accuracy criterion on the in-module posttest on the first attempt worsened with each subsequent module: 98% of participants met criterion on the first attempt at the Introduction posttest and only 50% met criterion on the first attempt at the Slope and Level Change posttest. On average, students who did not pass an initial posttest showed modest improvement after additional practice and a second attempt: the mean improvement from first posttest to second posttest was 1.33 questions for Level Change, 0.73 questions for Slope Change, and 2.00 questions for Slope and Level Change. However, it is unclear whether these gains can be attributed to the extra practice items provided, taking the test a second time, or a combination of the two. The proportion of participants meeting the 90% criterion on the first or second posttest attempt was above 85% for all modules except Slope and Level Change.

It is difficult to compare relative effectiveness of the modules given the different types and difficulty of questions in the modules (e.g., the Introduction module consisted
mainly of simple definition/identification questions, whereas the others required application of the module content to novel graphs), as well as the potential confound of fatigue as participants completed the entire training in one sitting. However, the data suggest that the Slope Change and the Slope and Level Change modules may have been less effective at teaching the targeted skills than the other modules.

**Accuracy on practice questions**

Another measure of the effectiveness of the computer-based instruction is the mean accuracy on practice questions within (a) each module and (b) each segment in each module. Each module contained a different number of questions in practice sets and the remediation loops, and participants answered varied numbers of questions based on their accuracy. Therefore, the mean accuracy is expressed as a percent of the mean number of questions answered in that module. Table 16 displays this information for the three modules directly related to graph discriminations.

In the Level Change module and Slope Change module, one question was equivalent to one response opportunity (e.g., “Do the data suggest that the intervention caused a change in the level of the behavior?”). In the Slope and Level Change module, one question contained two or three response opportunities (e.g., “Do the data suggest that the intervention caused a change in the level of the behavior?” AND “Do the data suggest that the intervention caused a change in the slope of the behavior?”). In this module, we used an all-or-none scoring method: participants had to answer all response opportunities correctly to get that question correct. Therefore, the scoring of the Slope and Level Change module was quite stringent and may underestimate how well participants were actually performing in this module. (On the in-module pre and posttest
for this module, each response was scored separately. Based on the accuracy percentages, it appears that the Level Change module was most effective, and the Slope Change module and the Slope and Level Change modules were similarly effective.

In addition to examining the overall effectiveness of each module, we analyzed instructional segments within each module in the same way to pinpoint aspects of the instruction that were particularly effective and those that were less effective. The mean accuracy on each instructional segment is presented in Table 16. Accuracy was high across (a) all segments of the Level Change module, (b) segments containing graphs with

Table 16

Effectiveness and Efficiency Measures of Modules and Instructional Segments

<table>
<thead>
<tr>
<th>Module</th>
<th>Accuracy on Practice Questions</th>
<th>Efficiency Index</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Level Change</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Instructional Segments</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Estimating level (low variability) and using band (high variability)</td>
<td>93%</td>
<td>1.30</td>
</tr>
<tr>
<td>2. Judging level change (low variability)</td>
<td>90%</td>
<td>1.59</td>
</tr>
<tr>
<td>3. Judging level change (medium variability)</td>
<td>99%</td>
<td>1.11</td>
</tr>
<tr>
<td>4. Judging level change (high variability)</td>
<td>90%</td>
<td>1.71</td>
</tr>
<tr>
<td><strong>Slope Change</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Instructional Segments</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Estimating the slope (low and medium variability)</td>
<td>96%</td>
<td>1.04</td>
</tr>
<tr>
<td>2. Estimating the slope (high variability)</td>
<td>73%</td>
<td>2.63</td>
</tr>
<tr>
<td>3. Judging slope change (low variability)</td>
<td>94%</td>
<td>1.41</td>
</tr>
<tr>
<td>4. Judging slope change (medium variability)</td>
<td>76%</td>
<td>2.33</td>
</tr>
<tr>
<td>5. Judging slope change (high variability)</td>
<td>65%</td>
<td>3.33</td>
</tr>
<tr>
<td><strong>Slope and Level Change</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Instructional Segments</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Estimating the level with slope (high variability)</td>
<td>100%</td>
<td>1.00</td>
</tr>
<tr>
<td>2. Judging whether actual data are better than predicted (high variability)</td>
<td>78%</td>
<td>3.40</td>
</tr>
<tr>
<td>3. Using prediction to judge change (high variability)</td>
<td>60%</td>
<td>3.04</td>
</tr>
<tr>
<td>4. Evaluating slope and level change (high variability)</td>
<td>61%</td>
<td>2.64</td>
</tr>
</tbody>
</table>
low variability in the Slope Change module, and (c) the segment on estimating the level when the data path has an increasing or decreasing slope. Accuracy was relatively low on the segments of the Slope Change module that contained graphs with high variability (2, 4, and 5). This suggests that the instruction may have been inadequate to teach participants how to accurately estimate the slope of a highly variable data path. An analysis of the apparent low accuracy in the final three segments of the Slope and Level Change module is complicated by (a) the low accuracy on estimating and comparing slopes of highly variable data paths in the Slope Change module and (b) the previously described all-or-nothing scoring in this module.

**Efficiency of Instruction**

The modules (and instructional segments within the modules) in the CBT included required practice sets and additional remediation questions that participants were directed to when they answered a question incorrectly in one of the practice sets. Therefore, each module had a minimum required number of questions that participants had to answer correctly to advance to the in-module posttest. In an efficient module, the average number of questions answered would be close to the minimum number of questions required, because this would indicate that participants were answering the questions in the required practice set correctly without being directed to many remediation questions. We divided the mean number of questions answered by the minimum number of questions required in the module to quantify the efficiency of each module. The resulting *efficiency index* describes the number of questions answered as a proportion of the minimum number of questions possible. The closer the index to 1, the closer the actual number of questions answered was to the minimum possible number of
questions. Table 16 includes the efficiency indices for each module and instructional segments within each module. In general, the pattern of these indices is very similar to the pattern observed in the accuracy of participants’ responses to practice opportunities. The Level Change module was the most efficient, with participants answering only about 1.5 times more questions than required; the Slope Change and Slope and Level Change modules were similar at 2.15 and 2.70 times more questions, respectively. Again, the scoring of the Slope and Level Change module affected the number of questions participants had to answer, because if one was scored as incorrect the participant was routed to the remediation loop. This makes it difficult to compare across modules without much more detailed analyses of specific response patterns in the Slope and Level Change module.

Across modules, the instructional segments that were most efficient (i.e., with the lowest index number) were those that contained graphs with low variability: each module had at least one segment with an efficiency index close to 1. The least efficient segments contained graphs with high variability. Overall, the least efficient segment required participants to judge whether the actual data were better than predicted based on the baseline data. Although this segment was in the Slope and Level Change module, efficiency was not artificially affected by the scoring guidelines, as these question slides contained only one response opportunity.

**Treatment Acceptability**

Participants in the lecture and CBT groups completed brief surveys about the trainings. On a 5-point Likert-type scale, with 1 representing *strongly disagree*, 3
representing *neutral*, and 5 representing *strongly agree*, participants rated their agreement with the following statements: “I enjoyed participating in the training,” and “I learned a lot from the training about interpreting graphs.” The means, standard deviations, and results of ANOVAs comparing these ratings are presented in Table 17. On average, participants in both groups were neutral about whether they enjoyed the training, and agreed that they learned a lot about interpreting graphs from the training. An ANOVA confirmed that the groups were not significantly different with regard to how much they said they enjoyed the intervention and how much they said they learned from the intervention.

The surveys also included an open-ended question about what could be done to make the training more engaging. Participants’ responses to this question were grouped into several categories and tallied. The most common responses from participants in the CBT group concerned the length and density of the training (*n*=15; “It was long, it would be good to break it up by doing it on different days”) and the remediation loops (*n*=12; “All of the extra practice questions got to be redundant”). The most common responses from participants in the lecture group concerned the perceived effectiveness of the videotaped lecture (*n*=7; “I learned a lot from the lecture”) relative to the reading (*n*=7; “The reading was super boring and I don’t think I learned anything from it”) and the
desire for some sort of interactive response opportunity to check understanding (n=7; “Maybe have something clickable where I can see if I’m getting it”). Interestingly, at least two participants in each group thought that some aspect of the other group’s intervention would make the instruction more effective and engaging.
CHAPTER V

DISCUSSION

The purpose of the current investigation was to add to the literature on training individuals to visually analyze AB graphs. Previous studies have relied heavily on teaching participants to use mechanical methods, which may not be practical for use in real-world settings and may not accommodate all types of data patterns. The training in the present study was designed to provide explicit instruction on the components of visual analysis (i.e., estimating slope and level), with the intent of developing accurate and flexible analysis. In addition, limitations of the assessment tools (e.g., lack of slope in the data paths, use of programmed parameters as the accuracy criterion) used in previous studies limit the conclusions that can be drawn about the trainings. In the present study, the assessment tool included (a) complex graphs with high variability and programmed slopes, (b) only graphs on which four independent experts agreed on the presence or absence of slope and level changes, and (c) separate questions about slope and level for each graph.

We evaluated and compared two methods of training individuals to visually analyze AB graphs: a traditional lecture format and a CBT. The instructional content was equated across the two methods; in other words, participants in the lecture and CBT groups were taught the same procedures for estimating slope and level using the same example graphs. Differences between the two interventions include (a) the media, (b) the reading completed by the lecture group, and c) extensive required practice opportunities and feedback included in the CBT.
The results of the study clearly and strongly suggest that, compared to no training, both interventions were effective at increasing accurate discriminations of slope and level changes across acquisition and reduction graphs. Participants in both intervention groups gained an average of approximately 13 questions correct from pretest to posttest (16.25%). However, the average posttest scores for both intervention groups were approximately 80% - neither treatment produced near-ceiling performance. Further, there was wide individual variability in both groups; only 51% of participants in the CBT group and 49% in the lecture group answered at least 80% of questions correctly on the posttest.

Effectiveness of the Lecture Intervention

It is perhaps unsurprising that participants who received training performed better on the posttest than those who did not receive training; however, the effectiveness of the lecture condition is notable given the mixed results of previous research. Stewart et al. (2007) found that a brief, videotaped lecture had little impact on the accuracy of six participants’ visual analysis. There are several potential explanations for the divergent results of Stewart et al. (2007) and the current investigation. The assessment tools differed between the two studies: Stewart et al. (2007) did not program slope in the assessment graphs, and counted any graph with a programmed effect size larger than 0 (e.g., 0.25) as demonstrating an effect. It is impossible to compare the difficulty of the two assessment tools, as the lack of slope might suggest that the graphs used by Stewart et al. (2007) were not as difficult as those included in the present investigation, but very small effect sizes used in that study could be quite difficult to discriminate visually. In addition to the assessment tools, the content of the lecture in Stewart et al. (2007) was
likely quite different from the lecture in the present study. Although both were based on the Cooper et al. (2007) chapter on visual analysis, it is possible that our lecture provided more examples and nonexamples of various types of data patterns and/or more guidelines on how to estimate slope and level. The lecture in Stewart et al. (2007) was only 6 min in duration, compared to 36 min in the present study.

The lecture condition in the present study may have been effective due to the addition of reading and practice items in the present study; Stewart et al. (2007) examined only the effects of the brief lecture without supplemental materials. Finally, it is possible that the differences are due to sampling error in the Stewart et al. study; Stewart et al. (2007) evaluated their lecture with only 6 participants, whereas 41 participants contacted the lecture condition in the present study.

The performance of participants in the lecture group in this study was more similar to that of participants in the lecture condition in Jostad (2011). In Jostad (2011), participants in the lecture condition improved from a mean of 61% on the pretest to a mean of 73% on the posttest after viewing a 20 min videotaped lecture on visual analysis. However, only 35% of lecture participants in Jostad (2011) reached 80% accuracy on the posttest compared with 49% of lecture participants in the current study. Although it is unclear why our results were more similar to those obtained by Jostad (2011) than by Stewart et al. (2007), the lectures in the present study and Jostad (2011) were closer in duration (20 min and 36 min), and it is possible that the assessment graphs may have been more similar given that Jostad (2011) did include slope changes.
Relative Effectiveness of the Two Interventions

Based on the varied results of previous research on the effectiveness of lectures, we anticipated that the CBT, with required active responding and feedback, would be more effective for teaching participants to conduct visual analysis. We expected that accurate discriminations of slope and level change would not be as effectively shaped by rules alone (i.e., descriptions and demonstrations in the lecture) as by rules and direct contingencies (i.e., descriptions and demonstrations plus opportunities to respond and receive feedback in the CBT). Accordingly, it was somewhat surprising that the performances of the lecture group and CBT group were so similar.

The lecture and CBT may have been comparably effective because the didactic content (i.e., the content of the lecture and the instructional frames in the CBT) was identical between the two conditions. Equating this content isolated the following differences between the two groups: (a) the media, or instructional delivery method, (b) the required practice with feedback in the CBT group, (c) the supplemental materials (i.e., reading and self-study items) provided to the lecture group.

Much has been written about the effects of media on learning (Clark, 1983, 1994; Kozma, 1994; Kulik, 1985). Clark (1994) claims that, aside from possible economic benefits, instructional media (e.g., lecture, computer, television) themselves have no differential influence on learning when instructional design is held constant. Meta-analyses of the research on media and learning suggest that media matters, and specifically that computers can be superior to live instruction (e.g., Kulik, 1985). However, most individual studies within these analyses did not control for instructional content across media conditions; when he re-analyzed the data, Clark (1985) found that
removing this confound resulted in no difference between groups. Although the two conditions in the present study were not exactly equivalent (the CBT group received forced practice opportunities with feedback), the didactic content was identical. Clark’s (1994) work suggests that we should not expect differences in learning outcomes when instructional content is equated but instructional delivery differs; however, one delivery system (e.g., computers) may be more efficient, economical, or disseminable than others (e.g., a live lecture). In the current study, both interventions were similar in the duration of training and the ease with which they could be disseminated.

Based on the gains made by participants in the CBT and lecture groups, it appears that the numerous practice opportunities within the CBT (M=162) did not provide any measurable benefit to participants in that group. One potential explanation for this is that the explicit instructions and demonstrations provided in the didactic content were sufficient to increase accuracy on the posttest without required active practice. However, the lecture group read portions of a chapter on visual analysis and received self-study materials in addition to viewing the lecture. It is impossible to know the separate contributions of each of these components; for example, there may have been gains attributable to the practice opportunities for the CBT group that were compensated by gains attributable to the reading for the lecture group, and we cannot conclude that the didactic content alone was responsible for the gains made by both groups. It is also possible that the graphs in practice items in the CBT may not have been similar enough to the graphs in the assessment. Although we attempted to select graphs for practice items that contained the features present in the graphs in the assessment, discriminations may not have generalized to the graphs in the assessment. However, the below-criterion
performance of participants in CBT on these practice items (see following discussion) complicates the analysis, because participants may not have mastered these practice items within the context of the CBT.

One potentially important difference in modality may have created a measurement artifact favoring the lecture group. The self-study materials were essentially flashcards with graphs printed on one side and the answer on the other; these were taken directly from practice opportunities in the CBT. Participants were told that they could, but did not have to, write on these self-study materials. Given that lecture participants could physically draw slope and level lines on these graphs, and the overall assessment was delivered via paper and pencil, it is possible that this afforded the lecture group an advantage in terms of generalizing to the posttest. However, less than half of the lecture participants drew slope and/or level lines on their self-study materials ($n=16, 39\%$), and the correlation between change scores and drawing these lines on the self-study materials was small, $r(39)=.067, p=.68$.

It is possible that the CBT group outperformed the lecture group on dimensions of the behavior that we did not measure. The practice opportunities and direct contact with contingencies (i.e., correction and praise) in the CBT may have generated behavior that maintained better over time compared to the lecture. Future researchers may wish to evaluate whether high rates of practice opportunities related to visual analysis, such as those presented in the CBT, supports maintenance. However, in real-world settings, it is unlikely that participants would learn visual analysis in isolation and not contact additional formal or informal opportunities to practice the skill.
Why Weren’t the Interventions MORE Effective?

Although participants in the CBT and lecture groups improved substantially relative to participants in the control group ($d=1.69$ and $d=1.66$, respectively) and attained mean scores of 77% and 79% on the posttest, it is important to acknowledge that only 51% and 49% of participants in each group performed at or above 80% on the posttest. These mean performance levels are better than those reported by previous researchers who trained individuals to conduct visual analysis with a structured approach (Colon, 2006; Jostad, 2011) or without a structured approach (Jostad, 2011); however, it is unclear why the interventions were not more effective. There was a great deal of individual variability in how participants responded to the interventions; even within pretest-score groups and intervention condition, some individuals improved substantially, others remained stable, and a few worsened from pretest to posttest. Figure 21 depicts this variability in a box-and-whisker plot of change scores by pretest score and group. The total ranges and interquartile range of change scores for the lecture and CBT groups are quite large, particularly for participants in the low and medium pretest groups. This variability is also evident in the standard deviations of change scores for each group (8.37 and 7.60 for the CBT and lecture groups, respectively). Given the active responding and mastery-based design of the CBT, and the relatively passive nature of the lecture condition, we expected that the CBT would reduce individual variability. However, participants varied widely in how they responded to both interventions; improvements of either instructional method should focus on strategies to increase the effectiveness of the interventions for a wider range of learners.
In terms of the didactic content that was common to both interventions, instructions on how to estimate level and slope were quite specific, but may not have been explicit enough to produce accurate discriminations on some of the more challenging items on the posttest. We chose not to teach participants to draw a split-middle line (e.g., White & Haring, 1980) or other procedure for constructing a trend line due to the previously discussed limitations of those approaches, but had participants learned such a process, their accuracy may have been higher on the posttest.

The difficulty of some items on the assessment may have lowered the functional ceiling of the test and suppressed participants’ accuracy percentages. Although there were 80 items on the assessment, the highest score attained by any participant was 73.

Figure 21. Distributions of change scores for participants within each pretest-score group and intervention group.
Some items may have been somewhat ambiguous, and would perhaps prove difficult to attain agreement across additional experts. The questionable validity of these items poses challenges for interpreting criterion-level performance and drawing conclusions about the overall effectiveness of the interventions. The validity of all items on the assessment could be evaluated by increasing the number of experts involved in testing each item; high agreement across more experts would increase our confidence in the extent to which the test items are valid (i.e., sufficiently clear examples of an effect or noneffect).

Participants’ motivation to learn the content must be considered as the effectiveness of the trainings is evaluated. In the present study, participants did not have an immediate incentive to improve their performance from pretest to posttest; they were compensated with extra course credit and a gift card for basic participation in the study (i.e., taking the pretest, completing the training, and taking the posttest), but were not required to perform at a certain level within the training or on the posttest to obtain this compensation. I observed a range of behaviors suggesting substantial individual variation in motivation in both intervention groups. For example, one participant in the lecture group highlighted and wrote notes on the assigned reading, while others in that group who appeared to skim through the reading much more quickly (e.g., 15 min) than other participants. If there were contingencies in place motivating criterion-level performance on the training and/or posttest, as would be present in the context of a course or job training, it is possible that the trainings might have had a greater impact on posttest scores. For example, participants may have attended to the material more closely and/or put more effort into mastering it if told that they were required to reach a certain performance level on the posttest in order to receive extra credit.
Improving the Interventions

A close examination of the item analysis and treatment acceptability data for each intervention group, along with the analysis of secondary dependent variables for the CBT, provide a myriad of data-based options for improving the trainings. In terms of the didactic content, which was the same across both interventions, the item analysis suggests that more specific instruction and/or practice with feedback was required on discriminations of small level changes with high variability and estimations of slope of data paths with (a) multiple trends and (b) one or two outlying data points. These skills are prerequisites for subsequent instruction on evaluating slope and level changes simultaneously, so it is difficulty to draw strong conclusions about the success of the final module.

When data were variable, participants were taught to imagine a band covering most of the data points in the phase. To improve discrimination of level changes, the training might include more demonstrations of how to adjust the width of this band based on the amount of variability in the data path. A more systematic method of estimating slope (e.g., the split-middle method) may produce more accurate discriminations of slope change as well as fewer misapplications of the rule about making predictions based on baseline data.

In addition to potential changes to the content of the instruction, the lecture intervention could potentially be improved by delivering the content live. This format would allow participants to ask questions throughout the lecture (instead of at the end, as typically occurred when participants asked questions in the study) and actively respond and receive feedback. However, this would also make dissemination of the content more
difficult. A component analysis of the features of the lecture intervention (i.e., reading, lecture, self-study materials) might identify one or more components that do not improve accuracy. This type of analysis could support the development of a more streamlined intervention package.

Similarly, the content and delivery of the CBT could be modified in several ways. Along with the previously described changes in the didactic instruction, additional practice items with the features of the more difficult graphs in the assessment (e.g., small level changes with high overlap, data paths with one or two outlying data points) could be included in the CBT to improve accurate discrimination of slope and level changes on these challenging graphs. In addition, the practice items in the current version were primarily multiple-choice questions; response opportunities in which participants estimate slope and level by “drawing” their own lines, or dragging and dropping lines into the correct place, would likely be more effective at teaching the targeted skills than the selection-based responding in the current training. Estimating the slope and level of one phase is the critical prerequisite skill for making discriminations of slope change and level change across two phases. In the CBT, participants primarily practiced this skill by selecting a) the appropriately-placed level band from three options or b) the appropriately-placed trend line from three options. Neither of these types of responses required participants to overtly construct slope or level lines for a phase; participants chose the best option from the three given lines. On subsequent two-phase graphs within the training, participants learned to visualize these lines for each phase and make comparisons between the phases. However, because of the selection-based responding in the segments on estimating slope and level of one phase, it is not clear that participants
mastered this skill in a way that was likely to generalize to later segments of the training, the posttest, or the real-world. In other words, participants practiced and demonstrated this important prerequisite skill (i.e., estimating slope and level) using a modality that was not matched to the target modality. Responses that align more closely in modality with the posttest (and real-world application of the skill) may be incorporated into the CBT with advanced programming.

The training might also be improved by exploring different ways of providing remediation. In the current study, an error on a single practice item directed participants to remediation items; after completing these, participants returned to attempt the practice items again. It is not clear whether it was necessary to route participants to the remediation items following a single incorrect response. The training may have been more efficient, just as effective, and potentially less frustrating if participants were allowed two incorrect responses on the practice items before being guided to the extra practice items. Alternatively, a learner could be assigned a single additional practice item after each error rather than a set of multiple items. The effectiveness of the training might also be improved by providing additional instruction tailored to the nature of the error. In the current version, participants received feedback when they made an error (i.e., they were shown the graph with slope and level lines along with the correct answer), but this feedback was not necessarily specific to the error that they had made. For example, in the Slope and Level Change module, some questions had three parts, but the feedback provided was the same regardless of which part participants answered incorrectly. Feedback about why their response was wrong, in addition to the correct answer, might enhance the effectiveness of the training. Neither of these design features
(remediation items after two incorrect responses and specific instruction based on the nature of the error) are available out-of-the-box on Adobe Captivate®, but could be included with more sophisticated programming.

It would be worthwhile to examine ways to increase the efficiency of the CBT by (a) improving the didactic instruction as previously described, (b) changing the types of responses, (c) adjusting the selection of practice items to better represent the most difficult types of items, (d) providing an appropriate number of practice items, and/or (e) including features that might make the CBT more engaging. The challenge of providing sufficient, but not excessive, practice in each skill area is influenced by the dichotomous responses and the complexity of the graphs. For example, participants have a 50% chance of answering a given yes/no question correctly if they have no knowledge of the content; thus a small number of correct responses are not a strong indicator of actual skill. In a similar vein, the one or two graphs included in those questions might not capture the range of variation in data patterns that participants would be exposed to in later modules or the assessment. Non-dichotomous response formats such as those in which students draw trend lines might allow for reduced practice items because the probability of responding correctly by chance would be much smaller than 50%.

The impact of participant motivation on the efficiency and effectiveness of the CBT are unclear, but feedback on the treatment acceptability questionnaires suggests that it would be worthwhile to examine ways to make the CBT more engaging. Graphs could be contextualized with additional real-world examples, and graphics, videos, or animations embedded to illustrate those examples. These changes may increase interest in the topic by connecting it to situations that participants might encounter in the real world.
In addition, the feedback in the CBT was somewhat dull (e.g., captions with the text, “Great job! That is correct”). Other features, such as audio and animation, could be included in the feedback to increase engagement.

Finally, it is not clear how the logistics of delivering the training affected its effectiveness. Participants completed the training in one sitting; in response to an open-ended question about improving the training, many commented ($n=15$) that they felt they would perform better if it were broken up. The different modules could be completed on separate days, but additional research would be required to evaluate whether this impacts participant performance.

**Limitations**

**Methodology and Interpretation**

As previously discussed, the validity of the assessment tool has been a limitation of most research on training visual analysis. We took steps to address the weaknesses of other assessments by including graphs with programmed slope and using expert consensus as the criterion for the selection of graphs and participant accuracy. However, several concerns related to the validity of our assessment remain. As previously discussed, some items on the assessment may have been relatively ambiguous examples of slope and/or level change, suggesting that unanimous agreement across four independent experts may not be sufficient to identify valid test items.

To control the features of the graphs in the assessment, we chose to generate graphs rather than use actual graphs. However, this limits the external validity of the study. At the root of this issue is the extent to which the assessment graphs resemble
actual data, and consequently, whether a participant’s performance is a valid estimate of their performance on real-world graphs. Real behavioral graphs rarely have 10 data points in each phase, and some of the patterns may not be commonly encountered (e.g., a phase change following a therapeutic trend in baseline). In many cases it could be argued that the assessment graphs are more difficult than would be encountered in the real-world, and that individuals who correctly discriminate changes in these graphs are better prepared to evaluate less complex data patterns that they may face in actual data. The graphs in the assessment tool may be more difficult than most graphs in published journal articles because presumably those published graphs have attained some consensus among reviewers about the presence of an intervention effect; as a result, often the changes in slope and/or level are quite large and unambiguous. Alternatively, the graphs in our assessment tool may be less difficult than real-world graphs. For example, all of our graphs contained ten data points per phase, and as a result, participants did not have to attend to the number of data points per phase and how the number of data points impacts the relative certainty of predictions that can be made. To evaluate the external validity of the data patterns in the current assessment tool (e.g., level, trends, and variability), we could provide experts with generated assessment graphs and actual graphs, and ask them to identify the source of each graph.

Another limitation to the interpretation of our results is the potential artifact created by giving the lecture group paper graphs to practice with prior to taking the paper-and-pencil posttest. Although less than half of participants wrote on these graphs, and the correlation between writing on the graphs and change scores was very small, it is
not clear whether the similar modality of the practice items and the posttest improved the scores of this group.

Effectiveness of the Treatments

Although the mean accuracy of each intervention group improved relative to the control group, there were individuals within each group (a) whose accuracy did not improve (or worsened) following the training and/or (b) whose accuracy on the posttest was below 80%. The previously described changes might reduce this variability, but it is possible that some individuals might require one-on-one, face-to-face instruction to make complex discriminations of slope and level change. Nonetheless, either of these interventions might be a suitable first tier of instruction on this topic, with non-responders receiving more individualized instruction.

Scope of the Study

Our research questions were designed to evaluate the immediate effects of the trainings; as such, participants took the posttest immediately following instruction, and we did not collect maintenance data. Therefore, the lasting effects of the trainings are unknown. Considering that we would like students, teachers, and researchers to continue to accurately discriminate slope and level changes in graphed behavioral data, it will be important to evaluate if either or both of the interventions produce gains that maintain over time.

As an initial training on the component skills of visual analysis, the trainings and assessment included AB graphs only. Although these are not true experimental designs, they are a building block of more complex graphs and are commonly used in some
clinical contexts. It is important to note that the trainings in the study were not designed to teach individuals how to evaluate a functional effect via visual analysis, but rather to discriminate the basic types of changes that might evidence a functional effect given replication of the effect over time.

Ideally, in both research and practice, data guide decision-making in an ongoing manner, rather than after all of the data are collected. Researchers have begun to investigate the variables that influence individuals’ decisions to change phases in the process of collecting data (Vanselow, Thompson, & Karsina, 2011). In the present study, as in most research on training visual analysis, participants were asked to look at completed graphs. As there are many situations in which individuals evaluate the effects of an intervention after data are collected (e.g., reviewing manuscripts, reading published journal articles, and evaluating data in a research presentation), instruction on visually analyzing both complete and incomplete datasets is valuable.

Finally, we only evaluated the effectiveness of the trainings with undergraduates. Other populations, such as teachers, practitioners, and graduate students, may respond differently. Although critical differences between these populations and our undergraduate participants with regard to the trainings should be identified empirically, it is possible that undergraduates have a more immediate history of learning through lectures than teachers or graduate students, or that teachers or graduate students may be more invested in learning to interpret graphs. In other words, we cannot generalize the results of this study to populations other than those with similar demographic characteristics, and future researchers should evaluate the effectiveness of both interventions with other populations.
Future Research

These results highlight several avenues for future research on training individuals to conduct visual analysis of behavioral data. Across both interventions, the outcomes on slope estimation might be improved by providing participants with additional heuristics or more concrete steps for estimating the slope of data paths with multiple trends and/or outliers. For example, participants could be taught to consider whether data points are part of a larger trend or whether they are outliers when estimating the slope of a data path. In the present study, we developed a visual discrimination approach that provided general guidelines for estimating slope and level and multiple examples and nonexamples of types of changes. Computational methods (e.g., the split-middle method) may produce more reliable estimates but also have validity concerns. Researchers could evaluate the reliability, validity (compared with expert opinion), and usability of different approaches to slope estimation, as well as whether training in one (e.g., split-middle method) eventually facilitates use of another, more practical approach (e.g., visual).

Specific to the lecture intervention, researchers may conduct component analyses to discover any relatively inactive elements of the intervention. These could be eliminated to create a more efficient training. It is possible that a live lecture would be more effective than the videotaped format evaluated in this study because the instructor could include active response opportunities; future researchers could investigate this question.

With regard to the CBT, the aforementioned changes to the remediation techniques (e.g., allowing more errors before remediation and providing specific instruction based on the type of error committed), practice items (e.g., including more
difficult graphs and changing the response requirement), and delivery (e.g., allowing individuals to complete the sections on consecutive days rather than in one sitting) could be evaluated empirically to determine whether any of these modifications improve participant accuracy on the posttest compared to the CBT modules implemented in this project.

As previously discussed, it will be crucial to evaluate whether the gains observed in the immediate posttest for both intervention groups maintain over an extended period of time, and whether they generalize to actual data. In addition, the trainings could be extended to provide instruction on analyzing data in true experimental designs (i.e., reversal and multiple baseline designs). The initial modules of the CBT, and the videotaped lecture, provide a foundation in the components of visual analysis, but do not teach participants to identify when a functional effect has been demonstrated. Additional modules and/or lecture content could be developed to address visual inspection in reversal designs (e.g., replication of the effect) and multiple baseline designs (e.g., vertical analysis).

**Conclusion**

Reliable visual analysis is necessary for valid interpretation of individual single-subject studies and systematic reviews of multiple studies. To date, research on agreement among visual analysts has generally found poor interrater reliability, even among experts. Although this poor reliability suggests that training in visual analysis should be systematic and research-based, few studies have been published on training individuals to conduct visual analysis.
This study contributes to the small body of literature on training visual analysis in several ways. Although its limitations must be acknowledged, the assessment in the present study may be a more valid tool for evaluating the effectiveness of training programs in visual analysis given the inclusion of slope and a reliance on unanimous expert opinion as the accuracy criterion. In addition, no previous training studies separately evaluated discriminations of slope and level change.

The trainings in the present study focused on producing accurate, practical and flexible discriminations of slope and level change through explicit instruction and opportunities to practice. Unlike previous training research, we did not provide participants with criterion lines or structured criteria. The trainings were relatively brief, easy to deliver, and produced sizeable mean gains on the posttest. Though there is still substantial room for improving the effectiveness of the trainings, these initial results are encouraging for both the lecture and the CBT. The item analysis and secondary dependent variables highlight potentially important starting points for improving the trainings. With continued refinement, both interventions have the potential for dissemination to graduate-level training programs for aspiring BCBAs and researchers. In addition, the trainings could be modified and implemented with teachers to improve data-based decision making.

The study also raises numerous questions for future researchers who wish to investigate the critical features of effective instruction in visual analysis. For example, what is the most effective, efficient, and valid method of estimating the slope of a data path, and how can we best train individuals to use that method? Studies examining these
and related research questions can inform the development of more systematic training approaches to improve the reliability and validity of visual analysis.
REFERENCES


APPENDICES
Appendix A

Treatment Acceptability Questionnaire
APPENDIX A

Treatment Acceptability Questionnaire

Training Visual Analysis
Treatment Acceptability Questionnaire—Computer-Based Training

Please answer the following questions. Do not put your name on this sheet.

On a scale of 1 (strongly disagree) to 5 (strongly agree), please rate how much you agree with the following two statements:

1. we enjoyed participating in the training.

   1 2 3 4 5
   Strongly disagree Disagree Neutral Agree Strongly agree

2. we learned a lot from the training about how to interpret data that is presented in graphs.

   1 2 3 4 5
   Strongly disagree Disagree Neutral Agree Strongly agree

3. What could be done to make the training more engaging?
Appendix B

Informed Consent
Appendix B

Informed Consent
CURRICULUM VITA

Katie Snyder, MS BCBA

359 S 100 E                                                                                                  katiesnyder@gmail.com
Smithfield, UT 84335                                                                                                (412) 508-0424

Education
Utah State University  Ph.D, expected May 2013, Disability Disciplines: Applied Behavior Analysis
Mercy College M.S. with distinction, 2005, Elementary Education
University of Virginia B.A. with distinction, 2002, History (Biology Minor)

Professional Certification
Board Certified Behavior Analyst, September 2011-present, Certificant #1-11-9737

University Teaching Experience
Spring 2012: Instructor, Department of Special Education and Rehabilitation, Utah State University
SPED 5060 - Consulting with Parents and Teachers (Undergraduate)

Fall 2010: Teaching Assistant, Department of Special Education and Rehabilitation, Utah State University
SPED 6720 - Educational Applications of Behavior Analysis (Graduate)

Spring 2010: Teaching Assistant, Department of Special Education and Rehabilitation, Utah State University
SPED 6710 - Concepts and Principles of Behavior Analysis in Education (Graduate)

Fall 2009: Teaching Assistant, Department of Special Education and Rehabilitation, Utah State University
SPED 5010 - Applied Behavior Analysis I (Undergraduate)

University Supervisory Experience
Fall 2012, Fall 2011: Supervisor, Department of Special Education and Rehabilitation, Utah State University
SPED 5410 – Practicum: Direct Instruction Reading and Language Arts for Students with Disabilities

Spring 2012: Supervisor, Department of Special Education and Rehabilitation, Utah State University
SPED 5410 – Practicum: Teaching Mathematics to Students with Mild/Moderate Disabilities

Publications


Manuscripts Under Review

Professional Presentations
Mace, F.C., DeLeon, I.G., Dixon, M.R., & Reed, D.D. (2012, May). In K. Snyder (Chair), *Bridging the gap between basic and applied research.* Panel discussion conducted at the 38th annual conference of the Association for Behavior Analysis International, Seattle, WA.


Poster Presentations

Workshops


Kelley, K.N., Snyder, K., & Higbee, T.S. (October, 2010). Effective teaching strategies. Workshop presented to Preschool Special Education Teachers, Paraprofessionals, and Related Service Providers at Utah’s Early Childhood Conference, Provo, UT.

Snyder, K. & Higbee, T.S. (June, 2010). Identifying reinforcers for individuals with disabilities. Workshop presented at the 8th annual Utah Conference on Effective Practices in Special Education and Rehabilitation, Logan UT.


Grants – Independently Written


Grants – Assisted in Preparation

Professional Experience
August 2011 – Present: Independent Behavioral Consultant, Logan, UT

August 2009 – March 2011: Graduate Case Manager, ASSERT Preschool Program, Utah State University, Logan, UT

August 2006 – June 2009: Behavior Specialist Consultant & Mobile Therapist, PLEA Agency, Pittsburgh, PA


September 2003 – June 2005: Special Education Classroom Teacher, PS 176X, Bronx, NY

Professional Memberships
Association for Behavior Analysis International
California Association for Behavior Analysis
Council for Exceptional Children
Division for Early Childhood
Division on Autism and Developmental Disabilities
Utah Association for Behavior Analysis

Professional Service
Student Editorial Board Member, Young Exceptional Children, January 2012 – May 2013

Reviewer for National Professional Development Center Autism Evidence-Based Practice Update, 2012

Student Representative to Association for Behavior Analysis International, 2011-2012 Academic Year

Reviewer for Journal of Applied Behavior Analysis, Young Exceptional Children, Behavior Analysis in Practice

Awards and Honors
Utah Regional Leaders in Neurodevelopmental Disabilities Trainee, 2010 – 2011
University of Utah & Utah State University
Vice President for Research Fellowship, August 2009 – May 2010
Utah State University

New York City Teaching Fellowship, June 2003 – May 2005
New York City Department of Education