

HIERARCHICAL LINEAR MODELING AGAINST THE “GOLD STANDARD”  
OF VISUAL ANALYSIS IN SINGLE-SUBJECT DESIGN

A Thesis

Submitted to the Graduate Faculty of the  
Louisiana State University and  
Agricultural and Mechanical College  
in partial fulfillment of the  
requirements for the degree of  
Master of Arts

in

The Department of Psychology

by  
Elizabeth S. Godbold  
B.A., Wake Forest University, 2005  
May 2008

## TABLE OF CONTENTS

List of Tables.....	iii
List of Figures.....	iv
Abstract.....	v
Introduction.....	1
Visual Analysis Research.....	3
Statistical Analysis Research.....	13
Problems and Limitations of Previous Research.....	18
Hierarchical Linear Modeling.....	23
Receiver Operating Characteristic Curves.....	24
Method.....	28
Graph Selection.....	28
Visual Analysis Survey.....	29
Participants.....	29
Materials and Procedure.....	31
Hierarchical Linear Modeling.....	33
Template Overlay.....	33
Data Analysis.....	34
Receiver Operating Characteristic Curves and Contingency Probability Tables.....	36
Results.....	41
Visual Analysis Survey.....	41
Demographics.....	41
Agreement.....	42
Ratings.....	42
Hierarchical Linear Modeling.....	42
Receiver Operating Characteristic Curves and Contingency Probability Tables.....	44
Discussion.....	51
Visual Analysis Survey.....	51
Hierarchical Linear Modeling.....	52
Receiver Operating Characteristic Curves.....	53
References.....	55
Vita.....	58

## LIST OF TABLES

1. Sample Graphs by Type and Author Statements of Certainty.....	30
2. Likert Scale Values and Corresponding Average Rating Ranges.....	38
3. ROC Curve Dichotomies for Visual Analysis and HLM.....	39
4. Percentage of Graphs within each Rating Range.....	43
5. AUCs and Significance Values for ROC Curves.....	45
6. Contingency Probability Table.....	50

## LIST OF FIGURES

1. Example survey graph with definitions.....	32
2. Graph questions.....	33
3. ROC curve of level: visual analysis ratings of 2 and higher vs. HLM.....	46
4. ROC curve of level: visual analysis ratings of 3 and higher vs. HLM.....	46
5. ROC curve of level: visual analysis ratings of 4 and higher vs. HLM .....	47
6. ROC curve of trend: visual analysis ratings of 2 and higher vs. HLM .....	47
7. ROC curve of trend: visual analysis ratings of 3 and higher vs. HLM .....	48
8. ROC curve of trend: visual analysis ratings of 4 and higher vs. HLM .....	48
9. ROC curve of the graph as a whole: visual analysis ratings of 2 and higher vs. HLM .....	49
10. ROC curve of the graph as a whole: visual analysis ratings of 3 and higher vs. HLM .....	49
11. ROC curve of the graph as a whole: visual analysis ratings of 4 and higher vs. HLM.....	50

## **ABSTRACT**

Visual analysis is the “Gold Standard” for single-subject data because of two assumptions: a low Type I error rate and consistency across raters. However, research has shown it less reliable and accurate than desired. Autocorrelation, variability, trend, lack of obvious mean shift, and differences in the physical presentation of graphs contribute to inconsistencies and higher error rates. Statistical analysis has been advocated as a judgmental aid to visual analysis, but an appropriate statistic has not been found. In the present study, the accuracy of Hierarchical Linear Modeling was compared to raters’ visual analysis of previously published data using Receiver Operating Characteristic curves. The statistic was established as a potentially useful judgmental aid; however, definite conclusions were hindered by low power.

## INTRODUCTION

Single-subject design exposes one individual (or several individuals, each serving as their own control) to different conditions of an independent variable and compares changes in a dependent variable across those conditions. Most designs begin with a baseline condition where the dependent variable is measured in the absence of any manipulation of the independent variable. Then, subsequent intervention phases involve some implementation of the independent variable while continuing measurement of the dependent variable (Kazdin, 1982). The baseline condition serves as a way to predict how the dependent variable would continue to appear if no independent variable was introduced. A comparison can be made between the baseline prediction and the intervention phase to see if the dependent variable continues in the same manner as the baseline phase or if it changes in some way with the manipulation of the independent variable. Most researchers and practitioners using single-subject designs visually analyze their data to determine if the dependent variable changes significantly across conditions. Researchers graph the data, and believe that if a change occurred, it should be obvious and noticeable in graph form (Kazdin, 1982).

Baer (1977) put forth the theoretical argument that visual analysis is superior to statistical analysis of single-subject data because of the different error rates associated with each analysis. The strength of visual analysis, for Baer and many other researchers, is the presumed conservative nature of visual analysis. Because changes in the dependent variable must be obvious and noticeable in graph form, researchers inspecting graphed data are less likely to notice subtle changes and should only conclude that dramatic differences are significant. This conservativeness would equate to lower Type I error, as researchers would be less likely to conclude incorrectly there was a significant change because they would only notice large,

clinically significant effects. In contrast, statistical analysis of the data might rate subtle changes as clinically significant even though there may have been no functional change in the dependent variable, leading to higher Type I error. Although the true Type I error rate cannot be calculated for single-subject designs, Baer states that the rate is “clearly much smaller than 0.05” – the accepted rate for statistical analysis of effectiveness (1977, p. 169). Baer does concede that a lower rate of Type I error would lead to a higher rate of Type II error (concluding there was no effect of the independent variable when there really was an effect), but Baer considers this a strength of visual analysis because any independent variables found to have an effect on behavior would be generally more “powerful, general, and dependable” than potentially weaker independent variables rated effective by statistical analysis (1977, p. 171).

Another presumption many researchers make about visual analysis is the consistency of judgments across different raters analyzing the same data set. If an effect is obviously present or obviously absent, raters should be able to agree on its presence or absence. These two assumptions – that visual analysis has a lower rate of Type I error and shows consistency across raters and applications – has led to it becoming the “Gold Standard” for analyzing single-subject data. However, this acceptance of visual analysis as the optimal method is not because it is truly more reliable or error free. Instead, it is simply the best method presently available to the field. In fact, studies into the consistency of visual analysis have found several factors inherent to single-subject data that lead to inconsistent visual analysis decisions – the presence of autocorrelation, variability, and trend in the data, as well as a lack of obvious mean shift across conditions. Other lines of research have found that different graphing techniques can also cause inconsistent visual analysis. In addition, researchers have attempted to estimate the true rate of

Type I error associated with visual analysis and found it to be potentially much higher than Baer originally assumed.

### **Visual Analysis Research**

Jones, Weinrott, and Vaught (1978) conducted the first study on the accuracy of visual analysis judgments. They compared the accuracy of visual analysis of significance with a statistical test of significance, time series analysis, using agreement indices. The authors also concentrated on the relationship between autocorrelation and the accuracy of the two analyses. Autocorrelation is the ability to predict a data point value from the data point immediately preceding it. Autocorrelation is almost always present in single-subject data (Busk & Marascuilo, 1988) and is a problem for many statistics because it violates the assumptions of independence on which they are based. Because autocorrelation is problematic for many statistical analyses, the authors wanted to determine if it was also a problem for visual analysis.

A non-random sample of graphs with “non-obvious” effects was selected from the *Journal of Applied Behavior Analysis*. Eleven judges experienced in visual analysis were asked to rate the meaningfulness of change in level across fifty-eight adjacent phases using the categories: “yes,” “no,” and “unsure.” Lag 1 autocorrelations were then calculated for each graph and placed into three categories: “low,” “moderate,” and “high.” The graphs were also tested using time-series analysis, a statistic that can accommodate the presence of autocorrelation and tests for changes in level or trend across phases. Statistical significance was categorized using the generated  $p$  values: “significant ( $p < 0.05$ ),” “nonsignificant ( $p > 0.10$ ),” and “unsure ( $0.05 < p < 0.10$ ),” a category discarded from further analysis as it was considered an ambiguous classification.

For the agreement indices, the researchers compared the inter-rater agreement (IRA) of the panel of judges. Calculated using the formula for agreement proportions, ( $P_A = A / A+D$ ), IRA ranged from 0.04 to 0.79 for each graph, with a median of  $P_A = 0.39$  – a conventionally unacceptable level of agreement, and one demonstrating the inconsistency of visual analysis. The second agreement index compared the agreement between each individual judge and the statistical classifications. Agreement was defined as a rater and the time series analysis both finding a significant change in level (“yes” and “ $p < 0.05$ ”). Raters classifying effects as “no” or “unsure” matched with the time series analysis classification “ $p > 0.10$ .” For this index, the best individual agreement between a judge and time series analysis was  $P_A = 0.65$  (average agreement was  $P_A = 0.50$ ), indicating that each judge agreed with time series analysis at or below chance levels. Therefore, not only did the group of raters disagree with each other at a high rate, time series analysis and visual analysis also disagreed at a high level.

In investigating the effects of autocorrelation, the researchers found that the agreement between time series analysis and visual analysis was inversely related to the presence of autocorrelation – the more autocorrelation present in the data (shown by a “moderate” or “high” autocorrelation ranking), the lower the agreement level ( $P_A = 0.50$  between visual and time series analysis for graphs with the highest level of autocorrelation). However, when there was little autocorrelation in the data, agreement rose to  $P_A = 0.73$ . Time series and visual analysis agreed most when there was low autocorrelation and both analyses classified the results as nonsignificant ( $P_A = 0.89$ ). The results showed that autocorrelation was a problem for statistical as well as visual analysis – when autocorrelation was present, raters tended to incorrectly view the autocorrelation as a significant change in trend when it was simply an artifact of the data. In addition, when time series found a significant effect, it was more likely to disagree with visual

analysis, demonstrating the more conservative nature of visual analysis. The authors concluded that because of the unreliability of visual analysis (as indicated by poor IRA), while statistical analysis may be less conservative, it is inherently more reliable – no matter who runs the statistical test, the result will always be the same – and advocated supplementing visual analysis with an appropriate statistical test.

A study by Matyas and Greenwood (1990) also found that both autocorrelation and variability had a negative effect on agreement levels. The authors used computer-generated graphs, which varied in respect to the amount of autocorrelation and random variability present in the data, as well as in the magnitude of effect sizes generated by the data. The amount of autocorrelation and variability present in the data was based on levels found in previously published data. Thirty-seven raters were able to respond to the data as showing no effect, a level change, a trend change, a level and trend change, or some other systematic change. Responses were split into the dichotomy of “conclusion of effect” versus “no effect.” Ratings of a change in level only or another type of systematic change fell into the “conclusion of effect” category, while those citing “no effect” were in the other. Ratings of a change in trend or a change in level and trend were considered incorrect conclusions – any change in trend was due to autocorrelation, so raters could only correctly judge the presence of an effect by citing a change in level or another systematic change.

False alarm rates (Type I errors) and miss rates (Type II errors) were calculated for all of the graphs as a function of the amount of autocorrelation and random variability in the graphs, as well as the effect size of each graph. The results showed that false alarm rates increased with higher autocorrelation and more variability and were as high as 84 percent. In contrast, there were very low miss rates (most less than 10 percent across varying levels of autocorrelation and

variability). There was a linear relationship between autocorrelation and variability, in that false alarms increased with more variability, but only when autocorrelation was also present, and autocorrelation increased false alarm rates, but only when variability was present. Overall, higher levels of autocorrelation and variability increased error rates in visual analysis. This high level of false alarms in visual analysis translates to a high level of Type I errors, and the authors concluded that perhaps visual analysis was not as conservative as generally thought. The authors proposed supplementing visual analysis with statistical analysis as a way to control false alarm rates, especially when autocorrelation or random variability was present in the data.

Ottenbacher (1990a) also investigated the accuracy of visual analysis. Using six computer-generated graphs, Ottenbacher varied the following factors across phases: mean shift, variability, slope, level, overlap, and autocorrelation. Sixty-one raters were asked to decide if there was a significant change in behavior across phases and could respond “yes,” “no,” or that they were “uncertain” about a change. Ottenbacher reported the number of raters responding in each category and found the percentage of raters responding “yes,” “no,” or “uncertain” to each graph. He then found the ratio of disagreement for each graph using the whole agreement calculation method (instead of using agreement percentages, he used disagreement percentages). Disagreement ranged from as high as 0.59 to as low as 0.08 for the graphs. Then the disagreement ratios, along with the percentage of raters classifying results as “uncertain,” were correlated with the values of the graphical features manipulated across each graph.

Variability and slope both had large positive correlations with the ratio of disagreement and degree of uncertainty associated with each graph (0.86 and 0.92 correlations with variability and slope, respectively, for disagreement, and 0.81 and 0.74, respectively, for uncertainty). The correlations indicate that the more variability and slope in the data, the more raters disagreed or

were uncertain of intervention effects. In addition, changes in mean shift and level had negative correlations with disagreement and uncertainty (-0.32 and -0.69, respectively, for variability and slope with disagreement, and -0.42 and -0.49, respectively, for uncertainty). When graphs showed large changes in mean shift or level across phases, raters were more likely to agree that effects were present. Other graphical features were only moderately correlated with disagreement and uncertainty, indicating they either did not have a large influence on raters or were overshadowed by other, more prominent features. Ottenbacher concluded that visual analysis showed unreliability in the face of certain types of data patterns, and that the technique could be helped by using statistical analysis when results were unclear or hard to interpret.

A study by DeProspero and Cohen (1979) also found that mean shift influenced visual analysis decisions, results similar to Ottenbacher (1990a). The authors used computer-generated graphs divided into different sets that had varying levels of graphical features across baseline and intervention phases. The manipulated features were changes in mean shift, changes in the magnitude of mean shift, the presence of variability, and the presence of slope. Over one hundred raters were asked to judge the experimental control demonstrated by the graphs on a scale of 1 to 100. While results indicated the raters used all of the graphical features in making their decisions, the factor affecting their decisions most was a lack of obvious mean shift – large changes in mean shift had very high ratings of experimental control, whereas smaller changes in mean shift or changes associated with more variability or trend had lower ratings. If trend and variability were constant across two different graphs, the graph with the larger mean shift received a more favorable rating.

Agreement between each pair of raters reviewing the same set of graphs was calculated using the Pearson product moment correlation. Average agreement was 0.61 – slightly above

chance. The authors also reported the range of ratings for each graph, and interestingly, the most “ideal” graph – one with a high degree of mean shift, little variability, and no slope or trend – received ratings ranging from 3 to 100 on the scale of experimental control. While the authors did not report the average agreement for each graph, it can be determined from the reported ranges that raters were fairly inconsistent in their determinations of experimental control – the widest reported range was a low of 0 and a high of 100 (covering the entire scale) whereas the smallest range was a low of 0 and a high of 16 (a more consistent result, but one found for a very “non-ideal,” obviously nonsignificant graph, which raters were consistently better at judging in other studies).

Mean shift and trend were also factors relating to rater agreement in Gibson and Ottenbacher (1988). Twenty raters were given twenty-four computer-generated graphs and asked to rate the significance of performance change across phases using a six-point Likert scale. Each graph showed different levels of various graphical features: mean shift, variability, level, slope, overlap, and autocorrelation. The results were analyzed using an interclass correlation approach (another approach to IRA) and the average interclass correlation among the raters was 0.60 – again, slightly above chance levels. Then, to compare specific features to rater judgments, ratings were dichotomized into “significant” (ratings of 3, 4, or 5 on the Likert scale) and “not significant” (ratings of 0, 1, or 2). First, ratios of disagreement for each graph were calculated in the same manner as Ottenbacher (1990a). Second, levels of “uncertainty” about data changes were calculated by determining the percentage of raters responding with a rating of 2 or 3 for a graph. Third, “confidence” was shown by raters responding to a graph with an average rating of lower than 1.5 or higher than 3.5 – meaning the raters picked, on average, a score on the Likert scale indicating a larger degree of confidence that there was or was not an effect present. All of

these measures – the ratio of disagreement, “uncertainty,” and “confidence” – were then correlated with the values of the manipulated graphical features using the interclass correlation approach.

Results were analogous to similar studies (such as Ottenbacher, 1990a). Higher ratios of disagreement correlated positively with higher levels of trend – the more trend, the more raters disagreed on intervention effects. Negative correlations were found between the ratios of disagreement and changes in mean shift and level – the more obvious the changes in mean and level, the more raters agreed on an effect. For these three features, the same correlation patterns were found for the measures of uncertainty: positive correlations between uncertainty and slope, and negative correlations between uncertainty and mean shift and level. For the confidence measure, higher levels of confidence correlated with larger mean shifts and larger changes in level. Changes in trend correlated with lower levels of confidence.

Overall, these five studies have shown that visual analysis is weakened by the presence of autocorrelated data, data that is highly variable, or data with a definite trend. Data with an obvious change in mean or level is helpful to visual analysis, but those positive factors are dampened if variability and trend are also present. All of the factors, autocorrelation, variability, mean level, and trend, are present in most single-subject data and yet influence raters differently, as shown by poor IRA across all studies (the median IRA across the studies was 0.63). Even studies where “ideal” graphs were used – those unlikely to be found in anything other than highly controlled research settings, with low variability, no trend, and large mean shifts across conditions – showed unreliability in visual analysis ratings across participants. Another line of research has looked into the physical features of graphs and found that how a graph is physically presented can also influence raters’ judgments of intervention effects.

Knapp (1983) found that several graphical features could change visual analysis judgments. Knapp created 135 graphs using three different graphing techniques (cumulative plot, log, or frequency polygon) and three different styles of presentation (baseline separated from intervention data by a space, vertical line, or connected). Graphs could also differ on the amount of mean shift across the baseline and intervention phase, with nine possible levels of mean shift ranging from high to low. No graph showed any trend or autocorrelation. Raters were asked to judge if a change occurred across the phases using the choices “yes” or “no,” and they were asked not to refer back to previously rated graphs. Each graph set was presented to each rater three times. Raters agreed with themselves, i.e., gave the same graph the same rating, 79 percent of the time. Higher levels of mean shift were associated with more intra-rater consistency. When mean shift was lower, ratings were less consistent, and graphing technique and presentation style had a significant effect on rater judgments. With lower mean shift, ratings differed across the type of graph used as well as the type of separation between baseline and intervention phases, even when the data itself did not change. In fact, raters were most likely to say a change had occurred across the phases if there was an obvious physical feature separating the phases, such as a vertical line. Knapp concluded that when data had a non-obvious mean shift, changing the graphing technique or presentation style changed the visual analysis judgment.

Other graphical features were investigated in two studies outlined in a 1998 article by Fisch. The first, by Greenspan and Fisch (as cited in Fisch, 1998), varied the number of data points in baseline and intervention phases and asked raters to judge the change across phases. Graphs either had five data points in each phase, ten in each phase, five in the baseline phase and ten in the intervention phase, or vice versa. The data were the same across the same phase types

except for the number of data points (baseline and intervention phases with only five data points simply had the last five points from ten point baseline and intervention phases removed). The raters' judgments could be based on changes in level, trend, level and trend, neither level nor trend, or another type of systematic change. False alarm rates (Type I errors) and miss rates (Type II errors) were calculated for all of the graphs. Across differing numbers of data points, raters showed higher levels of accuracy in detecting change for the graphs with unequal data points across phases. Raters were worst at identifying changes in level or trend in graphs with ten data points per phase.

Another study by Fisch and Schneider (as cited in Fisch, 1998) changed where the dataset was physically placed on the graph. The data could be located toward the top (away from the x-axis), the bottom (close to the x-axis), or in the middle of the graph. Raters had a higher number of correct responses when the dataset was placed closer to the top or bottom of the graph, even across different types of dependent variables. Data placed nearest the x-axis showed the highest proportion of correct responses. Fisch and Schneider concluded the frame of the graph provided an anchor on which the raters based their decisions, allowing them to detect changes more easily.

A further line of research was prompted because some studies found visual analysis to be more conservative than statistical analysis while other studies found the opposite result, leading to a study into the true rate of Type I error in visual analysis. Matyas and Greenwood (1990), previously discussed, attempted to estimate the Type I error rate of visual analysis and found it to range between 16 percent and 84 percent when autocorrelation and variability were present in the data (as based on the false alarm rate). However, when there was little autocorrelation in the data, Type I error was as little as 0 to 13 percent. Allison, Franklin, and Heshka (1992) used this estimate of Type I error as the basis for a study into the true amount of Type I error associated

with visual analysis. Whereas Baer (1977) assumed visual analysis was a more conservative approach to judging treatment effects (with a Type I error rate of less than 5 percent), Allison, et al., felt the way many researchers approach visual analysis inherently inflates the risk of Type I error. Many researchers graph each data point as it becomes available and then base decisions on whether to continue, discontinue, or change the intervention on these graphs. Named “Response-Guided Experimentation” by Edgington, the authors felt this approach would inflate Type I error.

To estimate this inflated Type I error rate, the researchers first decided on a 10 percent error rate as a conservative estimate of error in data with little autocorrelation or variability (again, based on the range of error rates for low-autocorrelated data found in Matyas and Greenwood). Then, the authors decided that a researcher making a decision about his data (that is, using Response-Guided Experimentation) over the course of ten data points might inspect the data at every other data point. Therefore, the researcher would make a decision about the treatment five times over the course of the ten data points. Based on Allison, et al.’s conservative estimate, the researcher would have a 10 percent potential error rate for each time a decision was made. Because the researcher could possibly make an incorrect decision five times total, with a 10 percent chance of an error each time, the real rate of Type I error across the entire dataset would increase to 25.9 percent. This Type I error rate given by the authors is much higher than the original “conservative” estimate provided by Matyas and Greenwood (10 percent) and even higher than the error rate touted by Baer (less than 5 percent). While acknowledging the estimate is simply plausible, and not a concrete pronouncement of the Type I error rate associated with visual analysis, the authors were still able to show that the Type I error rate may not be as small as originally believed.

Based on the above research, under certain conditions visual analysis is neither as reliable nor as accurate as desired. Factors frequently present in single-subject data – autocorrelation, variability, trend, and a lack of obvious mean shift – all contribute to inconsistent visual analysis. Visual analysis can be influenced by extraneous factors, like the physical presentation of data, and error rates may be much higher than would be acceptable to most researchers – perhaps as high as 25 percent. These considerations have led many researchers to advocate the use of a judgmental aid when conducting visual analyses of data. Judgmental aids are “stimulus-simplifying techniques and their products” that can supplement each other and give researchers and practitioners more confidence in their decisions (Michael, 1974, p. 647).

Statistical tests are an appropriate judgmental aid to researchers using visual analysis, as they are perfectly reliable and consistent, no matter who is running the test, and they have a known level of Type I error. They are also free from extraneous influences. Consequently, visual analysis and statistical analysis could both be considered judgmental aids for researchers and practitioners, with each receiving balanced consideration. Visual analysis is the aid of experimenter judgment. Using visual analysis, researchers can determine if the intervention effect is observable while staying “close” to the dataset and the research participant. Statistical analysis, in turn, is less experimenter-based, ridding decisions of the biases and extraneous influences associated with visual analysis. Statistical analysis ensures the reliability and consistency of decisions, especially in settings where complete control is impractical or data is unclear and a decision about an intervention must be made regardless.

### **Statistical Analysis Research**

Using statistical analysis as a judgmental aid to visual analysis has been, as previously stated, advocated by many researchers and several studies have investigated different statistics

that could be used with single-subject data. The Jones, et al., study (1978) outlined previously tested the utility of time series analysis as a supplement to visual analysis. However, the authors found time series analysis only agreed, at best, with visual analysis judgments 65 percent of the time. In fact, across all of the raters in the study, time series analysis showed better agreement with visual analysis when changes in the data were deemed nonsignificant by both tests. However, if raters are able to see clearly that the data did not change significantly, they really have no need for a supplement statistical test.

Ottensbacher (1990b) tested the Split-Middle Trend statistical analysis against raters' visual analysis judgments. The Split-Middle Trend is a celeration line approach generating a trend line based on the first data phase, normally baseline. If there is a change in performance across the phases, the proportion of data points above and below the line in the treatment phase will be different from the proportion above and below the line in the baseline phase. While not a strictly statistical test, the difference in proportion is compared to a statistical probability estimate of that difference occurring. Ottensbacher conducted this study in the same manner as his other 1990 study and his 1988 study with Gibson. He found that while IRA for the visual analysis judgments was near chance (an IRA consistent with other research), the statistic agreed with visual analysis slightly less than chance – 0.46 (calculated using the point-to-point method). Ottensbacher also used Contingency Probabilities and found visual analysis, as compared to the Split-Middle Trend, had a sensitivity of 0.50 and a specificity of 0.38 – both very low values.

Park, Marascuilo, and Gaylord-Ross (1990) tested the agreement between visual analysis judgments and Edgington's Randomization test. The authors presented forty-four randomly selected graphs from the *Journal of Applied Behavior Analysis* to five raters. The raters were instructed to judge the significance of change across phases in the graphs using the categories

“significant,” “nonsignificant,” and “unclear.” Inter-rater agreement was an average of  $P_A = 0.60$  (again, slightly above chance, but consistent with previous findings). Overall, the raters detected 48 percent of the significant effects present in the graphs. This finding surprised the authors because all of the phase changes had been previously found to be significant enough to warrant publication.

Fifteen of the graphs with a sufficient number of data points were subjected to a Randomization test. Randomization tests require that the intervention phase is begun at a random point in a study, but the introduction of the intervention is limited to a random point that ensures at least five baseline data values are collected while guaranteeing at least five intervention values will be collected (so in a study measuring the dependent variable twenty times, the intervention must be started within the interval of the sixth and sixteenth measurement). The test calculates the actual mean difference between baseline and intervention phases as well as all possible permutations of mean differences (i.e., all mean differences that could be found if the intervention was introduced at every other possible point within the interval of the sixth and sixteenth measurement). Then, each possible mean difference equal to or larger than the actual mean difference is tallied (including the actual mean difference) and divided by the total number of potential mean differences (in the twenty data point example, there are ten possible mean differences and one actual mean difference, for a total of eleven). The resulting number gives a  $p$  value which indicates the probability the actual mean difference found was due to chance (Edgington, 1992).

Agreement between rater judgments of these graphs and the Randomization test was  $P_A = 0.80$ . Most of the agreement stemmed from agreement about the non-significance of phase changes (67 percent). Agreement about the significance of phase changes was less important to

the overall agreement rate (13 percent). However, IRA for these fifteen graphs was  $P_A = 0.81$ . Five raters agreed on the tested graphs as a group at the same rate as the randomization test agreed with the raters – so it seems that instead of adding a statistical test to the data, a researcher could simply ask four colleagues to visually analyze a dataset. Both methods would have the same chance of agreement with the researcher’s original decision.

Stocks and Williams (1995) studied the accuracy of  $t$  tests of mean differences and piecewise regression against visual analysis judgments of over two hundred raters. The authors fabricated graphs with and without celeration lines and compared raters’ judgments of a systematic change across the phases to the appropriate statistical test of those phases ( $t$  tests when no trend was present in the data according to a linear regression and piecewise regression when there was trend present). Visual analysis judgments were based on a 10 percent probability that the change was due to error. Because the graphs had been intentionally created to show the presence or absence of systematic changes, the authors were able to compare the classification accuracy of rater judgments and the statistical tests. Overall, the statistics were better at classifying significant effects than raters, with raters being poor at detecting even large changes across phases (“large” defined as a difference of more than one standard deviation). However, when raters were given celeration lines to help their decisions, the advantage of statistical analysis all but disappeared.

A study by Ma (2006) attempted to develop a statistic similar to the Percentage of Non-overlapping Data (PND) approach that overcomes some of the limitations of PND calculations. Ma’s Percentage of Data Points Exceeding the Median of Baseline Phase (PEM) uses the median baseline data point to create a line where data values have an equal chance of falling above and below the line and extends the line into the treatment phase. Then, the percentage of data points

in the intervention phase falling above (or below) the line can be calculated. If the percentage deviates significantly from 50 percent, there is a significant effect of treatment, and effect sizes can be generated from the deviation. To test the viability of the PEM approach, Ma used previously published data from multiple behavioral journals and coded the original authors' analysis of graphs into three categories – a “moderate effect,” “noticeable effect,” or “little effect or improvement,” and correlated the findings of the PEM approach with these original visual analysis judgments. Using the Spearman correlation with the change across a pair of phases as the unit of analysis, the PEM approach correlated with original judgments 0.57. While PEM performed slightly better than PND (with a correlation of .49), the correlation is not high enough to warrant PEM taking the place of visual analysis, especially when basing important decisions on data judgments.

A 2006 study by Brossart, Parker, Olson, and Mahadevan attempted to create guidelines for using several different types of statistical tests as a supplement to visual analysis. The authors had fifteen raters make decisions about the effectiveness of interventions using computer-generated graphs. The raters based their decision on a five-point Likert scale. For each graph, an average rating of 1.0-2.9 was considered “Not Effective,” ratings of 3.0-3.5 were considered “Somewhat Effective,” and an average rating of 3.6-5.0 was considered “Very Effective.” Inter-rater agreement was calculated using inter-item analysis, with the raters as “items,” and was 0.89. Agreement between individual raters and the entire group was an average of 0.58, a result analogous to previous research.

Statistics tested included the Binomial Test on Extended Phase A Baseline – a measure of growth, the Last Treatment Day test – a measure of final level, Gorsuch's Trend Effect Size – a test of mean differences that controls for “expected” growth, the Center Mean plus Trend Model

– a test of mean differences that partials out overall trend, and Allison, et al.’s Mean plus Trend Model – a test of mean differences and growth over the intervention phase only. Effect sizes were generated for each graph using each statistic. The effect sizes generated by each statistic were then divided into three categories based on the visual analysis judgments. The first category contained the effect sizes for graphs rated “Not Effective” by raters, the second contained the effect sizes associated with “Somewhat Effective” ratings, and the third contained the “Very Effective” graphs. In this way, the authors were able to establish the effect sizes for each statistic that should be expected for a graph demonstrating a particular level of effectiveness as judged by visual analysis. Using these guidelines, data subjected to the statistics would generate effect sizes and researchers would know what general level of visual analysis would correspond to that effect size, without necessarily needing a high level of experience and ability in visual analyzing data. In addition, researchers unsure of their visual analysis decisions would be able to generate an effect size and see if it agreed with their decision about the significance of change across phases.

### **Problems and Limitations of Previous Research**

These studies demonstrate that researchers have attempted to identify statistics that could be useful supplements to visual analysis, a clear need in single-subject research. However, while many of the results were promising, all of the studies had limitations. Some of the studies tested statistics inappropriate or infeasible for single-subject data. Others had limitations inherent to the studies themselves. In addition, the research outlining the inconsistencies and problems associated with visual analysis also contains several problems and limitations. Most of these problems are design issues that could be improved. The following is a list of the problems and

limitations of the previous studies into visual analysis, followed by an explanation of the problems and limitations of research into potential statistical analyses.

Most of the research into visual analysis used fabricated, computer-generated data (DeProspero & Cohen, 1979; Gibson & Ottenbacher, 1988; Knapp, 1983; Matyas & Greenwood, 1990; Ottenbacher, 1990a). The authors chose to generate graphs fitting their conceptions of various data types – either data varied across different parameters, such as variability and level of mean shift, or data featuring only one type of parameter change, such as graphs showing a change in trend only (i.e., Fisch, 1998; Matyas & Greenwood, 1990). Most of these fabricated graphs also included little to no context for the data; in fact, several raters in the DeProspero and Cohen study (1979) refused to participate as they felt the rating task was “inappropriate” without context. Many of the studies, those using fabricated graphs and those using previously published data, used AB graphs with only one baseline and one intervention phase (Gibson & Ottenbacher, 1988; Knapp, 1983; Matyas & Greenwood, 1990). Most of the researchers justified the use of AB graphs because they argued that the basic AB design was the “building block” for all other research designs. However, when combining computer-generated data with basic AB designs, raters ended up judging graphs rare in real-world settings.

Sample sizes varied across all of studies. For the rater sample sizes, there was a range of 11 to 108 raters, with a median of 36. For the graph sample sizes, there was a range of 6 to 147 graphs, with a median of 24. Several of the studies, including the Ottenbacher (1990a) study that used only six graphs total, were limited to samples smaller than many researchers would desire.

In addition to using smaller sample sizes, some of the studies deliberately used inexperienced raters (Gibson & Ottenbacher, 1988; Ottenbacher, 1990a). Both of these studies discussed the use of inexperienced raters as a limitation of their findings, but argued that general

practitioners of visual analysis would probably have the same level of experience with visual analysis judgments as their raters. The authors make a reasonable argument; however, when trying to prove or disprove the accuracy of visual analysis, raters experienced in making these judgments would be the most sound group to sample. In contrast, the Jones, et al. (1978) study used experienced raters, but deliberately chose previously published graphs with “non-obvious” effects (the authors’ own words). In trying to prove or disprove the accuracy of visual analysis, a better graph sample would be representative and include obvious and non-obvious effects, instead of deliberately making rater judgments difficult.

Several of the studies showed ambiguity in the type of response required from raters. Terms were rarely defined. Some studies asked raters to judge graphs based on discrete categories of “yes,” “no,” “unsure,” “significant,” “nonsignificant,” while others relied on Likert scales. DeProspero and Cohen (1979) used a 100-point scale with no guidelines on which raters could base their judgments. In addition, the actual judgment question varied across the studies. Jones, et al. (1978) asked participants to rate if there were “meaningful changes” across phases, a consideration involving social validity, while DeProspero and Cohen (1979) asked participants to rate the graphs’ “demonstration of experimental control.” Knapp (1983) required raters to simply judge if “a change occurred,” while Gibson and Ottenbacher (1988), Ottenbacher (1990a), and Park, et al. (1990) asked raters if a “significant change in performance” occurred across phases. Stocks and Williams (1995) asked how “reasonably certain” raters were that a systematic change occurred across phases, and defined “reasonably certain” as a 10 percent probability that the change was due to error. These researchers often effectively gave raters a task analogous to statistical judgments of change, a decision visual analysts rarely, if ever, consider. Taken as a whole, the problems with previous visual analysis research can be

summarized as follows: previous studies used deliberately manipulated AB graphs of fabricated, de-contextualized data while giving often-inexperienced raters ambiguous instructions on rating the “meaningfulness” or “significance” of change.

Many of the studies into potential statistical analyses share the same problems as prior research into visual analysis: fabricated data (Ottenbacher, 1990a; Park et al., 1990; Stocks & Williams, 1995), AB graphs only (Brossart et al., 2006; Ottenbacher, 1990a; Park et al., 1990; Stocks & Williams, 1995), and ambiguous rater response requirements (Brossart et al., 2006; Jones et al., 1978; Park et al., 1990; Stocks & Williams, 1995). Two of the studies, Ottenbacher (1990b) and Stocks and William (1995), tested the accuracy of visual analysis against their chosen statistic, instead of the more appropriate approach of testing their statistic against the field’s “Gold Standard” of visual analysis.

The most pressing problems associated with studies into potential statistics involve the statistics themselves. Most of the statistics are either impractical or simply do not meet the assumptions of single-subject data. Others are too limited in their approach. Jones, et al.’s (1978) use of time series analysis is appropriate for single-subject data; however, it is limited by the need for 50 to 60 data points in each phase – a number rarely possible outside of strict behavioral research and very impractical in real-world settings. Randomization tests, used by Park, et al. (1990) are also fairly impractical, in that the statistic requires that the start point of interventions be randomly determined. A researcher or practitioner must be willing to accept a randomly decided time to implement an intervention. In addition, Randomization tests require at least twenty-five data points, and both the baseline and treatment phases must contain at least five data points each.

Ottensbacher's (1990b) Split-Middle Trend approach is also appropriate for single-subject data, but while it takes into account change in the proportion of data points above or below the trend line, it does not take into account the magnitude of those data points. Two graphs with the same proportion of data points above the line in the intervention phase would be rated as equal, even if one graph's data points were much further away from the trend line than the other (i.e., if they had a much greater magnitude). The same holds true for Ma's (2006) PEM test, in that it does not take magnitude of data points into account. Another problem for the PEM test is a lack of ability to consider trend or variability in the data.

The  $t$  tests and piecewise regressions used by Stocks and Williams (1995) require data with normal distributions, homogeneity of variance across phases, and the absence of autocorrelation. Stocks and Williams based their fabricated data on these assumptions, but in the context of real-world research, those assumptions would be rarely met. In addition, the two statistics are limited.  $T$  tests can only be used when there is no trend in the data, so it can only measure changes in level. While piecewise regression can consider both changes in level and trend, it is much less sensitive to changes in level than  $t$  tests.

These limitations are also problems for the statistics tested in Brossart, et al. (2006). The Binomial Test on Extended Phase A Baseline is a measure of trend only, whereas the Last Treatment Day test is a measure of final level only. Gorsuch's Trend Effect Size tests for mean differences (i.e., level) while semi-controlling for trend, and the Center Mean plus Trend Model tests for mean differences while completely controlling for trend. The Allison et al. Mean plus Trend Model tests for a change in level across baseline and treatment, but only considers trend in the treatment phase. Each of these statistics forces the user to decide whether a change in trend or level is more important to their decision-making.

## **Hierarchical Linear Modeling**

Previous research has shown the need for a statistical judgmental aid to visual analysis, and fortunately, the above limitations of previous research can be minimized through appropriate design changes. The present study attempted to lessen most of these issues, and by providing scenarios where visual analysts could be as accurate as possible, a statistic appropriate for single-subject data was examined as a potential complement to everyday visual analysis judgments. The chosen statistic, Hierarchical Linear Modeling (HLM), accommodated almost every type of single-subject design while controlling for the various assumptions of single-subject data. In addition, a more in-depth analysis using Receiver Operating Characteristic curves (ROC curves) was used to provide as much information as possible about the accuracy of HLM.

Hierarchical Linear Modeling tests for individual change over time while taking into account growth and initial level and can be easily computed using the HLM 6 Student Edition program (HLM6S, Raudenbush & Bryk, 2007). HLM was chosen for this study for several reasons. The first advantage of HLM is its ability to accommodate autocorrelation, if necessary, by testing and specifying the correct error term in the model equation. The second advantage of HLM is it tests for change over time for level and trend, and consequently, level and trend together. Other statistics previously tested against visual analysis either considered changes in just trend, just level, or an inadequate combination of the two (e.g., change in level and change in trend, but only change in trend during the intervention condition), forcing researchers to choose which factor was more important to study. In addition, HLM also accounts for the initial level of the target behavior, which is a practical feature for a statistic meant to be usable across a wide range of designs and setting. Participants receiving the same intervention but showing different initial levels of behavior can still be comparable.

The third reason to test HLM against Visual Analysis is it is usable on almost every single-subject design. The basic model used in this study could be modified to accommodate more complicated designs. In addition, HLM does not require a large number of data points.

The fourth advantage of HLM is that it can accommodate missing data or unequal intervals between measurements – another highly desirable feature. While not an overriding concern in this study, the ability to handle missing data is important for the general use of HLM. HLM can accommodate missing data if the data are assumed Missing At Random and the reason the data are missing is independent of any other data that are actually present. The Multiple Model-Based Imputation procedure can be used to estimate data values and standard errors for missing data satisfying the Missing at Random assumption (Raudenbush & Bryk, 2002).

### **Receiver Operating Characteristic Curves**

Previous research has focused on basic decision agreement between visual and statistical analyses. This study used Receiver Operating Characteristic curves and Contingency Probability tables to allow a more in-depth analysis of results. Receiver Operating Characteristic curves (ROC curves) can be calculated using a wide variety of software, such as MedCalc (Schoonjans, 2007), PASS (Hintz, 2007), and SPSS (SPSS Inc, 2007). ROC curves use a “Gold Standard” to investigate the classification accuracy of a new test and are based on dichotomies, such as “present” and “absent” or “significant” and “nonsignificant.” The curve is graphed by plotting True Positives (Sensitivity) on the y-axis and False Positives (1 - Specificity) on the x-axis. The True and False Positives are based on the classification accuracy of the new test as compared to the “Gold Standard,” giving an easy to visualize determinant of the accuracy of the test (Hopley & van Schalkwyk, 2006).

The better the ROC curve, the closer it will get to the upper left corner of the graph, meaning the test is generating more True Positives and fewer False Positives. The more useless the curve (meaning the lower its classification ability), the more closely the curve will approximate a diagonal forty-five degree line from the lower left to the upper right of the graph. A curve resembling this forty-five degree line means the test is about as accurate as random guessing in classifying test results (Langdon, 2006). For this study, the ROC curves tested the results of HLM against the results of the “Gold Standard” of visual analysis.

ROC curves were used chosen because they allow the user to determine the cost associated with any point on the curve. Cost is the level of False Positives generated by the test as a function of True Positives. As long as users are willing to tolerate a pre-specified chance of making the wrong decision (False Positives), the curve will show where the cut-off or threshold for finding a certain number of True Positives lies and how many False Positives will be incurred when using that threshold (Schoonjans, 2006).

There are multiple ways to quantify the results of a ROC curve. The first is by finding the highest point on the curve that corresponds with a set level of cost. This point is found by creating a straight line analogous to the desired cost and finding the highest point on the curve that meets this line. For an equal-cost point that gives an equal number of True Positives and False Positives, a forty-five degree line can be used, and the highest point on the curve meeting this line shows the point on the graph where the test will generate the same number of correct decisions and false alarms. Other lines can be used for finding the point where the test generates, for example, ninety percent True Positives and ten percent False Positives, or any other desired combination (Langdon, 2006).

Another way to judge the ROC curve is to use the Area Under the Curve (AUC). This number gives the percentage of tests classified correctly. The higher the curve, the more AUC, and the better the test is at classifying results. If the AUC is close to 0.5, the test is operating at chance levels (Hopley & van Schalkwyk, 2006). In addition, the MedCalc software can provide a *p* value for a given AUC indicating if the AUC is significantly different from 0.5 and is able to distinguish between groups accurately (Schoonjans, 2007).

More in-depth comparisons can be made using Contingency Probability tables. These tables allow the classification accuracy of HLM to be presented in concrete numbers and allow Overall Accuracy, as well as Positive and Negative Diagnostic Likelihood Ratios to be calculated, which are analogous to the Type I and Type II error rates used in other statistical tests. Few other studies have been able to provide a similar conclusion (for an example, see Matyas & Greenwood, 1990). The Positive Diagnostic Likelihood Ratio is the odds ratio that a significant HLM result will be observed for an intervention rated effective compared to the odds the same result will be observed for a non-effective intervention – that is, the Type I error rate that can be expected for HLM at a certain level of visual analysis (Technologies for Health Project, n.d.). The Negative Diagnostic Likelihood Ratio is the odds ratio a nonsignificant HLM result will be found for a non-effective intervention compared to the odds the same result will be found for an effective intervention – that is, the Type II error rate that can be expected for HLM at the same level of analysis (Technologies for Health Project, n.d.)

General comparisons can be made between AUCs, or the percentages of tests classified correctly if the AUCs are found to be significantly better than chance; however, it would be inappropriate to base concrete conclusions on these differences as the AUCs in this study were based on different Gold Standards because different visual analysis dichotomies were used.

Another approach to supplement comparisons between AUCs is to use the Contingency Probability tables and compare the different levels of Overall Accuracy generated by HLM. A significant difference would indicate HLM is more accurate for a particular factor or visual analysis level. However, Overall Accuracy would have to be weighed against considerations of Sensitivity and Specificity (exactly what AUCs measure), as Overall Accuracy alone can be artificially inflated by the prevalence rate of significant results for a particular dichotomy. In comparing the accuracy of HLM across different factors and visual analysis levels, differences in AUCs, Sensitivity and Specificity, and Overall Accuracy must all be considered to create the most appropriate comparison, as neither is entirely sufficient in itself (Alberg, Park, Hager, Brock, & Diener-West, 2004).

Study results were hypothesized for different levels of visual analysis and for different graphical features based on the different dichotomies analyzed using ROC curves. When using different visual analysis judgments to generate dichotomies, ROC Curve analysis was hypothesized to show HLM was an accurate test when visual analysis raters were somewhat certain of intervention effects. When visual analysis raters were reasonably or extremely certain of intervention effects, ROC curves would show HLM to be less accurate. In addition, when the accuracy of HLM was compared to visual analysis ratings of different graphical features (like level and trend), HLM would be more analogous to visual analysis ratings when analyzing each feature individually than together.

## METHOD

### Graph Selection

Previously published single-subject graphs were obtained from issues of psychological journals between January 2002 and December 2006. The graphs were selected from the following journals: *Behavioral Disorders*, *Behavior Modification*, *Child and Family Behavior Therapy*, *Journal of Applied Behavior Analysis*, *Journal of Special Education*, and *School Psychology Review*. Selecting graphs from these journals ensured the single-subject data covered areas of applied behavior analysis, clinical psychology, special education, and school psychology, making results more generalizable.

Graphs had to meet specific criteria to be included in the study. The graphs had to be legible enough to facilitate the calculation of data point values. The graphs also had to include a clear scale showing the rate or level of behavior on the y-axis and the number of sessions or time on the x-axis. Graphs with unequal scale intervals were not used. These requirements were consistent with other studies using previously published data, including Gresham, et al. (2004).

The first five graphs in each article were coded according to design type and the author's original interpretation of the graph. Almost all commonly used single-subject design types were included across single and multiple baselines: AB, reversal (ABAB), withdrawal (ABAB), multi-element designs, and "other" (ABCD, et cetera). Functional Analyses were also included, for eleven total design types. Changing Criterion designs were not included, as they do not contain a true baseline phase against which treatment effects can be compared.

The author's original interpretation of each graph, or "statement of certainty" was coded into three statement types: graphs identified as showing "unambiguous," "clear," or "certain" intervention effects ("extremely certain"), those with results considered "ambiguous," "unclear,"

or “uncertain,” (“moderately certain”) and those graphs identified as showing no effect (“not at all certain”). Inter-rater agreement was calculated for one-third of the coded graphs and was  $P_A = 0.93$ .

The complete database of graphs included 794 graphs from 268 articles. Once the database of graphs was coded, a stratified random sample was used to select graphs by design type and the author’s statement of certainty. Not every design type had a graph of each statement type (e.g., there were very few graphs coded as showing no effect, so not every possible design type was represented in this category). A power analysis run on the MedCalc program using the “demo” mode (Schoonjans, 2007), showed approximately forty graphs were needed across strata for adequate power in the overall analysis of study results. After graphs were initially selected using the strata, additional graphs were selected from some of the strata to achieve the desired level of power while also forming a representative sample of graphs closely matching the percentages of each graph type found in the database. There were thirty-nine graphs in the final sample (Table 1). While the journals from which graphs were obtained were not a strata used to select the graphs, the final sample had graphs from each and the percentages of graphs selected from each journal approximated the percentages found in the database.

### **Visual Analysis Survey**

**Participants.** An email requesting survey participation was sent to 5,074 Behavior Analyst Certification Board (BACB) members and 189 journal board members of the journals from which the graphs were obtained. BACB and journal board members were chosen because of their presumed familiarity with single-subject design and visual analysis. The participant sample was given an incentive for participation: one randomly selected participant who completed the entire survey won a fifty-dollar gift card.

Table 1

## Sample Graphs by Type and Author Statements of Certainty

Design type	<u>Author Statement of Certainty</u>			Total
	Not at all certain	Moderately certain	Extremely certain	
<hr/> Single baseline <hr/>				
AB	2.56%	2.56%	2.56%	7.69%
ABAB (reversal)	--	2.56%	5.13%	7.69%
ABAB (withdrawal)	--	2.56%	2.56%	5.13%
ABCD, etc	--	5.13%	12.82%	17.95%
Multi-element	2.56%	2.56%	5.13%	10.26%
<hr/> Multiple baseline <hr/>				
AB	2.56%	2.56%	10.26%	15.38%
ABAB (reversal)	--	--	2.56%	2.56%
ABAB (withdrawal)	--	2.56%	--	2.56%
ABCD, etc	2.56%	2.56%	5.13%	10.26%
Multi-element	--	2.56%	2.56%	5.13%
<hr/> Functional analysis <hr/>				
Functional analysis	2.56%	2.56%	10.26%	15.38%
<hr/> Total <hr/>				
Total	12.82%	28.21%	58.97%	100.00%

*Note.* Cells without data indicate no graphs in the database were of that graph type and certainty level and could not be included in the final sample. Cells with data indicate the percentages of graphs in the final sample.

**Materials and Procedure.** A visual analysis survey of the graphs was hosted by <http://www.QuestionPro.com> ("QuestionPro.com", 2007). The survey included twelve demographic questions, two questions about survey participation, and three questions about each of the thirty-nine graphs. To create the survey, each graph was reproduced using a high quality image and any definitions relevant to the graph were provided (e.g., baseline and treatment phases, dependent variables; Figure 1 provides an example). Participants had to answer each question and the graphs were randomly presented.

The three questions about each graph asked participants to judge changes in behavior in the graph based on changes in level, trend, and considering the graph as a whole. The questions appeared as follows:

Based on the information provided by the graph, how *certain* are you that the intervention(s) presented caused a change in behavior?

Considering changes in **LEVEL ONLY:**

Considering changes in **TREND ONLY:**

Considering the graph as a **WHOLE:**

For graphs showing Functional Analyses or Reinforcer Assessments, the question wording was changed to the more relevant:

Based on the information provided by the graph, how *certain* are you that the condition(s) presented distinguished a function for the target behavior?

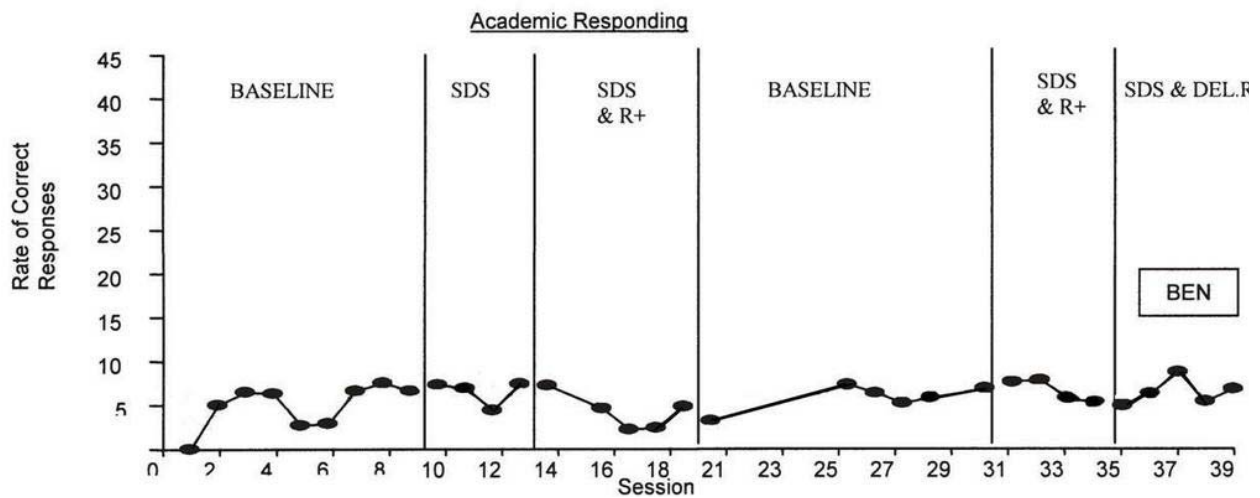
or

Based on the information provided by the graph, how *certain* are you that the assessment identified a preferred reinforcer?

The participants judged these changes using a five-point Likert scale. The Likert scale values were as follows, and the scale asked about certainty to ensure raters were not deciding on importance (a social validity decision) or significance (a statistical decision).

1. Not At All Certain
2. Somewhat Certain
3. Moderately Certain
4. Reasonably Certain
5. Extremely Certain

Figure 2 shows the question and scale as administered to participants.



**Behavior:** Digits correct on math worksheets.

**Baseline:** Completing math worksheets without any intervention.

**SDS:** Student verbalizations about fast and accurate performance before starting worksheets.

Figure 1. Example survey graph with definitions

**SDS & Del R+:** SDS & R+ condition, with the reinforcement delivered after a 1-hour delay.

Figure 1. Example survey graph with definitions

Based on the information provided by the graph, how **certain** are you that the intervention(s) presented caused a change in behavior?

	Not At All Certain	Somewhat Certain	Moderately Certain	Reasonably Certain	Extremely Certain
Considering changes in <b>LEVEL ONLY</b> *	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Considering changes in <b>TREND ONLY</b> *	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Considering the graph as a <b>WHOLE</b> *	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Figure 2. Graph questions

### Hierarchical Linear Modeling

**Template Overlay.** Data values were obtained from the graphs using a template overlay. The individual graphs were resized to fit an 8.5x11 sheet of paper with the y-axis sized to the appropriate inch or half-inch increment. The template overlay was created using a transparency sheet and marked lengthwise at each eighth of an inch. The overlays were then aligned with the zero point of the graph and the eighth-inch increments were marked starting with zero. Each data point was given the raw score value of the eighth-inch line closest to the center of the data point. These coded points were transformed back to the scale of the original graph using the following formula:

$$(Range\ of\ y\text{-axis} / Number\ of\ eighth\text{-inch}\ increments\ contained\ by\ the\ graph) * Raw\ score = Transformed\ score$$

If the graph was a multiple baseline design, each individual baseline within the overall multiple baseline graph was coded separately. Multiple dependent variables were also coded separately. The template overlay was used to ensure each graph was coded uniformly regardless of the scale used in the original graph. Transforming each raw score back to the original scale of the graph ensured the data remained accurate.

Inter-observer agreement was calculated for the graphs and was  $P_A = .96$ . One-third of the 39 graphs had 100% interobserver agreement, and no graphs had less than 90% agreement.

**Data Analysis.** Once data values were obtained for a graph, they were analyzed using a Level 2 model with the program HLM6S (Raudenbush & Bryk, 2007). The Level 1 equation focused on individual change over time. The Level 2 equation tested for change over time as a function of condition (baseline and intervention) for both level and trend. These values were allowed to vary randomly. The model below was used for basic single-baseline AB graphs, and was modified to accommodate designs beyond this level.

**Level 1 Model:**

$$Y_i = \beta + tx + e$$

$Y_i$  = individual graph,  $\beta$  = y-intercept,  $t$  = trend,  $x$  = constant (time),  $e$  = error

**Level 2 Model:**

Level:  $\beta = \text{condition} + e$

Trend:  $t = \text{condition} + e$

All of the graphs were analyzed using this Level 2 model. Multiple baseline graphs were run in sequential order (e.g., a graph with two baselines, each with one baseline and one treatment condition, was run in the order  $A_1B_1A_2B_2$ ). A more complex Level 3 model was considered to analyze the multiple baseline graphs, but was found to be incompatible with the graph types in the sample.

For this study, the initial level of behavior ( $\beta$ ) was centered at the mean of each baseline and treatment phase. The y-intercept was defined this way so there was a standard initial level of behavior for all of the graphs. Centering the y-intercept gave HLM the highest probability of achieving comparable results across graphs.

Several error terms ( $e$ ) were tested to determine which fit each graph's individual data best. The first error term, the First-Order Autoregressive model, assumed error was independent,

and autocorrelation or covariation might have been present in the data. This model specified that any data point depended on the preceding data point and incorporated this assumption into the calculation of the deviation between predicted values and model outcomes. The second error term, the Homogeneous error model, assumed equal variances at each time point and that all covariances were equal. This model did not factor in autocorrelation, and was a simpler model based on fewer underlying parameters. In both models, the overall HLM model was translated into the framework of Structural Equations Modeling and covariance was incorporated into the error term (Raudenbush & Bryk, 2002). A third potential error term, the Unrestricted model, was not used because the large number of data points required for this model rendered it unable to process the small number of data points associated with each graph.

HLM6S found the best error model for each graph's data by determining model fit, or the estimated deviance and degrees of freedom associated with each model. The best model was the one with the least deviance (indicated by the lowest deviance value) and the most degrees of freedom. HLM6S compared the difference in the deviance of each model to percentiles from a chi square distribution table as a function of the difference between the degrees of freedom. A significant  $p$  value given by the program indicated which model was a better fit to the data, and a  $p$  value not meeting the required significance level indicated the simpler error model (the Homogeneous model) was an appropriate predictor of model outcomes. Therefore, the Autoregressive model was used only when a significant  $p$  value indicated it was a better fit. If the  $p$  value showed the models were not significantly different, or if the Homogeneous model was a significantly better fit, that model was used instead.

To determine whether each graph showed a significant change in behavior across conditions, HLM6S performed a significance test for changes in level and trend. For either to be

considered statistically significant, HLM had to show specific results. For the first factor, level, to be significant, conditions had to have statistically different levels,  $p < 0.05$ . If the test revealed they were not significantly different, then there was no treatment effect on level in the graph,  $p > 0.5$ . For the second factor, trend, conditions had to have significantly different trends,  $p < 0.05$ . If HLM revealed they were not different, the results meant there was no trend across the conditions or statistically there was the same level of trend across conditions,  $p > 0.05$ . Therefore, for each graph, HLM showed whether there was a difference across conditions for level, trend, and as a result, the third factor of both level and trend together (the graph as a whole). For the graph to qualify as significant as a whole, both level and trend had to be statistically significant,  $p < 0.05$ . These criteria for overall significance were analogous to the rater's task of judging level, trend, and the graph as a whole, ensuring the test of statistical significance was comparable to the visual analysis decisions.

When graphs had multiple dependent variables, the variables were analyzed individually, and each variable had to meet the significance criteria outlined above for the graphs to be considered significant. In addition, if the graph included multiple treatments, the significance test indicated if there were a significant difference between baseline and one or more of the treatments, but did not indicate where this difference occurred. These general significance tests of treatments and dependent variables were analogous to the general conclusions given by the visual analysis raters.

### **Receiver Operating Characteristic Curves and Contingency Probability Tables**

As previously stated, ROC curves are based on dichotomies of “significant” versus “nonsignificant.” There were several potential dichotomies to test in this study based on the

factors level, trend, and the graph as a whole. Each visual analysis dichotomy for each factor was used as the “Gold Standard” against which the classification accuracy of HLM was tested.

The dichotomies were determined by using the average ratings given to each graph by the 96 visual analysis raters. Because of these average ratings, the Likert scale values used by the participants had to be extended into an average rating range surrounding each Likert scale value. For example, an average rating of 2.5 to 3.5 was analogous to the Likert scale value 3 (Table 2). In the Likert scale, a rating of 1 meant the raters were “Not at all Certain” of intervention effects, while a rating of 5 meant the raters were “Extremely Certain.” These values represented the absolute lowest and highest average rating possible. Therefore, the range associated with these values was confined within the lower and upper limits of the scale, and an average rating of 1 to 1.5 was analogous to the scale value 1, whereas an average rating of 4.5 to 5 was analogous to the scale value 5.

Five potential dichotomies were determined using the rating ranges (Table 3). Each dichotomy represented the lowest visual analysis rating accepted as significant. The significance criterion became more stringent as the average rating considered significant increased. Dichotomy 1 assumed any graph given an average rating of 1 or higher was significant by visual analysis standards, and all graphs met this criterion. Dichotomy 2 would only consider graphs given an average rating of 2 or higher significant (because of the range associated with each value, this dichotomy actually encompassed graphs rated 1.5 or higher), and so on. Dichotomy 5 required graphs to have an average rating of at least 5, and no graphs met this criterion.

Although five potential dichotomies were identified, only three could be analyzed using ROC curves. Dichotomies 1 and 5 could be not used because these dichotomies assumed that either all of none of the graphs were significant. ROC curves require at least one example of a

Table 2

Likert Scale Values and Corresponding Average Rating Ranges

Verbal Label	<u>Rating Scales</u>		<u>Number of Graphs</u>		
	Likert Value	Average Rating Range	Level	Trend	Whole
Not at all Certain	1	1 to 1.5	2	2	2
Somewhat Certain	2	1.5 to 2.5	6	10	8
Moderately Certain	3	2.5 to 3.5	12	9	10
Reasonably Certain	4	3.5 to 4.5	8	7	8
Extremely Certain	5	4.5 to 5	0	0	0

*Note:* “Number of graphs” indicates the graphs given an average rating within a particular range.

significant and nonsignificant test result in a potential dichotomy, and these dichotomies did not satisfy that condition. Therefore, HLM could be tested against three potential visual analysis dichotomies for each of the three factors.

HLM only provided one potential dichotomy (significant or nonsignificant) for level and trend, meaning HLM was tested only once against each visual analysis dichotomy for each factor. To test HLM’s classification ability when considering only change in level, HLM was tested against all three visual analysis dichotomies generated by the rater’s decisions based on level. The same test was used to judge HLM’s classification accuracy when considering only change in trend. Changes in the graph as a whole were tested against raters’ decisions made considering the entire graph. For this test of the graph as a whole, there were again three potential dichotomies of visual analysis against which to test HLM. There were also four potential results for HLM significance, formed by combining the separate HLM

Table 3

ROC Curve Dichotomies for Visual Analysis and HLM

	<u>Level</u>	<u>Trend</u>	<u>Whole</u>
Potential	<i>I</i> : 1 (all)	<i>I</i> : 1 (all)	<i>I</i> : 1 (all)
Visual Analysis	2: 1 to 1.5 vs. 1.5 to 5	2: 1 to 1.5 vs. 1.5 to 5	2: 1 to 1.5 vs. 1.5 to 5
dichotomies	3: 1 to 2.5 vs. 2.5 to 5	3: 1 to 2.5 vs. 2.5 to 5	3: 1 to 2.5 vs. 2.5 to 5
(based on	4: 1 to 3.5 vs. 3.5 to 5	4: 1 to 3.5 vs. 3.5 to 5	4: 1 to 3.5 vs. 3.5 to 5
rankings)	5: 4.5 to 5 (none)	5: 4.5 to 5 (none)	5: 4.5 to 5 (none)
Total dichotomies	3	3	3
Potential	Significant	Significant	Significant: Yes and yes
HLM	or	or	or
dichotomies	Not significant	Not significant	Not significant: No and no
(based on			Yes and no
significance test)			No and yes
Total dichotomies	1	1	1

judgments of the graph as a whole (both level and trend significant, neither level nor trend significant, level significant and not trend, or trend significant and not level), but only one was considered significant – the conclusion that both level and trend were significant. The other three possibilities were all considered nonsignificant. Therefore, even when considering the graph as a whole, there was still only one dichotomy for HLM.

Once each HLM dichotomy was tested against its relevant visual analysis dichotomies, more in-depth analyses were generated for specific dichotomies using Contingency Probability tables if the ROC curve analysis revealed that HLM was significantly more accurate than chance (i.e., if the AUC was greater than and statistically different from an AUC of 0.5). These analyses included the Overall Accuracy of HLM for that particular dichotomy, as well as the Positive and Negative Likelihood Ratios, calculated as follows:

Overall Accuracy =

$$(TP + TN) / (TP + FP + TN + FN)$$

Positive Diagnostic Likelihood Ratio =

$$[TP / (TP + FN)] / [FP / (FP + TN)] \text{ or Sensitivity} / (1 - \text{Specificity})$$

Negative Diagnostic Likelihood Ratio =

$$[FN / (TP + FN)] / [TN / (FP + TN)] \text{ or False Negative Rate} / \text{True Negative Rate}$$

## RESULTS

### Visual Analysis Survey

**Demographics.** Less than two percent of the total sample responded to the email and completed the survey, but of those who began the survey, 30.4% completed it, giving 98 survey responses. The responses from the first two participants were dropped after their comments about the survey led to the survey instructions being clarified. Therefore, there were 96 participants.

Average survey completion time was 36 minutes. The majority of participants were female (76%) and most were between the ages of 25 and 34 (49%). All of the respondents at least had a Bachelor's degree, with most holding a Master's degree (58%). Almost 35% had a Doctorate.

Most of the participants were not journal board members (89%), but of those who did serve journals, the majority served the *Journal of Applied Behavior Analysis* (13%). Over 30 different journal boards were represented. All participants held a current BACB certification or had been previously certified (94% and 6%, respectively).

Several questions asked participants about their experience with single-subject design, and almost 63% of participants had used single-subject designs in some capacity for over five years. Ninety-eight percent had been using single-subject designs over one year. The majority used single-subject designs in school (38%), clinical (31%), or research (28%) settings.

Participants were asked to provide any comments or suggestions they had about the survey, and although most participants did not provide any, several comments were consistent across the 29 participants who did respond. Most of the comments were about the length of the survey (21%) or the difficulty of rating graphs in an artificial manner (21%) or the instructions were unclear (18%). Some participants commented that the graphs appeared fake, (3%) or that

they did not find the definitions provided for each graph helpful (3%). In addition, several mentioned the need for graph presentation to be randomized (9%), and were unaware they were presented randomly.

**Agreement.** Because of the large number of participants, Pearson  $r$  correlations were used to judge the consistency of participant ratings, in lieu of agreement calculations like proportion agreement or Kappa. When judging level,  $r = 0.46$ ,  $p < 0.05$ , when judging trend,  $r = 0.46$ ,  $p < 0.05$ , and when judging the graph as a whole,  $r = 0.43$ ,  $p < 0.05$ .

**Ratings.** Participants judged the graphs based on level, trend, and the graph as a whole. Most of the graphs were given a rating of 3 across the three factors. Because average ratings were used, the rating of 3 encompasses the average rating range of 2.5 to 3.5. Only two graphs were rated lower than 1.5, and no graphs were rated higher than 4.5 (Table 2).

The range of ratings given to each graph was determined for each factor (Table 4). For level, less than 3% of the 39 graphs were given a 1 or a 2 rating. Five percent of the graphs were rated across the range 1 to 4. Ninety-two percent of the graphs were rated across a range of 1 to 5, meaning that for these graphs, at least one respondent rated the graph a 1, at least one respondent rated it a 2, et cetera, across the entire range possible (1 to 5). The same pattern held for ratings based on trend (85% of graphs were rated across the entire range possible, 1-5), and for judging the graph as a whole (90% of graphs were rated across the entire 1-5 range).

### **Hierarchical Linear Modeling**

During data analysis, several limitations for HLM were found. Simple single-baseline AB designs could be not analyzed, as HLM required more than one baseline and treatment phase (multiple-baseline AB designs contained multiple baseline and treatment phases and were analyzed sequentially, so HLM could accommodate that design type). In addition, the graphs

Table 4

Percentage of Graphs within each Rating Range

Rating range	<u>Percentage of graphs rated within range</u>		
	Level	Trend	Whole
(2) 1 to 2	2.56%	2.56%	5.13%
(3) 1 to 3 or 2 to 5	--	2.56%	2.56%
(4) 1 to 4	5.13%	10.26%	2.56%
(5) 1 to 5	92.31%	84.62%	89.74%

*Note:* Cells without data indicate no graphs were rated across the corresponding range needed to have a large number of overall data points or phases with at least three data points if there were a small number of overall data points. Phases with less than three data points were analyzed successfully if the graphs had a large number of overall points. Another limitation was HLM required at least some variability in the data. HLM models variability, so data showing no variability violated the model and could not be tested. However, HLM did not require that all phases showed variability, just that variability was present in at least one of the compared phases (e.g., baselines could show no variability if treatment did show variability).

Because of these limitations, eleven graphs had to be dropped from analysis, including five single-baseline AB graphs, five functional analyses, and one “other” (a multiple-baseline A<sub>1</sub>B<sub>1</sub>/A<sub>2</sub>). These graphs were all dropped because they only had one baseline and one treatment phase (HLM considers each experimental phase of a Functional Analysis as one instance of one treatment phase). Therefore, 28 graphs were successfully analyzed using HLM. Nineteen of the

graphs fit the Homogeneous error model, while the remaining nine were analyzed using the First Order Autoregressive model.

The required  $p$  value for graph significance was set at  $p < 0.05$ . When considering only level, 2 graphs were nonsignificant, and 26 were significant. For trend, 17 were nonsignificant and 11 were significant. Because both level and trend had to be significant for the graph to be considered significant as a whole, a more stringent criterion, analyzing the graph as a whole generated the most nonsignificant graphs (18) and the fewest significant graphs (10).

### **Receiver Operating Characteristic Curves and Contingency Probability Tables**

The ROC curve analyses were generated by the MedCalc software in “demo” mode (Schoonjans, 2007). Limitations of the “demo” mode did not affect ROC curve calculations. A power analysis found that forty-one graphs were required to detect an AUC of 75%; however, only twenty-eight graphs were available for each curve.

ROC curves were generated for all twenty-eight graphs across the three factors and the three dichotomies, for nine ROC curves total. For level, the dichotomy split at an average rating of 2 (therefore extending the ratings included down to an average rating of 1.5) showed HLM classified graphs correctly 73% of the time (AUC = 0.731, Table 5, Figure 3). The significance of the AUC was  $p > 0.05$ , indicating HLM was not significantly more accurate than chance. For an average visual analysis rating of 3, the AUC = .625,  $p > 0.05$  and for an average rating of 4, AUC = .55,  $p > 0.05$  (Table 5, Figures 4 and 5).

When judging the accuracy of HLM against visual analysis judgments of trend, graphs considered significant at an average rating of 2 and higher generated AUC = .712. For an average rating of 3, AUC = .552, and for an average rating of 4, AUC = .714. All of these AUCs had a significance level of  $p > 0.05$  (Table 5, Figure 6, 7, and 8).

When judging the accuracy of HLM against ratings of the graph as whole, graphs considered significant at an average rating of 2 and higher generated  $AUC = .692, p > 0.05$ . For graphs considered significant at an average rating of 3 and higher,  $AUC = .622, p > 0.05$ . For graphs considered significant at an average rating of 4 or higher,  $AUC = .775, p < 0.05$ , indicating that HLM was significantly better than chance at accurately classifying these graphs (Table 5, Figures 9, 10 and 11).

Table 5  
AUCs and Significance Values for ROC Curves

ROC curve dichotomy	Level	<u>AUC</u>	
		Trend	Whole
1.5 to 5 <i>(average rating of 2)</i>	.731	.712	.692
2.5 to 5 <i>(average rating of 3)</i>	.625	.552	.622
3.5 to 5 <i>(average rating of 4)</i>	.55	.714	.775*

\* $p < 0.05$

The only significant AUC was generated when judging graphs as a whole and when using the most stringent visual analysis dichotomy (an average rating of 4),  $AUC = .775, p < 0.05$ . Therefore, a Contingency Probability table was generated for this graph only (Table 6). Sensitivity (true positives) was 75% at the equal cost point of the curve (6 of 8 graphs with a visual analysis rating of 4 or higher were correctly classified as significant by HLM) and Specificity (true negatives) was 80% (16 of 20 graphs with a rating lower than 4 were classified correctly as nonsignificant by HLM). Overall Accuracy, calculated using the prevalence of true

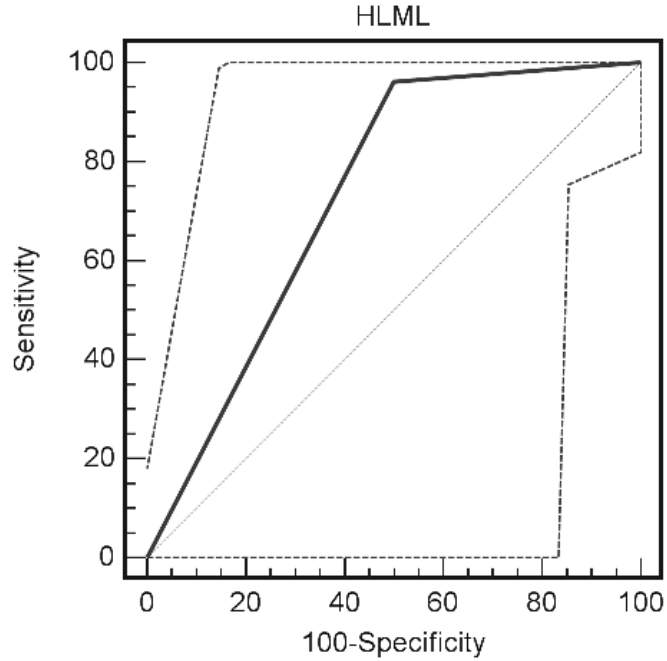


Figure 3. ROC curve of level: visual analysis ratings of 2 and higher vs. HLM. The thick black line is the ROC curve. The dotted black lines represent the 95% Confidence Interval surrounding the curve. The grey line represents an AUC of 50%, or chance.

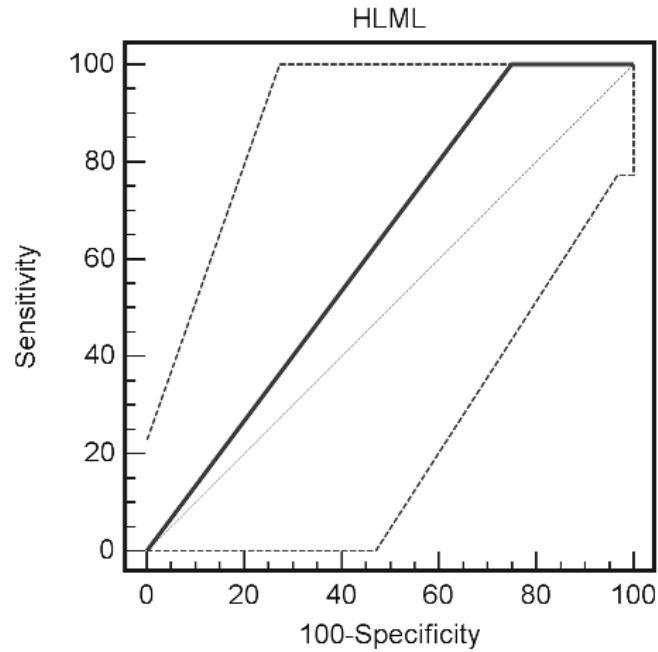


Figure 4. ROC curve of level: visual analysis ratings of 3 and higher vs. HLM

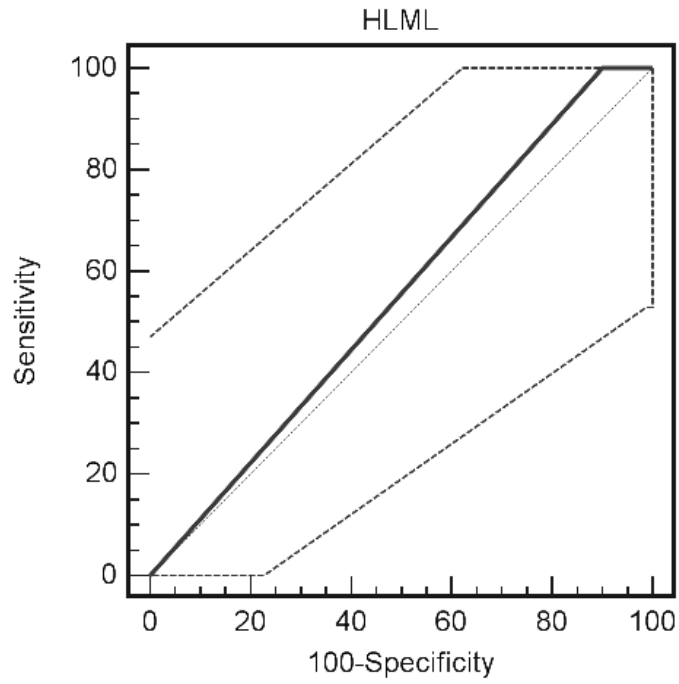


Figure 5. ROC curve of level: visual analysis ratings of 4 and higher vs. HLM

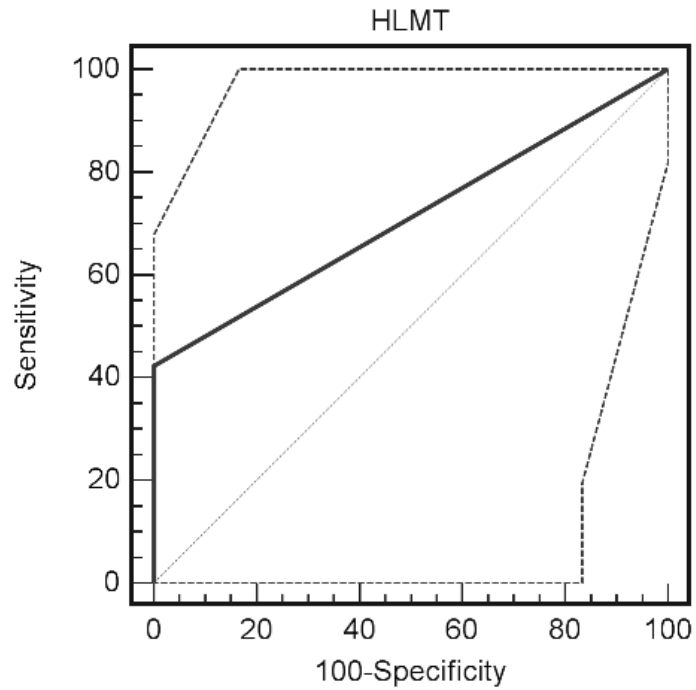


Figure 6. ROC curve of trend: visual analysis ratings of 2 and higher vs. HLM

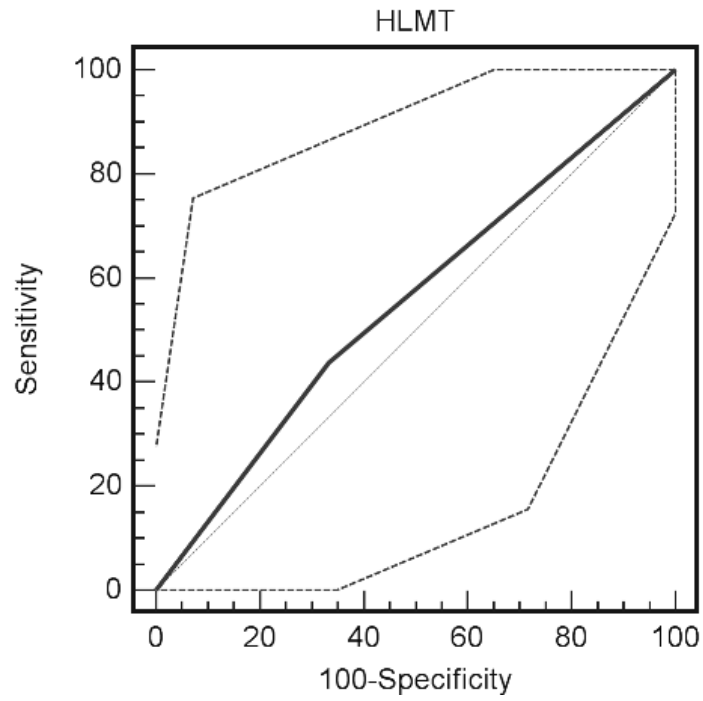


Figure 7. ROC curve of trend: visual analysis ratings of 3 and higher vs. HLM

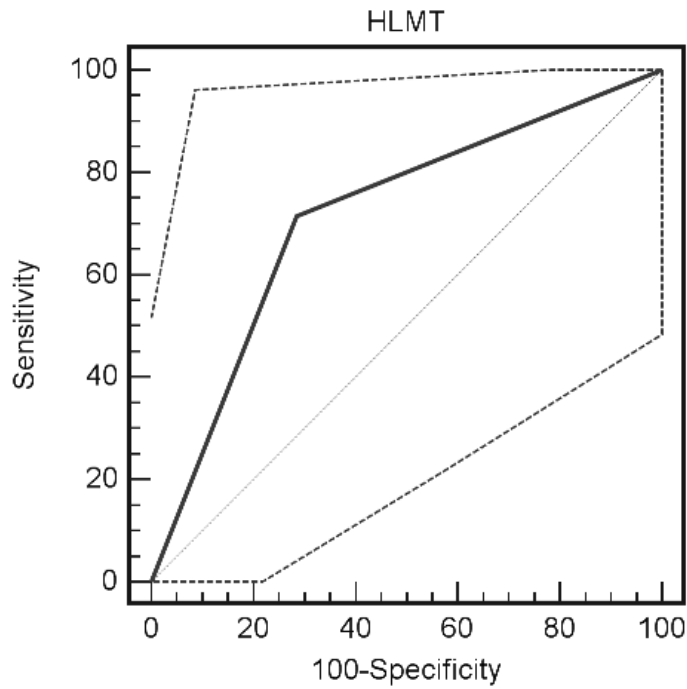


Figure 8. ROC curve of trend: visual analysis ratings of 4 and higher vs. HLM

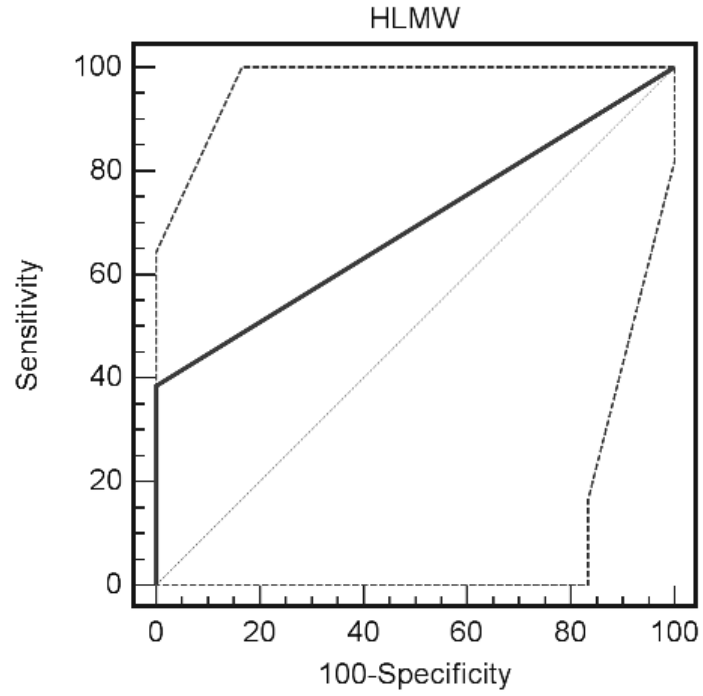


Figure 9. ROC curve of the graph as a whole: visual analysis ratings of 2 and higher vs. HLM

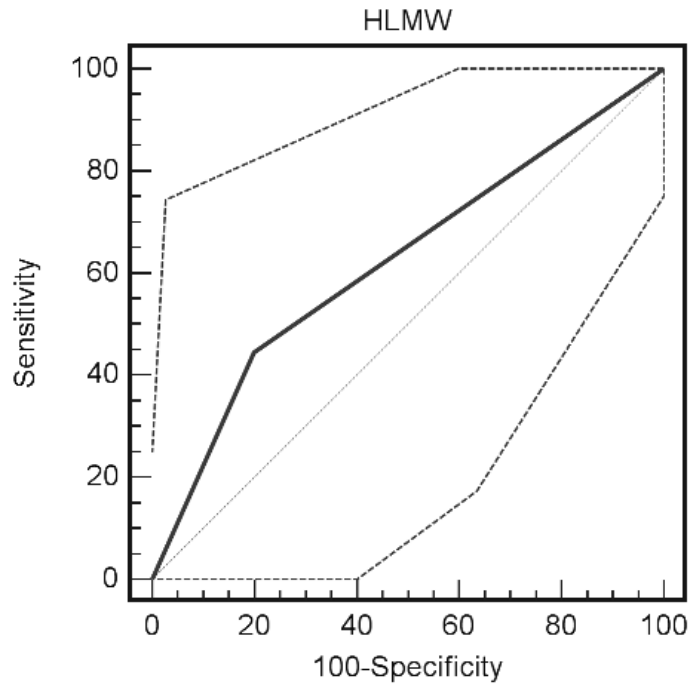


Figure 10. ROC curve of the graph as a whole: visual analysis ratings of 3 and higher vs. HLM

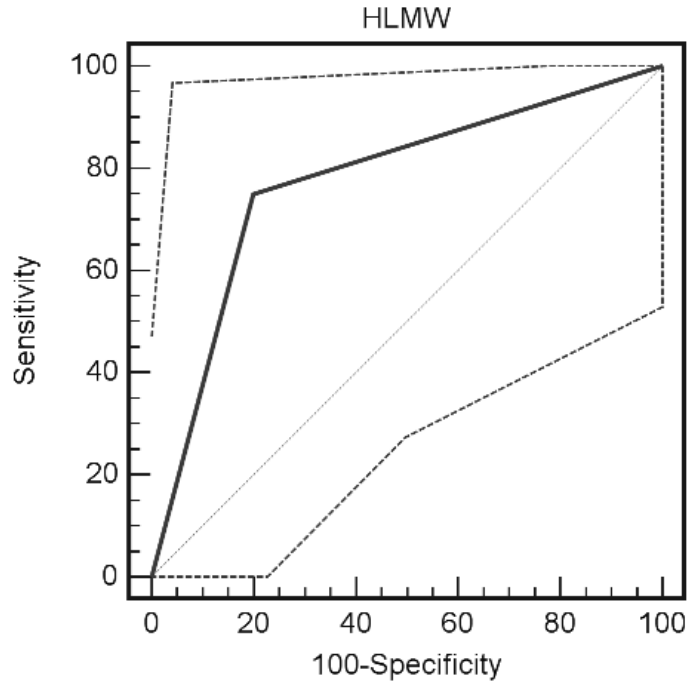


Figure 11. ROC curve of the graph as a whole: visual analysis ratings of 4 and higher vs. HLM

Table 6

Contingency Probability Table

<u>Gold standard</u>	<u>New test</u>	
	<i>p</i> < 0.05	<i>p</i> > 0.05
<i>(Visual analysis)</i>	<i>(significant)</i>	<i>(nonsignificant)</i>
Rating ≥ 4	6	2
<i>(significant)</i>	<i>(True positives/sensitivity)</i>	<i>(False negatives)</i>
Rating < 4	4	16
<i>(nonsignificant)</i>	<i>(False positives/1-specificity)</i>	<i>(True negatives/specificity)</i>

positives and true negatives, was 78.57. The Positive Diagnostic Likelihood Ratio, equivalent to Type I error was 0.4, and the Negative Diagnostic Likelihood Ratio or Type II error, was 0.3.

## DISCUSSION

### Visual Analysis Survey

Survey participants appeared to be a group highly qualified to judge single-subject design graphs. The majority of participants had either a Master's or a Doctoral degree, and had been using single-subject design graphs for five years or more in research and applied settings. In addition, they were all formerly or currently certified as Board Certified Behavior Analysts or Board Certified Associate Behavior Analysts.

The average ratings for each graph indicate a conservativeness that is often assumed with visual analysis but has not necessarily been found in previous research. No graph received the highest average rating possible (5) and no graph received an average rating higher than 4.5, meaning no graphs were rated at the "Extremely Certain" level – even though 57% of the twenty-eight graphs included ROC curve analysis were coded as showing "unambiguous," "clear," or "certain" intervention effects by the journal article authors. In fact, the highest average rating given to a graph was 4.48. In spite of these lower than expected average ratings, the expected wide range of ratings given to each graph was found, with 89% of graphs rated over the entire range of possible ratings (1-5) across the three factors.

Pearson  $r$  correlations of graph ratings also followed findings of previous research. The moderate  $r$  obtained with the correlations indicates only a modest relationship among raters' graph ratings. Although the correlation coefficients do not indicate agreement per se, they do indicate that the raters did not give each graph consistent ratings. Again, this lack of consistency is comparative to previous research into visual analysis agreements.

## **Hierarchical Linear Modeling**

The study found several limitations of HLM when using single-subject design graphs. The most stringent limitation was the inability to analyze graphs with only single baseline and treatment phases, meaning that basic AB designs and Functional Analyses could not be included in the analysis. Fortunately, these design types do not appear to constitute the majority of single-subject designs used currently (combined they represent 22% of the graph database), and HLM was appropriate for all other design types (78% of graphs published in the selected journals from 2002-2006). In addition, these non-usable designs could potentially be modified by adding phases if authors wished to analyze their data statistically.

The necessity of analyzing multiple baseline graphs sequentially is also a limitation. This method does not completely follow the assumptions behind this design type (i.e., phases run more concurrently than sequentially), but HLM did give a suitable analysis of the graphs as determined by comparing the HLM output and the original dataset. Future research may identify a way to analyze the data more comparable to multiple baseline principles.

In spite of these limitations, HLM was found to be a generally robust and usable statistic with single-subject data. Most design types (nine of the eleven coded) could be analyzed appropriately with HLM, with HLM able to satisfy assumptions of the data like accommodating missing data or the presence of autocorrelation. HLM was also able to accommodate varying numbers of phases and varying amounts of data within phases. In addition, the twenty-eight graphs successfully analyzed covered a wide range of participants, behaviors, treatment types, and single-subject areas like clinical and school psychology and applied behavior analysis. HLM may be the most appropriate and usable statistic found thus far in the statistical analysis of single-subject data.

## Receiver Operating Characteristic Curves

The limitations of HLM inherently limited the ROC curve analysis used to judge the accuracy of HLM against the “Gold Standard” of visual analysis. Because graphs had to be excluded from the HLM analysis, fewer graphs were available to generate the ROC curves than required, resulting in a substantial decrease in power. Eight of the nine ROC curves generated were unable to differentiate the accuracy of HLM from chance, and these findings were likely due to a lack of power. This problem stemmed from the decision to use the minimum number of graphs necessary to generate ROC curves in the visual analysis survey and HLM analysis. The maximum number of graphs participants could rate in one sitting was about 40 graphs, based on previous research, and this number coincided with the number required for the ROC curve analysis. Additional graphs were not included in the survey because they appeared unnecessary.

For the eight ROC curves showing HLM was not significantly better than chance, HLM ranged from 55% accurate to 71% accurate, with mist at 60% or better. These AUCs show that HLM *could* be a useful tool for visual analysis judgments, if future research can satisfy the necessary power requirements and demonstrate HLM is better than chance at classifying single-subject data as significant or nonsignificant.

The ROC curve showing HLM was significantly better than chance (and, in fact, almost 78% accurate when compared to visual analysis) shows this result is robust because it occurred even with very low power. This ROC curve indicates that when judging the graph as a whole, and when raters are “Reasonably Certain” of intervention effects (an average visual analysis rating of 4), HLM is an effective analysis method that corresponds to visual analysis judgments on almost four of five graphs. This agreement rate between HLM and visual analysis is higher than most agreement rates among raters in previous research, and is the same level of agreement

visual analysis raters had with themselves in the Knapp (1983) study, when raters judged the same graph multiple times. At this level of visual analysis certainty, raters may or may not require a statistical tool like HLM (as they were “Reasonably Certain” of intervention effects), but for those who would like a “second opinion,” HLM could be a valuable asset to visual analysis judgments and possibly more consistent than asking another visual analyst.

Future research should focus on providing enough power to the ROC curve analysis while maintaining a qualified sample of raters and a variety of graph types and subjects. HLM may appear more or less helpful when tested against a wider range of graphs and single-subject datasets. HLM has been supported in this study as a potentially useful and practical tool for visual analysis judgments, but its full possibilities have yet to be demonstrated because of the lack of power of the present study.

## REFERENCES

- Alberg, A. J., Park, J. W., Hager, B. W., Brock, M. V., & Diener-West, M. (2004). The use of "Overall Accuracy" to evaluate the validity of screening or diagnostic tests. *Journal of General Internal Medicine, 19*, 460-465.
- Allison, D. B., Franklin, R. D., & Heshka, S. (1992). Reflections on visual inspection, response guided experimentation, and Type I error rates in single-case design. *Journal of Experimental Education, 61*, 45-51.
- Baer, D. M. (1977). "Perhaps it would be better not to know everything". *Journal of Applied Behavior Analysis, 10*, 167-172.
- Brossart, D. F., Parker, R. I., Olson, E. A., & Mahadevan, L. (2006). The relationship between visual analysis and five statistical analyses in a simple AB single-case research design. *Behavior Modification, 30*, 531-563.
- Busk, P. L., & Marascuilo, L. A. (1988). Autocorrelation in single-subject research: A counter-argument to the myth of no autocorrelation. *Behavioral Assessment, 10*, 229-242.
- DeProspero, A., & Cohen, S. (1979). Inconsistent visual analysis of intrasubject data. *Journal of Applied Behavior Analysis, 12*, 573-579.
- Edgington, E. S. (1992). Non-parametric tests for single-case experiments. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Fisch, G. S. (1998). Visual inspection of the data revisited: Do the eyes still have it? *The Behavior Analyst, 21*, 111-123.
- Gibson, G., & Ottenbacher, K. (1988). Characteristics influencing the visual analysis of single-subject data: An empirical analysis. *The Journal of Applied Behavioral Science, 24*, 298-314.
- Gresham, F. M., McIntyre, L.L., Olson-Tinker, H., Dolstra, L., McLaughlin, V., Van, M. (2004). Relevance of functional behavioral assessment research for school-based interventions and positive behavioral support. *Research in Developmental Disabilities, 25*, 19-37.
- Hintz, J. (2007). PASS: Power analysis and sample size [computer program]. Kaysville, UT: NCSS Statistical Software.
- Hopley, L., & van Schalkwyk, J. (2006). The magnificent ROC (receiver operating characteristic curve). Retrieved December 1, 2006, from <http://www.anaesthetist.com/mnm/stats/roc/Findex.htm>

- Jones, R. R., Weinrott, M. R., & Vaught, R. S. (1978). Effects of serial dependency on the agreement between visual and statistical interference. *Journal of Applied Behavior Analysis, 11*, 277-283.
- Kazdin, A. E. (1982). *Single-case research designs: Methods for clinical and applied settings*. New York: Oxford University Press.
- Knapp, T. J. (1983). Behavior analysts' visual appraisal of behavior change in graphic display. *Behavioral Assessment, 5*, 155-164.
- Langdon, W. B. (2006). Receiver operating characteristics (ROC). Retrieved November 27, 2006, from <http://www.cs.ucl.ac.uk/staff/W.Langdon.roc/>
- Ma, H. H. (2006). An alternative method for quantitative synthesis of single-subject researches: Percentage of data points exceeding the median. *Behavior Modification, 30*, 598-617.
- Matyas, T. A., & Greenwood, K. M. (1990). Visual analysis of single-case time series: Effects of variability, serial dependence, and magnitude of intervention effects. *Journal of Applied Behavior Analysis, 23*, 341-351.
- Michael, J. (1974). Statistical inference for individual organism research: Mixed blessing or curse? *Journal of Applied Behavior Analysis, 7*, 647-653.
- Ottenbacher, K. J. (1990a). Visual inspection of single-subject data: An empirical analysis. *Mental Retardation, 28*, 283-290.
- Ottenbacher, K. J. (1990b). When is a picture worth a thousand *p* values? A comparison of visual and quantitative methods to analyze single subject data. *The Journal of Special Education, 23*, 436-449.
- Park, H. S., Marascuilo, L. A., & Gaylord-Ross, R. (1990). Visual inspection and statistical analysis in single-case designs. *Journal of Experimental Education, 58*, 311-320.
- QuestionPro.com. (2007). QuestionPro Survey Software. Retrieved November 26, 2007, from <http://www.questionpro.com>
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications, Inc.
- Raudenbush, S. W., & Bryk, A. S. (2007). HLM 6.03 [computer program]. Lincolnwood, IL: Scientific Software International.
- Schoonjans, F. (2006). ROC curve analysis in MedCalc. Retrieved November 27, 2006, from <http://medcalc.be/manual/mpage06-13b.php>

Schoonjans, F. (2007). MedCalc [computer software] (Version 9.3.9.0) [Demo]. Mariakerke, Belgium: ISV/Software Solutions.

SPSS Inc. (2007). SPSS 15.0 [computer program]. Chicago: SPSS Inc.

Stocks, J. T., & Williams, M. (1995). Evaluation of single subject data using statistical hypothesis tests versus visual inspection of charts with and without celeration lines. *Journal of Social Service Research, 20*, 105-126.

Technologies for Health Project. (n.d.). Accuracy of diagnostic tests. Retrieved December 6, 2006, from <http://www.rapid-diagnostics.org/accuracy.htm>

## VITA

Elizabeth Godbold graduated *magna cum laude* with a Bachelor of Arts degree in psychology from Wake Forest University in 2005. She decided to pursue a career in school psychology after working with children with academic and behavioral difficulties. Elizabeth began her studies at Louisiana State University in August 2005 under Dr. Frank M. Gresham.