2012

Hierarchical Linear Modeling versus visual analysis of single subject design data

Elizabeth Godbold Nelson
Louisiana State University and Agricultural and Mechanical College

HIERARCHICAL LINEAR MODELING VERSUS VISUAL ANALYSIS OF SINGLE SUBJECT DESIGN DATA

A Dissertation
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
Doctor of Philosophy

in

The Department of Psychology

by
Elizabeth Godbold Nelson
B.A., Wake Forest University, 2005
M.A., Louisiana State University, 2008
May 2012

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Visual analysis is the "gold standard" for single-subject design data because of a presumed low Type I error rate and consistency across raters. However, research has found it less accurate and reliable than typically assumed. Many statistics have been proposed as aids for visual analysis, but most suffer from limitations either due to methods of investigation or problems inherent to the statistics themselves. Several researchers have proposed the use of Hierarchical Linear Modeling to analyze single-subject data because it can withstand violations of assumptions often present in single-subject data that other statistics cannot. In addition, HLM is similar to the actual data structure of single-subject designs as it allows predictors to be nested within different levels of analysis. Godbold (2008) tested the accuracy of HLM against visual analysis ratings of the same data and found HLM to be a potentially useful statistical aid. The current study rectified the limitations of the 2008 study and extended the applicability of HLM to more types of single-subject designs. HLM was again shown to be a viable statistic across a wide variety of design types including single and multiple baseline designs. Comparisons between two HLM models indicated a longitudinal HLM model was more accurate as compared to visual analysis than a simpler non-longitudinal 2-level model, however, more research is warranted.

INTRODUCTION

Single-subject designs are used in basic and applied settings across psychology, education, medicine, and business. They may be known as single-case designs, small-*n* designs, or *n* of one trials, among other names (Miller, 2003). These research designs typically begin with a discrete, observable dependent variable (DV) repeatedly measured in the absence of an independent variable (IV) for a single participant. Then, the IV is systematically introduced while measurement of the DV continues. Differences in the DV after the introduction of the IV are evaluated to determine the effect, if any, of the IV. Single-subject designs are often chosen when researchers want to demonstrate experimental control over the IV or to show individual differences in response to the IV. In contrast, large *n* or group design studies typically focus on the average response of a group of individuals to an IV, which can mask individual differences and imperfect experimental control. Single-subject research considered a viable alternative to group designs and the use of single-subject research is supported by the United States Department of Education (Sparks, 2012).

To interpret the data provided by single-subject designs, measurement of the DV is graphed with the unit of time on the x-axis (sessions, days, weeks, etc.) and DV values on the y-axis; manipulations of the IV are distinguished by phase lines (Figure 1). The graphs are judged based on visual data patterns composed of level, trend, and variability. Level in single-subject design refers to the average value of the DV across a particular phase and trend refers to a linear increase or decrease in the data pattern over time. Variability refers to how much the data within a phase deviates from the level or trend (Horner, et al., 2005).

A basic single-subject design is the ABAB design, where A represents the Baseline or "no treatment" phase (i.e., no manipulation of the IV) and B represents the Treatment phase, or

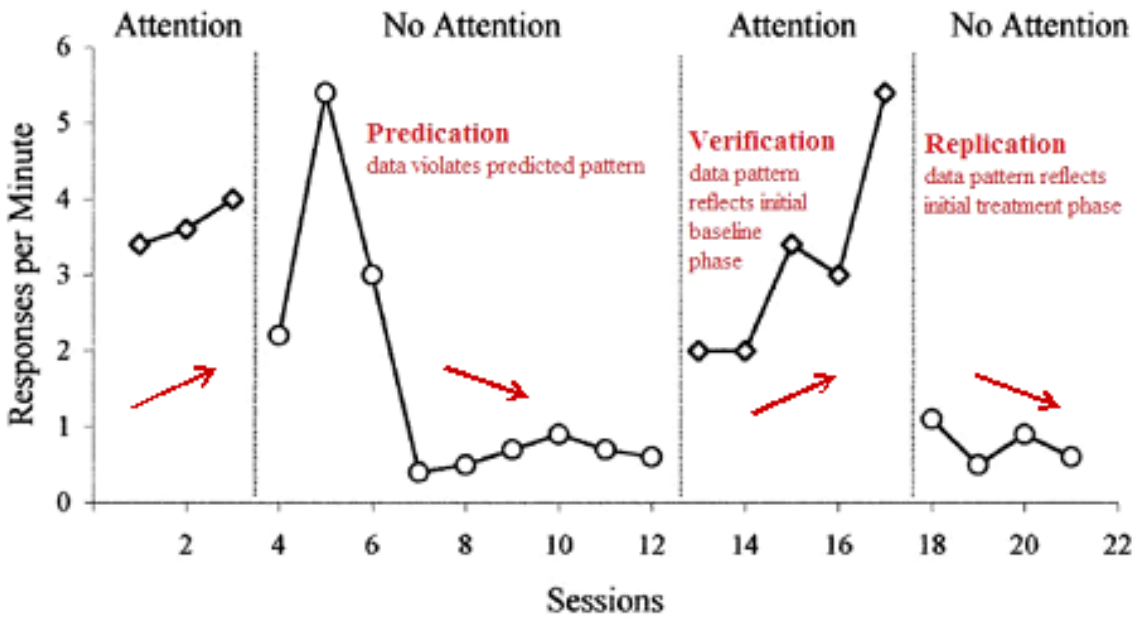

Figure 1. Rate of self-injurious behavior during the analysis.

Figure 1. Sample single-subject design with principles of visual analysis. This figure illustrates a single-subject design graph and the principles of predication, verification, and replication.

when the IV is manipulated (Figure 1). This design can be judged based on the foundations of prediction, verification, and replication, as opposed to statistical analyses as with large *n* designs. Prediction means one would expect the first data point after the introduction of a new IV (the first B phase) to remain at the same level and trend as in the previous phase (the first A phase) if the IV had *no* effect. Therefore, *violating* prediction indicates a potential effect of the IV, because a change in the data pattern is indicated. When the IV is withdrawn (i.e., the second A phase), verification further strengthens the presumed effect of the IV if the data returns to the same patterns as before the treatment was implemented, so both A phases resemble each other. Replication occurs when the IV is again implemented (the second B phase), and the data points return to the same pattern as the original treatment phase, so both B phases resemble each other.

The *sine qua non* or "gold standard" for interpreting data patterns is visual inspection. Researchers believe if a change occurred due to the IV, either in level, trend, variability or a

combination, it should be visually obvious and noticeable in graph form (Kazdin, 1982) without requiring further statistical investigation.

For Baer (1977), this obvious and noticeable change in behavior is the strength of visual analysis. Visual analysis is presumed to be more conservative and have a lower Type I error rate (false positives) than statistical analysis because visual analysis reveals only "powerful, general, and dependable" effects (p. 171), i.e. effects can be seen in a graphed data pattern. Conversely, statistical analysis of data could reveal subtle or weak changes in behavior that may be statistically but not clinically significant, leading to higher Type I error rates. The "true" Type I error rate cannot be calculated for single-case designs; however, Baer maintained the rate would be much smaller than .05, the conventional rate for statistical analysis, based on the presumed need for visually distinct differences in phase data patterns. The necessary increase in Type II error rates (false negatives) would be higher for single-subject research designs, but would not be a weakness when considering the desire for clinically significant results.

The use of visual analysis is based on this presumption of experimental control, conservativeness, and lower Type I error as well the assumption that visual analysis is consistent and reliable within and across those judging graphed data. In meeting these criteria, visual analysis would seem a true "gold standard" without any need for a judgmental aid such as statistical analysis. Indeed, Michael (1974) suggests statistical analysis would only serve to "abbreviate" the complexity of single-subject data and that the experimental control shown by rigorous single-subject design is the preeminent method to reduce experimental error and nuisance variables. Many experimenters have therefore seen little need for statistical aids based on the "gold standard" status of visual analysis (Baer, 1977; Johnston & Pennypacker, 1993; Michael, 1974; Parsonson & Baer, 1992; Sidman, 1960).

In contrast to these assumptions, an arresting line of research has shown problems with these assumptions about visual analysis: (a) visual analysts are consistent judges against others and even against themselves and (b) visual analysis is more conservative than statistical analysis, and consequently, has a lower rate of Type I error than statistical analysis.

Visual Analysis Research

Inter-Rater Agreement. Jones, Weinrott, and Vaught (1978) conducted the seminal study of visual analysis accuracy. Raters were presented with graphs showing "non-obvious" effects selected from the *Journal of Applied Behavior Analysis* (JABA) and asked to assess the meaningfulness of change in level across adjacent phases using the categories "yes," "no," and "unsure." The inter-rater agreement between judges ranged from .04 to .79, with a median of .39 – a conventionally low level of agreement.

DeProspero and Cohen (1979) extended the line of research using computer-generated graphs with varying levels of mean shift, variability, and slope. Raters were asked to judge experimental control using a scale of 1 to 100. The raters in this study also showed a low level of agreement at .61, slightly above chance. In addition, the graph deemed most "ideal" – showing large mean shift and little variability or slope – received ratings ranging from 3 to 100.

Matyas and Greenwood (1990) also used computer-generated graphs showing varying levels of autocorrelation and variability. Raters were asked to respond to the data with either a "conclusion of effect" or "no effect," and Type I and Type II errors were calculated as a function of the amount of autocorrelation and random variability in the graphs. Type I error rates increased with autocorrelation and variability and were as high as 84 percent.

Gibson and Ottenbacher (1988) and Ottenbacher (1990a) each used computer-generated graphs with varying degrees of mean shift, variability, slope, level, overlap (data points in one

phase within the range of data points in another phase), and autocorrelation (the ability to detect an individual data point from the one immediately preceding it, often present in single-case data). Though the studies differed on the responses available to participants (a 6-point Likert scale versus three categorical labels) the results also corroborated previous research. For Gibson and Ottenbacher (1988), the average interclass correlation among raters was .60. Ottenbacher (1990a) presented disagreement ratios, which ranged from .08 to .59. In addition to levels of agreement and disagreement among raters, the researchers also calculated the relationship between the manipulated graph features and rater agreement. In both studies, variability and slope showed large positive correlations with disagreement among raters. Mean shift and level changes showed negative correlations – the more obvious the changes in mean and level, the more raters agreed on an effect and the more certainty the raters showed. Other 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. In both studies, the authors concluded visual analysis showed unreliability when used with certain data patterns.

Graph Properties. Overall, the results of these studies have shown visual analysis can be inconsistent among raters and further weakened by the presence of autocorrelation, variability, or definite trend in the data (*median agreement* = .60). Other researchers have studied the effects of the physical features of the graphs themselves. Knapp (1983) used varying techniques and presentation styles to create graphs with varying mean shift but no trend or autocorrelation. Presentation styles were more likely to affect graphs with lower mean shift and raters were most likely to say a change occurred if there was an obvious physical feature (i.e., a physical line) between the phases. Other researchers seeking ways to improve visual analysis

accuracy found rater judgment did not necessarily improve with the addition of physical aids (i.e., celeration lines; Normand & Bailey, 2006; Stocks & Williams, 1995). In fact, the results of neither study detected an increase in rater accuracy, indicating the effects of presentation styles vary.

Fisch (1998) outlined two studies that also manipulated graphs physically. Fisch and Schneider used graphs with datasets placed away from the x-axis (toward the top), close to the x-axis, or in the middle. Raters gave correct responses more often when datasets were framed by the top or bottom of the graph, with those near the x-axis showing the highest proportion of correct responses. Fisch also described a study by Greenspan and Fisch where the researchers varied the number of data points in baseline and intervention phases (five in each, ten in each, or five in one and ten in the other). Raters showed the highest level of accuracy when the number of data points was unequal across phases. Raters showed the lowest level of accuracy when judging graphs with ten data points per phase.

Type I Error Rates. In addition to being affected to data and graph properties, visual analysis may also be less conservative and more prone to Type I error than previously assumed. As discussed previously, Matyas and Greenwood (1990) found the Type I error rate of visual analysis to range between 16 and 84 percent when variability and autocorrelation were present. However, when autocorrelation was not present, Type I error ranged from 0 to 13 percent. Allison, Franklin, and Heshka (1992) used this estimate as the basis for a study into the true amount of Type I error inherent in visual analysis. The researchers first decided 10 percent was a conservative estimate of error based on the results of Matyas and Greenwood. Then, assuming a researcher made a treatment decision based on available data at every other data point (five times over ten data points), the researcher would have a 10 percent total error rate each time a

decision was made. Consequently, the real rate of Type I error across the ten data points would increase to 25.9 percent, much higher than the original conservative estimate of the authors (10 percent) and the rate touted by Baer (less than 5 percent). Although this error rate is simply plausible and not a blanket assertion, the research practices of many employing single-subject designs may cause visual analysis to be less conservative than originally believed. This consideration is especially important as the flexibility of single-subject designs, that is, the ability to make experimental decisions throughout the research project, is considered an advantage over large group designs and is often employed and encouraged by clinicians and researchers (Miller, 2003).

Statistical Analysis Research

Statistical tests have been proposed as judgmental aids by many researchers, including those from the aforementioned studies, due to their known level of Type I error and because they are reliable and consistent (resulting in the same conclusion every time) regardless of who conducts the test. Statistical tests would seem especially useful when visual analysis is known to be problematic (e.g., in the presence of variability, autocorrelation, or lack of obvious mean shift). The use of statistical tests does not mean a replacement for visual analysis, instead, it can be viewed as an aid to visual analysis in several ways by (a) enhancing reliability and consistency, (b) giving researchers and clinicians a means to corroborate visual analysis decisions, especially when considering important treatment decisions, (c) providing an empirical "check" for researchers and clinicians, either by forcing them to examine data more closely when contrasting decisions arise or by tempering the tendency of some researchers to overestimate treatment effects, and (d) providing a common metric for discussing effects across participants, studies, and treatments.

Parametric and Non-Parametric Statistics. In their landmark study, Jones, et al. (1978) tested the utility of time series analysis in judging single-case design effects. The same graphs viewed by visual analysis raters were analyzed with time series analysis and classified as "significant" ($p < 0.05$; equivalent to a "yes" rating by the judges) and "non-significant" or "unsure" ($p > 0.05$; equivalent to a "no" or "unsure" rating) graphs. The average agreement between the judges' visual analysis decisions and the statistical decisions was $P_A = .50$ (best was $P_A = .65$), indicating each judge, on average, agreed with the statistical decisions at chance levels; agreement was lower for highly autocorrelated time series data.

Other non-parametric statistics have also been assessed as an aid or replacement to visual analysis: randomization tests (Edgington, 1992; Park, Marascuilo, & Gaylord-Ross, 1990), split-middle techniques (Ottenbacher, 1990b) *t* tests of mean differences, and piecewise regressions (Stocks & Williams, 1995). However, each statistic was found either to have low levels of agreement with visual analysis ratings or to be problematic when used with single-case design data, as they were either not conducive to single-subject research methodologies or single-subject data violated the assumptions of the statistics.

Effect Size Indices. A current trend in statistical aids is effect sizes. Many researchers have either described or investigated the utility of numerous effect sizes that aggregate into three categories. The first category is based on standardized mean differences, such as *Cohen's d* and McGraw and Wong's *Common Language Effect Size* (Parker & Hagan-Burke, 2007), as well as the *Binomial Effect Size Display* based on Cohen's *d* (Parker & Hagan-Burke, 2007). These effect sizes quantify the average effect of treatment relative to observed variability (Higgins & Green, 2011).

A second category, regression-based effect size indices (e.g., R^2), has been studied in

Olive and Smith (2005) and Parker and Hagan-Burke (2007), along with five models investigated in Manolov, Solanas, and Leiva (2010): Gorsuch's *trend effect sizes*, White, Rusch, Kazdin, and Hartmann's *d*, Center, Skiha, & Casey's *mean plus trend difference model*, and Allison and Gorman's *mean plus trend difference model*. A study by Brossart, Parker, Olson, and Mahadevan (2006) also investigated the models by Gorsuch, Allison and Gorman, and Center, Skiha, and Casey, as well as the White and Haring *binomial test on extended Phase A baseline* and White's *Last Treatment Day*. Very generally, regression-based effect sizes are linear estimations of slope and level fitted to single-case data in different phases of a design. Any differences in fit are then compared to determine a treatment effect (Swaminathan, Horner, Rogers, & Sugai, 2012).

Visual effect sizes are based on displays of data, such as *Percentage of Non-overlapping Data Points* (PND; Brossart et al., 2006; Manolov, Solanas, & Leiva, 2010; Parker & Hagan-Burke, 2007), *Percentage of All Non-overlapping Data Points* (Manolov et al., 2010; Parker & Hagan-Burke, 2007), *Percentage of Data Points Exceeding the Mean* (PEM; Brossart et al., 2006; Manolov et al., 2010; Parker & Hagan-Burke, 2007), and the *Improvement Rate Difference* (Parker & Hagan-Burke, 2007; Parker, Vannest, & Brown, 2009). These effect sizes are based on selecting a certain parameter and comparing the number of data points that fall within and outside of the parameter; parameters could include the number of data points overlapping between two phases, or the number of treatment phase data points overlapping the median value of data points in the baseline phase (Wendt, 2009).

Research into effect sizes has included both hypothetical and real comparisons to visual analysis. Despite real interest, effect sizes may not be fully appropriate for single-case design data for many of the same reasons as the parametric and nonparametric statistics previously

studied. Research designs were often problematic as well, as discussed in the next section.

Problems and Limitations of Previous Research

Visual Analysis Studies. Although there is large body of research into visual analysis and possible statistical aids, many of the studies (and statistics themselves) suffer concrete limitations. Most of the research of visual analysis accuracy (and statistical aids when utilizing a rater component) used computer-generated graphs fitting specific data types conceived by the authors (Brossart et al., 2006; DeProspero & Cohen, 1979; Gibson & Ottenbacher, 1988; Knapp, 1983; Matyas & Greenwood, 1990; Ottenbacher, 1990a; Parker et al., 2009). Graphs sometimes contained little or no context for the data (e.g., DeProspero & Cohen, 1979) and subsequently, raters sometimes refused to participate in the task. In addition, many researchers gave raters ambiguous response requirements about the degree of change seen in the graphs – some asked for social validity judgments ("meaningful change" or a "significant change in performance," Gibson & Ottenbacher, 1988; Jones et al., 1978; Ottenbacher, 1990a; Park et al., 1990), some "experimental control" judgments (DeProspero & Cohen, 1979), and some asked about certainty (Brossart et al., 2006; Stocks & Williams, 1995).

Additionally, some researchers used discrete response categories ("yes," "no," or "unsure") and others used Likert scales. DeProspero and Cohen (1979) used a 100-point scale with only "Low" and "High" end markers as guidelines. Kahng, et al., (2010) updated DeProspero and Cohen's study with experienced raters (previous and current JABA board members) and very precise instructions about the scale and response categories available, and found an agreement level (IRA) of .93 between judges, indicating the ambiguity in scales and response terms in the original study likely had a direct, negative impact on agreement levels.

Many studies of both statistical and visual analysis used AB graphs with only one

baseline and one treatment phase (Brossart et al., 2006; Gibson & Ottenbacher, 1988; Knapp, 1983; Manolov et al., 2010; Matyas & Greenwood, 1990; Parker et al., 2009). The authors considered these graphs defensible as they are the cornerstone of other designs, but when combined with prefabricated data, raters judged graphs rarely seen in real-world settings. In addition, some graphs were selected or created to show certain types of effects (i.e., "obvious" or "non-obvious" effects), whether fabricated or previously published, creating a non-representative sample of real-world data (e.g., Jones et al., 1978).

Several studies were conducted with limited sample sizes in either graphs or participants and sometimes had as few as eleven raters or six graphs (Ottenbacher, 1990a; Parker & Hagan-Burke, 2007); some researchers deliberately used inexperienced raters (Gibson & Ottenbacher, 1988; Ottenbacher, 1990a).

Statistical Aid Studies. Research into statistical aids for visual analysis has also been limited by statistical and design challenges. The statistics were either impractical as they required a large number of data points, the absence of autocorrelation, the independence of error terms, or a random start point (Jones et al., 1978; Park et al., 1990; Stocks & Williams, 1995) or simply did not meet the assumptions of single-case data (Stocks & Williams, 1995). Other researchers did not capture the full complexity of single-case data as they focused on just mean shift or just changes in trend (Brossart et al., 2006; Jones et al., 1978; Ottenbacher, 1990b). In Ottenbacher (1990b) and Stocks and William (1995), the researchers tested the accuracy of visual analysis against their chosen statistic, instead of the more appropriate approach of testing their statistic against visual analysis.

Additionally, much of the time, the statistics simply did not agree with visual analysis at acceptable levels to become a suitable judgmental aid; agreement was often around chance

levels:

- Split Middle Trend: *IRA* = .46 (Ottenbacher, 1990b)
- Time Series Analysis: P_A = .50 (Jones et al., 1978)
- Randomization test: P_A = .67 for non-significance, P_A = .13 for significance, when all data were previously published and found to be significant (Park et al., 1990).

The various effect size statistics investigated also had their own specific limitations. The first limitation was the calculation of effect sizes. Users had to decide which conditions to contrast (e.g., baseline vs. treatment) and which data points to use (e.g., all points in a phase vs. the final three). Another limitation was the nature of the effect size itself – although the effect size resulted in a number indicating the strength of the effect, there was no set scale by which to judge results, and interpretations could change based on the practical implications of the data. Effect sizes, especially regression-based contrasts, also make the same assumptions as many of the previously tested statistics: independence, normal distributions, and equal variance, all of which are typically violated in single-case design data (Parker et al., 2009).

Studies comparing effect sizes to visual analysis had limitations as well. Several researchers gave no details into how visual analysis was conducted nor the criteria used to compare effect sizes against visual analysis (Brossart et al., 2006; Manolov et al., 2010; Olive & Smith, 2005; Parker & Hagan-Burke, 2007). The authors of these studies discussed the accuracy of using the effect sizes without providing concrete evidence of their utility. In two studies researchers compared their effect sizes to visual analysis using a very small number of graphs (e.g., 5-10; Olive & Smith, 2005; Parker & Hagan-Burke, 2007), which likely led to low power. Parker et al. (2009) and Parker and Hagan-Burke (2007) compared IRD to other effect sizes, not to visual analysis, so no comparison to the "gold standard" of single-case analysis could be

made.

Researchers who did directly equate effect sizes and visual analysis found results similar to the other statistics in that agreement was around chance levels. The average R^2 found by Brossart, et al., (2006) was 0.46. PEM and PND had Spearman correlations of 0.57 and 0.49 with visual analysis, respectively (Brossart et al., 2006).

Multilevel Modeling

An alternative approach to the statistical methods described above is Hierarchical Linear Modeling (HLM), a multilevel modeling method that allows predictors to vary within nested levels (Raudenbush & Bryk, 2002). Instead of all predictors modeling the outcome variable in the same level, as in a linear regression, an HLM model could have student academic achievement scores modeled by student factors like socioeconomic status, nested within school level variables like average school achievement, and further nested within district level factors like urban versus rural locations. Nesting these predictors is appropriate, as explained by Osborne (2000), because students from high achieving schools in urban districts are likely more similar (due to shared experiences and geographic factors) than they are to students from low achieving schools in rural districts. Because HLM can nest variables, it also provides for cross-level interactions between predictors. Other statistics would require a decision about the lowest level unit of analysis to use. For example, other statistics might require only school-level and higher predictors be used, excluding student-level data and potentially misestimating the relationships between the predictors (Osborne, 2000). HLM is also able to accommodate error within each nested level while partitioning out the error from other levels (Uekawa, 2012) and can use various error terms with random and nonrandom effects (Raudenbush & Bryk, 2002).

As well as modeling the effects of predictors on a single outcome measure, HLM can also

accommodate analyses of multiple observations over time (Raudenbush & Bryk, 2002). In the school-based example described above, repeated measures of achievement scores could be modeled with the addition of a time-based predictor. Other benefits of this type of model include accounting for both the initial level and growth of the outcome variable over time, accommodating missing data and unequal intervals between measurements, and not needing a large number of data points to run successfully. HLM can also accommodate autocorrelation by testing and specifying the correct error term in the model equation, thereby overcoming a problem encountered by many other statistics and one inherent to single-subject designs.

Several models have been proposed for single-subject designs. One of the most basic proposed by Kyse, Rindskopf, and Shadish (in submission) models single-subject outcome data (i.e., measurement of a DV within a given graph) using predictors representing the occasions when measurements were taken, the intercept of the outcome when the time point equals zero, the rate of change or slope in the DV over time, and any unexplained variance. This type of basic model takes into account multiple observations, the initial level of the outcome variable, and any trend or slope across subsequent time points. HLM is similar to visual analysis in that both level and trend across time are accounted for within the model (level in this model has a different definition than the typical single-case definition of level, but HLM can be adjusted to better reflect the visual analysis use of this concept). HLM also accounts for variability within and across nested levels.

Beyond the simple model described above, predictors can be added to account for the effects of person-level variables like gender or intelligence test scores, or study-level predictors like differences in experimental procedures. One predictor of interest to almost all single-subject designs is the effect of condition on the outcome variable. Kyse et al. (in submission), Van der

Noortgate and Onghena (2007) and Waddell, Nassar, and Gustafson (2011) each propose the effect of condition be modeled on the lowest level with the explanation that it is a time-based predictor; however, Godbold (2008) and Ployhart and Vandenberg (2010) propose condition be modeled as a higher level predictor, with the outcome variable and the effect of time nested within experimental conditions.

Even though HLM appears ideally suited to analyze single-subject data, exploration into the utility of HLM for single-subject design analysis has been limited. Kyse at al. (in submission) presented models for single-baseline ABAB and multiple-baseline AB designs, which were also included in a detailed instructional manual available from the authors (Nagler, Rindskopf, & Shadish, 2008). These models were much like those shown in Van der Noortgate and Onghena (2007) and Waddell et al. (2011) and were purely explanations of how to model simple single-subject designs with one to two studies as examples. Little research exists to support its widespread use as a statistical aid.

Godbold (2008) investigated the utility of HLM with eleven types of single-case designs from previously published graphs and compared the results to visual analysis ratings of the same graphs. In doing so, the study attempted to determine the utility of HLM as a statistical aid while also rectifying many limitations of previous studies of statistical and visual analysis. Ninety-six raters highly trained in visual analysis (94% of participants were certified by the Behavior Analyst Certification Board and 6% were previously certified) were presented with a questionnaire of 39 graphs. These graphs were randomly selected from six major research journals by strata (single-subject design type and the author's original interpretation of each graph, e.g., extremely, moderately, or not at all certain of an effect of the IV on behavior) to ensure a representative sample of previously published data.

Raters were asked to judge each graph based on level, trend, and the graph as a whole using a 5-point Likert scale ("Not At All Certain" to "Extremely Certain") to determine how certain they were the interventions presented caused a change in behavior. Overall rater reliability, calculated using Pearson *r*, was .46 for level, .46 for trend, and .43 for the graphs as wholes, $p < 0.05$, indicating a high level of variance between raters. These results corroborated previous research, which found even expert analysts could produce substantial variance in ratings.

These graphs then underwent statistical analysis using HLM. The specific model accounted for change over time in level, trend, and both combined, with the time points designated as those on the graph's x-axis. These variables were nested within the experimental conditions in the graph. Error terms were tested and the term that best fit each graph's data was used. One specifically accommodated autocorrelated data; the other was used when little autocorrelation was present. HLM successfully modeled 28 of the graphs, indicating the utility of the model in fitting single-subject data.

HLM's classification accuracy was then compared to visual analysis using proportional agreement and conditional probability tables. The comparisons were made individually for level, trend, and the graphs as wholes. HLM was most accurate at classifying significant effects in regards to level (sensitivity increased as average visual analysis ratings increased and specificity decreased) and most accurate at classifying non-significant effects in regards to trend (sensitivity decreased as average visual analysis ratings increased and specificity increased). When considering level and trend, HLM's accuracy reflected the dual criteria necessary for significance (level and trend both had to be significant for graphs as wholes to be) and followed the same basic pattern as when considering trend alone by most accurately classifying non-

significant effects. Overall, HLM was found to be less conservative than visual analysis when classifying level effects and more than or equally as conservative as visual analysis in classifying trend effects. In addition, when judging the graphs as wholes, HLM showed almost the same level of agreement with visual analysis as found between visual analysts in a Knapp (1983) study (P_A=76% and 80%, respectively), with the added benefit of statistical reliability (Godbold, 2008).

Although the results found in Godbold (2008) are promising, limitations of this study are the focus of the present study. The first major limitation is the use of graphs containing multiple IVs and DVs. Raters were asked to rate level, trend, and the graph as a whole while considering the effects on *all* DVs and IVs present in the graph. To compensate, HLM analysis was conducted analogously, with "significance" only occurring if HLM rated the changes in each DV significant across all IVs. Graphs' data were thus "averaged" for both visual and statistical analysis. This limitation presents a potential point of ambiguity and many raters commented on the impractical nature of the rating. In addition, asking raters to judge the graph as "whole" may not have been equivalent to combining the HLM results for both level and trend. Judging the graph as whole may have led raters to consider other aspects of the graphs beyond level and trend, such a physical presentation, which were not included in the HLM analysis.

Other limitations of the study were restrictions of HLM found when analyzing the single-single data, which reduced the usable sample of graphs and the power of any subsequent analyses. HLM required multiple instances of each phase type to model change correctly, so each graph had to include at least two baseline and treatment phases (AB graphs and functional analyses could not be used). In addition, to model trend correctly, graphs had to contain at least three data points per phase or a large number of overall data points. HLM also required some

data variability for the models to run correctly, so graphs with little variability could not be accommodated.

Another restriction was multiple baseline graphs were accommodated by running the Level 2 model sequentially (e.g., $A_1B_1A_2B_2$ for a graph with two AB baselines) as opposed to adding another predictor to model the effects of different baselines. The length of the visual analysis survey (forty minutes) could have lessened the size of the respondent sample and ceiling effects were found with the 5-point Likert scale, as no graphs were rated at the high end of the scale.

Purpose and Rationale of the Current Study

The current study was designed rectify many of the limitations of the above study while extending the rater pool as well as the graph types accommodated by HLM. First, and most significantly, raters were asked to evaluate the effect of one IV and one DV per graph and HLM analyzed each dataset accordingly. Second, in addition to asking raters to judge the graph as a "whole," they were also asked about the combined effects of level and trend to equate the analyses further. The visual analysis survey was shortened and the potential pool of raters expanded to increase the number of potential respondents, and the Likert scale was expanded to limit the ceiling effects found previously.

To extend research into HLM, two different methods of HLM modeling were used: HLM2 and the Hierarchical Multivariate Linear Model (HMLM). Kyse (in submission) solely used HLM2; HMLM was used in Godbold (2008). The two models differ in that HLM2 is a simple two-level linear model, whereas HMLM is a longitudinal model that allows error term specification and can accommodate incomplete data. Because both models can successfully analyze single-subject data, the current study investigated the utility and accuracy of each model

to determine if one of the models was more suited to single-subject data. Sufficient power was ensured by creating an initial sample of graphs fitting the requirements of HLM as found in Godbold (2008). Additional graphs were then selected to investigate the extension of HLM to single-case data not accommodated by the original model (e.g., functional analysis).

METHOD

Visual Analysis Survey

Participants. Survey respondents were recruited from school psychology and applied behavior analysis list serves, discussion forums, and social media websites, including four National Association of School Psychologists[SM] list serves, four Applied Behavior Analysis International® Special Interest Group list serves, two general autism and verbal behavior discussion forums, and three social media groups for school psychologists and applied behavior analysts. These groups had average memberships of 1,600 individuals for approximately 19,000 total members. However, it is expected not all members read the recruitment posting and some individuals were likely members of multiple groups. Additional participants (approximately 200) were recruited from psychology and applied behavior analysis graduate programs and internship sites. The above groups were chosen for recruitment based on their educational diversity and members' presumed familiarity with single-case data and visual analysis. Participants had two weeks to respond to recruitment postings.

Graph Database and Selection. A database of 794 graphs from 268 articles was generated using single-subject design articles published in psychological journals between January 2002 and December 2006. The journals were *Behavioral Disorders, Behavior Modification, Child and Family Behavior Therapy, Journal of Applied Behavior Analysis, Journal of Special Education,* and *School Psychology Review.* These journals publish single-case designs in the areas of applied behavior analysis, special education, and clinical and school psychology.

Graphs had to meet specific criteria to be included in the database: legibility, clear scales showing the rate or level of behavior on the y-axis and time or sessions on the x-axis, and scales

with equal intervals. These criteria are consistent with past studies using previously published data, such as Gresham, et al. (2004). The first five graphs from each article were coded according to design type (single and multiple-baseline AB, ABAB reversal, ABAB withdrawal, multielement, and "other" designs, as well as functional analysis graphs). Coders were trained on the different types of graphs as well as the basic graph criteria listed above. Functional analysis graphs were coded separately as they generally utilize a consistent and accepted experimental method; graphs coded as "multi-element" often included aspects not typical to a basic functional analysis – baseline phases, reversals, and replications. Changing criterion designs were not included, as they generally do not contain a true baseline or alternate condition against which treatment effects can be compared.

Graphs were also coded by the author's original statement of certainty about the behavior change present in the graph. The author's original statement of certainty was coded as the graph showing 1) a clear, unambiguous, certain effect, 2) an ambiguous, unclear, uncertain effect, or 3) no effect. Coders were trained on each rating aspect and inter-rater agreement was measured for 33% of coded graphs. Inter-rater agreement was 93% [*Agreement = (Number of agreements/Number of agreements + Disagreements)*100*].

A stratified random sample was used to select the graphs for survey inclusion based on design type and the author's statement of certainty (Table 1). Because not every graph type had a graph for each statement (there were very few graphs coded as showing "no effect"), 80% of potential graph types could be used. The final sample contained 80 graphs, and tests using G*Power indicated this sample size would be more than adequate for finding a modest effect size ($\beta = 0.80$; Faul et al., 2007). Fifty-three graphs (none from the original study) were selected. An additional 27 graphs from the original study were also included. Only graphs containing one

DV and condition comparison (i.e., baseline versus treatment or treatment A versus treatment B) were selected from the original study to control for any effects of additional data on rater judgments. These graphs were included to allow for a more meaningful comparison of rater accuracy between the two studies,

Survey Descriptions, Images, and Questions. Information about each graph was prepared in short paragraph form and contained the following items: (a) operational definitions of all DVs in the graph, (b) indication of the type of DV data collection (continuous recording, momentary, partial interval, whole interval, etc.), (c) method of data collection (frequency, intensity, or duration), (d) duration of data collection (5-min, 10-min, all day, etc.) (e) data presentation (rate, count, percentage, etc.), (f) indication of the frequency of measurement (by session, daily, weekly, etc.), and (g) experimental procedures for all phases presented in the graph. DVs and conditions were labeled using the names provided on the graph. The information about each graph was evaluated by two doctoral-level BCBAs for accuracy and understanding and edited as needed.

Images of each graph were reproduced using high quality images. The images were modified to remove trend or average lines to prevent any effect of graphical aids on survey ratings. Additionally, x and y-axis labels were added or moved, if necessary. For example, if a graph was selected from a figure containing multiple graphs, the label was moved so it could be seen next to the selected graph. In addition, if necessary, captions were modified to remove extraneous information. Example modifications were removing descriptions of unused graphs or additional participant names. Again, these changes generally occurred with figures containing multiple graphs. These modifications were done to ensure all necessary information was present for raters and potentially confusing information was removed.

Table 1.

Sample Graphs by Type and Author Statements of Certainty

	Author Statement of Certainty			
Design type	No effect	Uncertain effect	Certain effect	Total
SB AB	--	1.25	1.25	2.5
SB ABAB (reversal)	--	1.25	3.75	5
SB ABAB (withdrawal)	--	5	8.75	13.75
SB ABC, etc.	--	7.5	18.75	26.25
SB Multi-element	1.25	3.75	7.5	12.5
MB AB	2.5	3.75	11.25	17.5
MB ABAB (reversal)	--	2.5	--	2.5
MB ABAB (withdrawal)	--	1.25	2.5	3.75
MB ABC, etc.	--	2.5	6.25	8.75
MB Multi-element	--	1.25	1.25	2.5
Functional analysis	1.25	1.25	2.5	5
Total	5	31.25	63.75	

Note. SB indicates a single-baseline design; MB indicates a multiple-baseline design. Cells with data indicate the percentages of graphs in the final sample. -- indicates no graphs in the database were of that graph type and certainty level. $N = 80$.

Graph images were uploaded to the survey site using set size parameters. Each graph was made as large as possible while still fitting comfortably on a computer screen (approximately 650 pixels by 400 pixels). However, although the images were a consistent size, graphs with one baseline appeared more magnified than graphs with multiple baselines.

The visual analysis survey was hosted by http://www.psychdata.com, a paid hosting

service meeting Institutional Review Board research standards (Psychdata™, 2012). The survey included 10-15 demographic questions and one question about survey participation. To reduce participant response effort to the survey, the 80 graphs were randomly divided into 16 sets of 5 graphs each; the graphs were then randomly ordered within each set. At the conclusion of the demographic questions, each participant was presented with one of the sixteen sets using quota assignment. Participants were randomly assigned to a set until each set had at least 10 participants.

For each graph, DVs and target conditions were randomly selected to create a more specific comparison point for both visual analysis ratings and HLM. No selection was made if a graph contained one DV and either one baseline and one treatment condition or two treatment conditions. If a graph contained multiple DVs, one was selected as the target variable. Additionally, if a graph contained a baseline phase and multiple treatment conditions, one treatment condition was randomly selected to compare to baseline. If a graph contained no baseline but multiple treatment conditions, two treatment conditions were selected.

Participants were given detailed instructions and if necessary, the target comparison and DV for each graph was clearly stated. The participants were then asked to judge the target comparison based on changes in level, trend, both level and trend, and all aspects of the graph. The questions about level, trend, and both level and trend was included to be analogous to the HLM tests of each graph; the question about all aspects of the graph was included to help identify if participants changed their ratings due to other graphical features, such as variability. The ratings given to these four questions could then be compared to the statistical analysis of the graphs.

The questions appeared as follows:

Based on the information provided by the graph, how certain are you that the *Target 1* condition caused a change in behavior as compared to *Target 2* for the participant?

Based on changes in **level** only

Based on changes in **trend** only

Based on changes in **both level and trend**

Based on **all aspects** of the graph

For graphs with multiple DVs, the question was changed:

Based on the information provided by the graph, how certain are you that the *Target 1* condition caused a change in *Dependent Variable* behavior as compared to *Target 2* for the participant?

The question also changed for graphs with multiple baselines:

Based on the information provided by the graph, how certain are you that the *Target 1* condition caused a change in behavior as compared to *Target 2* across *all participants (settings, behaviors, etc.)*?

Participants judged these changes using a six-point bipolar Likert scale. Ceiling effects were found with the 5-point Likert scale used previously, as no respondents said they were "Extremely Certain" about intervention effects in the graphs (Godbold, 2008). The 6-point scale was chosen to help alleviate any range constraints imposed by the 5-point scale. Discrete response categories were also included for each numeral to increase validity and reliability (Weng, 2004). An even-numbered scale was chosen to create a "forced-choice" situation for participants because they could not pick a "neutral" or "uncertain" central position, as a large number of participants choosing a scale midpoint would prevent the results from being accurately compared to HLM. The Likert scale values were as follows:

(-3) Very Certain it ***did not*** cause a behavior change

(-2) Certain it ***did not*** cause a behavior change

(-1) Slightly Certain it ***did not*** cause a behavior change

(1) Slightly certain it ***did*** cause a behavior change

(2) Certain it ***did*** cause a behavior change

(3) Very Certain it ***did*** cause a behavior change

The questions all referenced participants' certainty that a change in behavior occurred to ensure raters did not solely rely on social validity judgments or the significance of the change. Figure 2 shows the images, descriptions, and questions as presented to participants.

Hierarchical Linear Modeling

Data Extraction. DigitizeIt© (Bormann, 2010), a computer software program, was used to extract data point values from each graph. DigitizeIt© uses scanned image files to provide data values based on the x and y-axes. Shadish, Brasil, Illingworth, White, and Galindo (2009) found such programs have good reliability and validity when graphs are clear and program users are trained and able to detect different phases, treatments, etc., in single-case designs.

To extract data, graphs were scanned into the DigitizeIt© program and magnified to at least 200% of the original image. The x and y-axes were defined, and data points were noted by clicking in the middle of the desired data point (Figure 3). If the cursor could not be centered on a data point, coders moved their cursor either one pixel up or to the right, depending on whether the problem with centering was vertical or horizontal, so any "error" (although typically only by one pixel) was consistent across coders.

Data point values were then converted by the program and exported to spreadsheet software. The values extended to several decimal points but were reduced to two decimal points

1. Based on the information provided by the graph, how certain are you that the *No Attention* condition presented caused a change in behavior as compared to *Attention* for the participant?

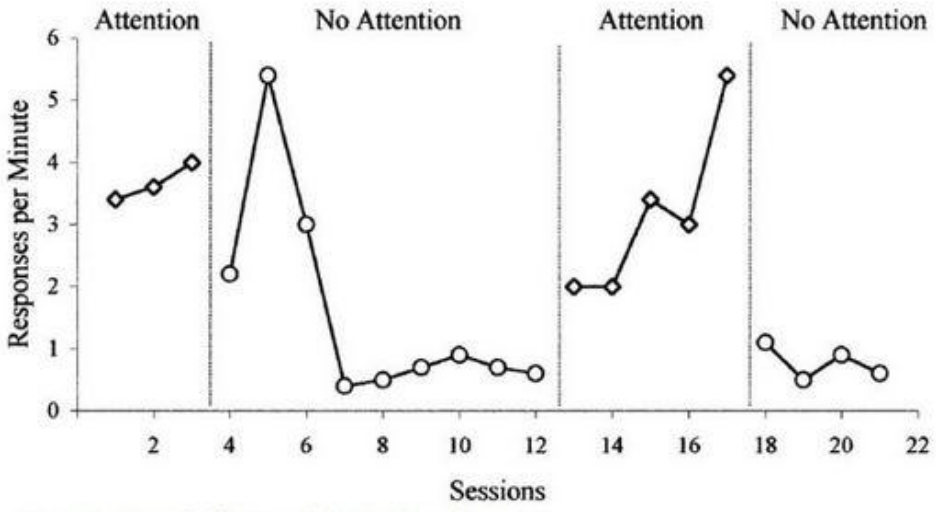

Figure 1. Rate of self-injurious behavior during the analysis.

77-2-3-2 A

Behavior

Responses per Minute: Responses were instances of self-injurious behavior during 10-min sessions in an analogue setting. Self-injurious behavior was defined as any instance of the participant striking any object with her head, forceful contact between her head and hand, or closing her teeth around any portion of her wrist or hand. Total instances of self-injurious behavior divided by total minutes (10 min).

Phases

Attention: The participant was instructed to play quietly with low to moderately preferred toys while the therapist read a magazine. Contingent on self-injurious behavior, the therapist gave a brief verbal reprimand on a Fixed Ratio 1 schedule.

No Attention: The participant was instructed to play quietly with low to moderately preferred toys while the therapist read a magazine. No attention was given to the participant regardless of behavior.

	(-3) Very Certain it *did not* cause a behavior change	(-2) Certain it *did not* cause a behavior change	(-1) Slightly Certain it *did not* cause a behavior change	(1) Slightly Certain it *did* cause a behavior change	(2) Certain it *did* cause a behavior change	(3) Very Certain it *did* cause a behavior change
* Based on changes in **level** only	○	○	○	○	○	○
* Based on changes in **trend** only	○	○	○	○	○	○
* Based on changes in **both level and trend**	○	○	○	○	○	○
* Based on **all aspects** of the graph	○	○	○	○	○	○

Figure 2. Example survey graph with definitions and questions. This figure illustrates questions as presented in the visual analysis survey.

in the case of percentages, responses per minute, or other behaviors not conducive to whole numbers. In cases where whole numbers were known to be used in the study (e.g., frequency counts), the data were rounded to the nearest whole number. The entire data file was then labeled with the appropriate DV and conditions.

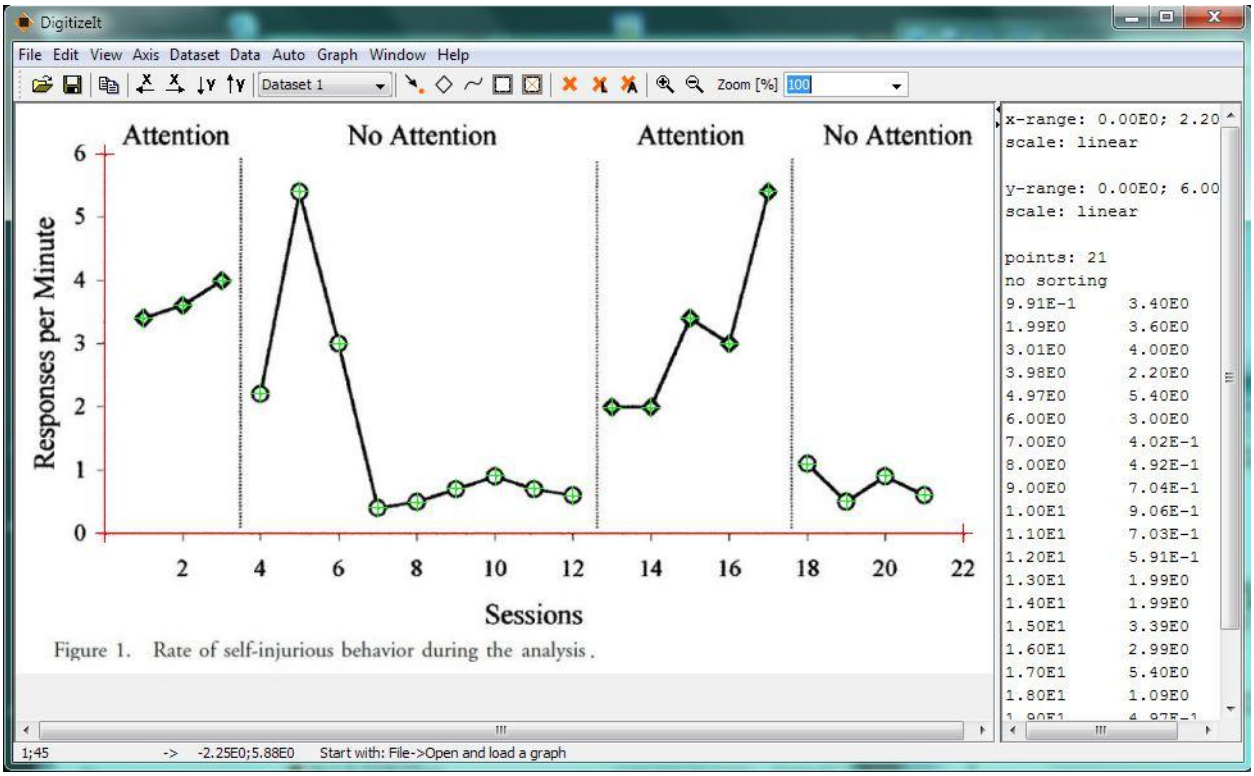

Figure 3. Example data point extraction. This figure illustrates the use of the DigitizeIt© data extraction program.

Inter-rater agreement was calculated on 34% of graphs using interclass correlations. Pearson *r* was selected due to the large number of data points and the continuous nature of the data point values (Garson, 2011; Wuensch, 2007). Average data point value agreement was *r* = .99; the range was *r* = .85 to *r* = 1.00. Ninety-six percent of graphs had *r* values greater than .90.

Single-Baseline Model. Statistical analysis of graph data was conducted using the HLM7 Student Edition program, (HLM7S; Raudenbush, Bryk, & Congdon, 2011). Data were analyzed using a 2-level HLM2 model and a 2-level HMLM model. Both models included the same predictors within the same basic structure and all single-subject designs were tested with this model:

Level 1:

$$Data_i = \pi_{0i} + \pi_{1i}(Time_{ti}) + e_{ti}$$

Level 2:

$$\pi_{0i} = \beta_{00} + \beta_{01}(Condition_i) + r_{0i}$$

$$\pi_{1i} = \beta_{10} + \beta_{11}(Condition_i)$$

In these equations, the outcome variable, $Data_i$, was the single-case data extracted from a particular graph. Each data point within the graph was coded as belonging to a particular phase or group (graphs with ABAB designs had four sequential groups). The data points were then coded according to where they fell within the graph using the time points from the graph's x-axis ($Time_{ti}$) and by the conditions they were nested within – either Baseline versus Treatment or Treatment 1 versus Treatment 2 ($Condition_i$).

In the Level 1 equation, π_{0i} and π_{1i} represent Level 2 equations, as explained below. $Time_{ti}$ is the effect of time on $Data_i$ as coded from the x-axis, and e_{ti} is an error term accounting for any variance not associated with the predictors $Time_{ti}$ and $Condition_i$.

The first Level 2 equation, π_{0i}, represents the main effect of $Condition_i$ on group means: β_{00} is the average intercept of the groups and $\beta_{01}(Condition_i)$ represents the relationship between group mean and $Condition_i$. The variable r_{0i} represents the error term, which was allowed to randomly vary in this equation (Decoster, 2002).

The second Level 2 equation, π_{1i}, represents both a main effect and an interaction: β_{10} represents the average group slope, or the main effect of trend, and $\beta_{11}(Condition_i)$ represents the relationship or interaction between group slope and $Condition_i$ (Decoster, 2002). Note that this equation is considered to have nonrandomly varying slopes as evidenced by the lack of an error term. No additional variation was allowed in slope because HLM determined there was no significant variation beyond that accounted for by the addition of $Condition_i$ to the model (Sarkisian, 2007b).

These equations can be combined into the following equation, called a Mixed Model (Raudenbush et al., 2011):

$$Data_i = \beta_{00} + \beta_{01}(Condition_i) + \beta_{10}(Time_{ti}) + \beta_{11}(Condition_i)(Time_{ti}) + r_{0i} + e_{ti}$$

Multiple-Baseline Model. The model for multiple baseline designs included an additional predictor, $Baseline_i$, to account for any variation associated with a graph containing more than one baseline. $Baseline_i$ was included as a Level 2 predictor because, as a theoretical consideration, graph conditions are not necessarily nested within different baselines (or vice versa) as would be the case with a 3-level model. For example, consider a traditional 3-level student-classroom-school model. If the target classrooms in each school were identical, with all variables held constant across each classroom, including the teacher, would classroom still be appropriate to nest under school, or would be it more appropriate to consider the effects of both at an equal level? The experimental procedures of different single-subject conditions are typically held constant across different baselines, even when, for instance, the experimental setting or the participant changes, and so can be argued as a similar situation. $Baseline_i$ as a Level 2 predictor accounted for variation due to different baselines without being nested above or within $Condition_i$ potentially inappropriately; the variable was also coded to include information about the order of baselines within the graph. The equations for multiple baseline graphs were as follows:

Level 1:

$$Data_i = \pi_{0i} + \pi_{1i}(Time_{ti}) + e_{ti}$$

Level 2:

$$\pi_{0i} = \beta_{00} + \beta_{01}(Condition_i) + \beta_{02}(Baseline_i) + r_{0i}$$

$$\pi_{1i} = \beta_{10} + \beta_{11}(Condition_i) + \beta_{12}(Baseline_i)$$

Mixed Model:

$$Data_i = \beta_{00} + \beta_{01}(Condition_i) + \beta_{02}(Baseline) + \beta_{10}(Time_{ti}) + \beta_{11}(Condition_i)(Time_{ti}) + \beta_{12}(Baseline_i)(Time_{ti}) + r_{0i} + e_{ti}$$

Model Selection. The single and multiple-baseline models are analogous to "Slopes-and-Intercepts-as-Outcomes" or "Cross-Level Interaction" models (Sarkisian, 2007b). This model is not the most parsimonious allowed in HLM, but is instead a general model accounting for the effects of both $Time_{ti}$ and $Condition_i$ on level and trend and is potentially applicable across a wide variety of single-subject design types. Additionally, random variation is allowed for both intercepts and slopes, and cross-level interactions are included (e.g., $Condition_i$ * $Time_{ti}$). One approach to investigating the appropriateness of this model begins with an Unconditional model where $Data_i$, the outcome variable, is modeled by only intercepts and group means:

Level 1:

$$Data_i = \pi_{0i} + e_{ti}$$

Level 2:

$$\pi_{0i} = \beta_{00} + \beta_{01}$$

Predictors are then added to the model, and after each addition, variances are examined for an increase in explained variance due to each new predictor (Sarkisian, 2007b); predictors remain in the model only if they significantly increase the explained between and within-group variance. The models may increase to and beyond the Slopes-and-Intercepts model depending on the predictors available.

This approach was conducted with eight randomly chosen datasets to test if the addition of $Time_{ti}$, $Condition_i$, and $Baseline_i$ would increase the explained Level-1 and Level-2 variance

for both HLM2 and HMLM models. For both HLM2 and HMLM, addition of $Time_{ti}$ to the Level-1 equation explained an average of 56% more within and 49% more between-group variance in $Data_i$ over the Unconditional model; the addition of $Condition_i$ to both Level-2 equations explained an average of 9% more within and 44% more between-group variance than the model with just $Time_{ti}$ as a predictor. Two multiple-baseline graphs were also tested, and the addition of $Baseline_i$ to the model with $Time_{ti}$ and $Condition_i$ did not explain anymore within group variance than but did explain 3% more between-group variance, for 53% more within and 98% more between-group variance explained than the Unconditional model. The Slopes-and-Intercepts-as-Outcomes model was deemed an appropriate model as the addition of the chosen predictors consistently explained more variance than simpler models.

Predictor Centering. HLM allows for predictors to be centered in one of three ways: uncentered (the predictors are expressed as raw scores), grand mean centered (the predictors are expressed as deviation around a grand mean), and group mean centered (the predictors are expressed as deviations around a group mean; Decoster, 2002). The choice of centering affects the interpretation of the intercept and slope terms. In this study, uncentered predictors would mean the value of π_{0i}, the intercept term, is the predicted value of $Data_i$ when $Time_{ti}$ equals zero. To center the predictor, a value is subtracted from $Time_{ti}$ to make this zero point more meaningful to the research question. With grand mean centering, π_{0i} represents the group mean on *Data* adjusted for the entire graphs' average value on $Time_{ti}$. With group mean centering, π_{0i} would represent each group's mean on *Data* adjusted for the individual group's average values on $Time_{ti}$ (Decoster, 2002; International, 2012a).

$Time_{ti}$ was group mean centered to create a standard initial level of behavior for each group and across each graph. Additionally, group mean centering meant the average value of

Data within each phase was considered in the model and the effect of $Condition_i$ on the average level of a phase could be tested, which is most analogous to the concept of level in single-case design. Group mean centering also helped reduce any potential collinearity between predictors; high collinearity in the model (or a predictor being linearly dependent on another predictor, International, 2012b) can negatively affect HLM model specification. The variables $Condition_i$ and $Baseline_i$ were uncentered, as the zero point for these variables was meaningful without centering.

Dependent Variable Distributions and Overdispersion. The HLM2 model analysis required the distribution of the outcome variable to be specified. The Poisson distribution is more appropriate than continuous for count or rate DVs (Agresti, 1996) and was specified appropriate. Another distribution, Binomial, was either not appropriate for the data or, for a very small number of datasets, the articles did not provide enough information to correctly specify the Binomial distribution within the HLM7S program. In these instances, the continuous distribution was specified instead.

When using the Poisson distribution, overdispersion should be considered (Agresti, 1996). If the data display greater variability than expected by the distribution overdispersion occurs, and assumptions about relationships between the mean and variance of the distribution can be violated. When running a model accounting for overdispersion, HLM7S will list the within-subject variance in the output file. If this variance is greater than 1.0, the data are likely overdispersed, and the overdispersion model should be used. If the within-subjects variance is less than 1.0, the data are not overdispersed and the simpler model not accounting for overdispersion can be used (Raudenbush, 2004). To test for and accommodate potential overdispersion, datasets using the Poisson distribution were analyzed using both models, and the

appropriate model was chosen based on output.

Error Term Modeling. The HMLM analysis allowed for the specification of the error term model. Two models for the error terms were tested, the Homogeneous model and the First-Order Autoregressive model. The Homogeneous model is based on fewer underlying parameters than the First-Order Autoregressive model and assumes that covariances are equal and there are equal variances at each time point. The First-Order Autoregressive model assumes independent variances and that autocorrelation might be present in the data (Raudenbush & Bryk, 2002). HLM7S modeled $Data_i$ using each error term and indicated the more appropriate model by showing the deviance and degrees of freedom associated with each model. The better model was indicated by a lower deviance term and higher degrees of freedom (Raudenbush & Bryk, 2002). HLM7S also provided a chi square statistic indicating whether the fit of the models was significantly different. In cases where the models were significantly different, the better fitting model was used. In cases where the models were not significantly different, the Homogeneous model was used, as it was the more parsimonious model.

Hypothesis Testing. HLM7S may be used for hypothesis tests to determine the effect of a variable or variables on the outcome data. In both the HLM2 and HMLM models, the effect of $Condition_i$ (i.e., Baseline and Treatment or Treatment 1 and Treatment 2) was tested using three hypothesis tests. The first two tests were single parameter and tested the effect of $Condition_i$ on level and trend separately, $H_o: \pi_{0i} = \emptyset$ and $H_o: \pi_{1i} = \emptyset$. Each tested the hypothesis that level (or trend) was not significantly different for each group due to $Condition_i$, that is, phases were either similar in their group means or in their slopes (Sarkisian, 2007a). The third hypothesis $H_o: \pi_{0i} = \pi_{1i} = \emptyset$ tested the effect of $Condition_i$ on both level and trend simultaneously ("Introduction to multilevel modeling using HLM," 2012; Sarkisian, 2007a).

HLM7S provided chi square statistics, degrees of freedom, and *p* values for each hypothesis test, which led to three conclusions about each graph (a) HLM level: whether level was significantly different due to $Condition_i$ across groups ($p \leq 0.05$), (b) HLM trend: whether trend was significantly different ($p \leq 0.05$), and (c) HLM level and trend: whether both level and trend were significantly different ($p \leq 0.05$). Figure 4 provides a sample output of an HLM

Results of General Linear Hypothesis Testing - Test 1

	Coefficients	Contrast
For INTRCPT1, β_0		
INTRCPT2, γ_{00}	3.345001	0.0000
COND, γ_{01}	-2.000388	1.0000
For SESSION slope, β_1		
INTRCPT2, γ_{10}	0.701667	0.0000
COND, γ_{11}	-1.085590	0.0000

χ^2 statistic = 22.237330
Degrees of freedom = 1
p-value = 0.000037

Figure 4. Hypothesis test results. This figure shows an example output for the effect of condition on level in a single-baseline HMLM model.

hypothesis test for the effect of $Condition_i$ on level in an HMLM model. The coefficient for INTRCPT2, γ_{00} represents the average level of the Baseline group (due to the group mean centering chosen for $Time_{ti}$). The coefficient COND, γ_{01} represents the average change in Level due to the effect of treatment, and in this case, shows a decrease in level ("1.00000" under "Contrasts" indicates which predictor is being tested). The *p*-value 0.00037 indicates there was a significant change in level between the two conditions.

Comparison Analyses

HLM and Visual Analysis Comparisons. The visual analysis and HLM analysis results each contained several different components. The four visual analysis questions and three HLM conclusions generated four different comparisons for study across both models, for eight total

comparisons (Table 2). Comparisons were chosen based on their utility and appropriateness.

Table 2.

Visual Analysis Rating and HLM Model Comparisons

	VA Level	VA Trend	VA Level & Trend	VA All Aspects
HLM Level	Comparison 1	--	--	--
HLM Trend	--	Comparison 2	--	--
HLM Simultaneous	--	--	Comparison 3	Comparison 4

Note: -- Indicates this comparison was not analyzed.

Contingency Probability Tables. Contingency probability tables are often used in medicine to compare new diagnostic tests to a "gold standard," or an established test already used for the same diagnosis (Deeks, 2006). Measures of diagnostic accuracy are ideally suited to the current study, as a "new" test (HLM) was compared to the "gold standard" of visual analysis. Calculating the tables required the results of both HLM and visual analysis to be sorted into dichotomous variables. The results of the three HLM hypothesis tests for both the HLM2 and HMLM models were sorted using"1" for a significant result (any $p \leq 0.05$ from the hypothesis tests) and "0" for a non-significant result (any $p > 0.05$). Visual analysis was separated into two dichotomies using participants' average ratings for each graph. Because ratings 1-3 were given the qualifiers "(Very, Slightly) Certain a change did *not* occur," any average ratings falling between 1.0 and 3.494 (the designated midpoint between 1 and 6) were collapsed into a "non-significant" category coded as "0." Average ratings falling between 3.495 and 6 were collapsed into a "significant" category coded as "1."

A contingency probability table was generated for each comparison for both HLM2 and HMLM (Table 2). The tables allowed the calculation of Sensitivity (true positives, or the probability HLM resulted in a significant *p* value when analyzing a graph rated "significant"),

Specificity (true negatives, or the probability HLM resulted in a non-significant *p* value when analyzing a graph rated "non-significant"), and Overall Accuracy, the proportion of the total correction classifications (true positives and true negatives) compared to all classifications.

The Positive Likelihood Ratio (PLR) was also calculated and was a measure of how much more likely HLM was to return a significant *p* value in graphs given "significant" visual analysis ratings compared to those given "non-significant" ratings. The Negative Likelihood Ratio (NLR) was how much less likely HLM was to return a non-significant *p* value in graphs rated "significant" in the survey as compared to graphs rated "non-significant." Likelihood ratios greater than one would indicate a result from HLM was associated with a significant visual analysis rating, and a ratio less than one would indicate the HLM result was associated with a non-significant rating; ratios equal to or close to one would have no comparison utility (Beardsell, Bell, Rumbold, & Robinson, 2009). The statistics were calculated as follows, where *TP = True Positives*, *TN = True Negatives*, *FP = False Positives* and *FN = False Negatives*:

$$Overall\ Accuracy = \frac{TP + TN}{(TP + FP + TN + FN)}$$

$$Sensitivity = \frac{TP}{TP + FN} \qquad Specificity = \frac{TN}{TN + FP}$$

$$PLR = \frac{Sensitivity}{1 - Specificity} \qquad NLR = \frac{1 - Sensitivity}{Specificity}$$

Chi Square Statistics. Chi square statistics were chosen to quantify the accuracy of HLM to visual analysis as well as to provide a method for comparing both HLM models to each other. Chi squares test how likely the distribution of observed categorical data across different variables is due to chance. The null hypothesis of the chi square is that the observed distribution is due entirely to chance and the compared variables are not associated. Observed data counts

are compared to the counts that would be expected if the variables were independent. The more the observed counts deviate from the expected counts, the more likely the variables are dependent and associated in some way (Connor-Linton, 2003).

Chi squares are appropriate for the categorical data generated by the visual analysis ratings and HLM hypothesis tests (Howell, 2010). Additionally, the use of both Pearson chi squares and Linear-by-Linear chi squares can result in an assessment of whether the underlying relationship between HLM and visual analysis is linear or curvilinear. In general, chi squares do not account for the order of categorical variables (rows and columns can be rearranged without affecting the outcome); however, the ordinal nature of the visual analysis data makes order an important consideration, especially as not accounting for it can potentially decrease power (Agresti, 1989). Chi squares were chosen because they can accommodate this type of ordinal data given certain adjustments.

The data from the HLM analysis were classified dichotomously as with the contingency probability tables discussed previously. The average visual analysis ratings were separated into more categories than used previously to account for the ordinal nature of the data – six categories total, based on the average ratings of each graph (Table 3). These categories were then assigned an ordered metric (their corresponding Likert scale value) before being analyzed in the chi square tests (Figure 5). Initially, Pearson chi squares were conducted using an ordered metric reflecting all six visual analysis Likert scale values (1 – 6). However, due to the large number of cells created by having two HLM categories and six visual analysis categories, the chi square's assumption of at least five expected counts in a given cell were often violated, rendering these test unusable. An Exact *RxC* Contingency Table could have been used with this number of cells and with the small expected counts, but would not have provided as much information about the

		Visual Analysis						
		1	2	3	4	5	6	Total
HMLM	$p > 0.05$	5	15	4	1	3	2	30
	$p < 0.05$	2	5	10	3	16	14	50
Total		7	20	14	4	19	16	80

Figure 5. Example chi square test. This figure illustrates use of the ordered metric for visual analysis within the chi square tests.

variables' relationship. Therefore, two of the visual analysis categories were "rebinned," or re-sorted following the procedure outlined in Kirman (1996). Categories 1 and 6 were combined with their neighboring categories to create four visual analysis categories. Rebinning these categories was not likely to change the test results significantly, as few graphs (an average of 4.84% for each visual analysis factor) were rated within Categories 1 and 6 (Table 3).

After rebinning, the tests were conducted again. The results of the Pearson chi square tests were then compared to Linear-by-Linear chi squares (with the variables categorized in the same way) to provide an assessment of the relationship underlying the variables. Results of the Pearson chi squares indicated whether the variables were independent, whereas the Linear-by-Linear chi squares indicated if the relationship was linear, i.e., as visual analysis ratings increased, the likelihood of a significant HLM result increased. Typically, the Linear-by-Linear chi square is subject to the same principle as the Pearson chi square in that rearrangement of the columns and rows does not affect the outcome; however, the ordered metric for ratings allowed for the direction of the linear relationship to be tested (Howell, 2010).

Utilizing the Pearson and Linear-by-Linear chi squares also provided a test for a non-linear relationship (e.g., curvilinear) between the variables (Howell, 2010). By subtracting the Linear-by-Linear chi square statistic and degrees of freedom from the results of the Pearson chi square test, a curvilinear relationship was tested. The resulting "deviation from linear" *p* value

Table 3.

Likert Scale Values and Corresponding Average Rating Categories

Verbal Label	Likert Value/ χ^2 Metric	Average Rating Category	Percentage of Graphs in Original Categories				Percentage of Graphs in "Rebinned" Categories			
			L	T	LT	AA	L	T	LT	AA
Very Certain No Change	1	1 to 1.494	1.25	1.25	1.25	1.25				
							8.75	11.25	10	8.75
Certain No Change	2	1.495 to 2.494	7.5	10	8.75	7.5				
Slightly Certain No Change	3	2.495 to 3.494	26.25	27.5	26.25	27.5		same		
Slightly Certain Change	4	3.495 to 4.494	22.5	31.25	23.75	23.75		same		
Certain Change	5	4.495 to 5.494	32.5	26.25	30	30				
							42.5	30	40	40
Very Certain Change	6	5.495 to 6	10	3.75	10	10				

Note: L indicates the visual analysis factor level, T indicates the factor trend, LT indicates the both level and trend, and AA indicates all aspects. $N = 80$.

indicated if a linear relationship was present in the data (a non-significant *p* value) or if there was a curvilinear relationship (a significant *p* value; Howell, 2010).

In addition to the tests of independence and linearity, the Cramer's *V* Coefficient was used to test the degree of the relationship between the variables. The *V* Coefficient is appropriate for contingency tables larger than 2x2 and is not affected by sample size (Crewson, 2012). The statistic takes *df** into account, which is not the *df* associated with the chi square statistic (*R*-1)(*C-1*) but is instead the smaller of the two values (Gravetter & Wallnau, 2007). Cramer's *V* ranges from 0.0 to 1.0 and can be interpreted with the following guidelines when *df**= 1: 0 to .1

indicates little association between the variables, .1 to .3 indicates low association, .3 to .5 indicates a moderate association and greater than .5 indicates high association (Gravetter & Wallnau, 2007).

RESULTS

Survey Demographics

Participants who completed all survey questions were included in the results. Of the 298 participants who started the survey, 168 (69%) began answering questions about the graphs, and 168 (56%) completed the entire survey. Survey completion time was measured by participants clicking the first and last "submit" buttons on the survey website, and so was actually a measure of the amount of time the survey was open in a web browsers. Therefore, to determine completion time ten outliers were removed using scatterplot analysis. Average completion time was 15.5 min.

Most participants were women (64%) between the ages of 25 and 35 (53%) with Master's degrees (48%). All participants held at least a Bachelor's degree, and 32% held doctorates. The majority of their professional work was in the United States (88%, representing 33 states and the District of Columbia), with 4% working in Canada and 8% working in other countries, including Brazil, China, Israel, Mexico, New Zealand, Saudi Arabia, Venezuela, and six European Union member countries.

Thirty-five percent of participants were Behavior Analyst Certification Board® certified (23% Board Certified Behavior Analysts®, 11% Doctoral-level Board Certified Behavior Analysts®, and 2% Board Certified Assistant Behavior Analysts®). Nineteen percent were working toward their certifications, and 45% did not hold any board certifications. Most participants holding certifications earned their credentials between one and three years prior to the survey.

Participants were asked about their experiences with single-subject designs. Almost 52% had been involved in professional work judging single-subject design for five years or more (27% over 10 years, and 1 participant over 30 years). Ninety-five percent had been using single-

case designs in a professional capacity for at least one year. Professional positions held by participants included professors (19%), students (19%), case managers (11%, defined as being responsible for client treatment, including direct treatment and supervision of direct care or line staff), consultants (11%, defined as providing indirect treatment via a consultee such as a teacher or parent), school psychologists (10%), pre- or post-doctoral interns (7%) and researchers (4%). Sixty-three percent of participants used single-case designs professionally in school settings, 54% in research or experimental studies, 27% in home settings, 21% in clinical settings, 20% in center-based settings, 12% in residential settings, and 3% in industrial or business settings.

Sixty-five percent of respondents were affiliated with a college or university, including Applied Behavior Analysis programs (20%), school psychology programs (17%), behavior analysis or experimental behavior analysis programs (6%), education or special education (5%), child or adult clinical programs (5%), and 3% each with medical or psychiatric programs.

Most participants were not journal board members (80%). Participants who did serve as journal board members currently or in the past (13% and 7%, respectively), served on the boards of the Journal of the Experimental Analysis of Behavior (33%), the Journal of Applied Behavior Analysis (23%) , the Behavior Analyst Today (19%), and the Analysis of Verbal Behavior (10%).

Visual Analysis Ratings

Average Ratings. Each of the 16 sets of graphs had at least 10 raters (62.5%), with an average of 10.5 raters and a maximum of 12 (12.5%). The most frequent Likert scale rating for a graph across all sets and visual analysis factors (level, trend, level and trend, and all aspects) was 6; the average rating was 3.96. The lowest average rating for any graph was 1.45 and the highest was 5.90. When splitting the visual analysis ratings by the midpoint of the Likert scale (as with

the analysis for contingency probability tables), participants rated their certainty change did or did not occur; based on level, they were at least "slightly certain" a behavior change did occur in 65% of the graphs. For trend, participants rated their certainty this way for 61.25% of graphs and for level and trend and all aspects, 63.75%.

When the average ratings were split into four categories for further analysis as with the chi square tests (Table 3), for all visual analysis factors, the smallest percentage of graphs had average ratings between 1 and 1.494 (1.25% for each visual analysis factor). The visual analysis factor trend had more graphs given an average rating between 3.495 and 4.494 than any other rating (31.25%). The remaining three factors each had the most graphs given average ratings between 4.495 and 5.494, with 32.5% for level and 30% each for level and trend and all aspects (Table 3).

Rating Ranges. The range of ratings given to each graph was determined as one way to examine rater consistency (Table 4). Across all the graphs and visual analysis factors, 10% of

Table 4.

Percentage of Graphs within Each Rating Range

Rating Range	Level	Trend	Level and Trend	All Aspects
1	10	2.5	7.5	10
2	17.5	15	18.75	17.5
3	13.75	16.25	13.75	12.5
4	33.75	35	31.25	31.25
5	25	31.25	28.75	28.75

Note: Numbers represent the percentage of graphs within each range. $N = 80$.

graphs had ratings that ranged by just one Likert scale value. The majority (58.75%) had ratings

representing either four or five values. The same pattern held for the four individual visual analysis factors, with the smallest percentage of graphs in each factor having rating ranges of one Likert scale value. The majority of graphs for each factor were rated across four or five scale values, with 58.75% for level, 66.25% for trend, and 60% for level and trend and all aspects.

Agreement. Intraclass Correlation Coefficients (ICCs) were used to judge the consistency of participant ratings for each set of graphs. ICCs range from .0 to 1.0, with 1.0 representing perfect agreement. Conventions for interpreting ICC's are similar to Cohen's Kappa (Garson, 2011), with .40-.59 representing some inter-rater agreement, .60-.79 representing moderate agreement, and .80 to 1.0 representing high agreement. Because each set of five graphs had an independent group of raters drawn from a larger pool (i.e., the ten to twelve participants randomly selected for a particular set), a two-way mixed model ICC was used. Absolute agreement was also specified to test if the raters used the same absolute score (as opposed to being judged consistent if ratings were "relatively" similar as with a Consistency ICC; Garson, 2011). The average measures coefficient, the reliability of the mean of all raters, was chosen because it can be a more reliable analysis than a single measures ICC and because visual analysis ratings were going to be averaged across all raters in the comparison analyses between visual analysis and HLM (Garson, 2011; Romberg, 2009). Agreement for the 16 sets across all of the visual analysis factors ranged from .56 to .97, with an average of .81. Fifty percent of sets had coefficients greater than or equal to .80. For level, the average ICC across all sets was .83, for trend .74, for level and trend, .84 and all aspects, .83. More detailed information is presented in Table 5.

Hierarchical Linear Modeling

HLM analyses were found to have fewer limitations than previously thought. For the

Table 5.

Average Measures Intraclass Correlation Coefficients by Visual Analysis Factor

Set Coefficients	All Factors	Level	Trend	Level and Trend	All Aspects
Average	.81	.83	.74	.84	.83
Minimum	.56	.29	.53	.67	.55
Maximum	.97	.98	.97	.98	.98
ICC ≥ .80	50%	70%	50%	63%	63%

Note: N = 16 sets.

HLM2 models, all graphs except AB graphs were accommodated, for a final sample size of 68. For the HMLM models, all graphs except two were accommodated, for a final sample size of 78. After a visual inspection of the two graphs, they were found to have little to no variability in their datasets, which is a potential reason for their lack of accommodation by HMLM.

When modeling graphs using HLM2, 57% of the graphs were modeled using the Poisson distribution and 65% were analyzed using the single baseline model. When modeling the graphs using HMLM, 78% used the Homogeneous error term model, and 72% were modeled using the single baseline equations.

For graphs analyzed successfully using both models, $n = 66$, under the HLM2 model, 65% of graphs showed a significant effect of condition on level ($p < 0.05$) as compared to 71% under the HMLM model. For trend, 32% of graphs had p-values less than 0.05 with HLM2 compared to 30% of graphs with HMLM. The simultaneous level and trend hypothesis test indicated 62% and 76%, of graphs had $p < 0.05$ for the effect of condition, respectively, for HLM2 and HMLM.

Comparison Analyses

Contingency Probability Tables. Contingency probability tables (Tables 6-8) show the

Table 6.

Contingency Probability Table for Comparison 1

Statistic	HLM2		HMLM	
Overall Accuracy	.80	(.68-.89)	.86	(.75-.92)
Positive Likelihood Ratio	2.83	(1.59-5.25)	3.55	(2.36-14.71)
Negative Likelihood Ratio	.20	(.09-.45)	.11	(.04-.27)
Sensitivity	.86	(.88-.93)	.92	(.84-.97)
Specificity	.70	(.52-.82)	.74	(.59-.83)

Note: The 95% Confidence Interval is designated by parentheses. HLM2 n = 68, HMLM n = 78.

Table 7.

Contingency Probability Table for Comparison 2

Statistic	HLM2		HMLM	
Overall Accuracy	.52	(.39-.61)	.58	(.47-.64)
Positive Likelihood Ratio	1.24	(0.79-1.72)	1.58	(1.05-1.99)
Negative Likelihood Ratio	0.67	(0.26-1.44)	0.38	(0.12-0.93)
Sensitivity	.71	(.52-.87)	.82	(.62-.94)
Specificity	.43	(.34-.50)	.48	(.41-.53)

Note: The 95% Confidence Interval is designated by parentheses. HLM2 n = 68, HMLM n = 78.

results of Comparisons 1-4. Comparisons 3 and 4 were combined as the classification accuracy of HLM2 and HMLM was the same when using this dichotomy. In general, both HLM models were most accurate when modeling the effect of condition on level (Overall Accuracy = .80 for

Table 8.

Contingency Probability Table for Comparisons 3 and 4

Statistic	HLM2		HMLM	
Overall Accuracy	.66	(.54-.77)	.80	(.70-.87)
Positive Likelihood Ratio	1.57	(1.02-2.49)	4.42	(1.97-13.13)
Negative Likelihood Ratio	0.47	(0.23-0.97)	0.24	(0.16-0.43)
Sensitivity	.76	(.65-.85)	.80	(.73-.85)
Specificity	.52	(.36-.66)	.82	(.63-.94)

Note: The 95% Confidence Interval is designated by parentheses. HLM2 n = 68, HMLM n = 78.

HLM2 and .86 for HMLM) and least accurate when judging the effects of condition on trend (Overall Accuracy = .52 and .58). The HMLM model consistently showed higher levels of Overall Accuracy than the HLM2 model across all of the comparisons.

When comparing the PLRs and NLRs, HLM2 was 2.83 times more likely to classify a graph as significant based on level (Comparison 1) when it was also rated "significant" by visual analysis than it was to classify a "non-significant" graph as significant. HLM2 was .20 times less likely to classify a graph as non-significant when it was rated "significant" using visual analysis than a graph rated "non-significant" (Table 6). These likelihood ratios follow the basic pattern desired for an accurate test. HMLM also showed the same desirable pattern when classifying graphs based on trend, and even more so when classifying graphs in Comparisons 3 and 4 (PLR = 4.42, NLR = 0.24, Table 8). Across all visual analysis factors, HMLM showed higher PLRs and lower NLRs than HML2.

Sensitivity also showed the same pattern as Overall Accuracy and the likelihood ratios, with higher numbers of true positives in Comparison 1 for both HLM2 and HMLM (Sensitivity

= .86 and .92, respectively, Table 6); HMLM showed higher levels than HLM2 across all four comparisons. In terms of specificity, the HMLM model showed the highest accuracy when classifying graphs based on level and trend (.82; Table 8). Specificity decreased for both models when classifying graphs based on trend (.42 and .48 for HLM2 and HMLM, respectively), and remained low in the HLM2 model when classifying graphs based on level and trend (.43).

Chi Square Tests. Chi square tests were conducted for all four comparisons. Comparisons 3 and 4 could not be combined as with the contingency probability tables, as the higher level of detail in the chi square tests (four visual analysis categories instead of two as above) resulted in graphs being classified slightly differently between the two comparisons. Results from the Pearson tests were significant for all four comparisons and both HLM models ($p < 0.05$ and lower, Tables 9-12), except in Comparisons 2 and 3 with the HLM2 model. These

Table 9.

Ordinal Visual Analysis Level Ratings versus HLM Level Dichotomies

χ^2 Test	HLM2		HMLM	
Pearson χ^2	21.41	(0.001)*	39.06	(0.001)*
Linear χ^2	15.73	(0.001)*	34.89	(0.001)*
χ^2 Deviation	5.68	(0.05)	4.196	(0.05)
V	.57		.71	

Note: *p* values are indicated by parenthesis. For all Pearson χ^2 tests, $df = 3$, for Linear χ^2 tests, $df = 1$, for Deviance tests, $df = 2$, for Cramer's *V*, $df^* = 1$. HLM2 $n = 68$, HMLM $n = 78$.
* indicates a significant *p* value.

comparisons were the classification of graphs based on trend and level and trend, and as smaller effects, could have simply hindered by low power from the smaller HLM2 sample size [Pearson χ^2 (3, n=68) =4.15, $p = 0.25$; Pearson χ^2 (3, n=68) =6.95, $p = 0.07$). Howell (2010) suggested

Table 10.

Ordinal Visual Analysis Trend Ratings versus HLM Trend Dichotomies

χ^2 Test	HLM2		HMLM	
Pearson χ^2	4.15	(0.25)	9.01	(0.029)*
Linear χ^2	2.99	(0.08)	8.77	(0.003)*
χ^2 Deviation	1.16	(0.05)	0.24	(0.05)
V	.25		.34	

Note: *p* values are indicated by parenthesis. For all Pearson χ^2 tests, *df* = 3, for Linear χ^2 tests, *df* = 1, for Deviance tests, *df* = 2, for Cramer's *V*, *df** = 1. HLM2 *n* = 68, HMLM *n* = 78.
* indicates a significant *p* value.

Table 11.

Ordinal Visual Analysis Level and Trend Ratings versus HLM Level and Trend Dichotomies

χ^2 Test	HLM2		HMLM	
Pearson χ^2	6.95	(0.07)	29.83	(0.001)*
Linear χ^2	6.50	(0.011)*	29.04	(0.001)*
χ^2 Deviation	0.45	(0.05)	0.79	(0.05)
V	.32		.62	

Note: *p* values are indicated by parenthesis. For all Pearson χ^2 tests, *df* = 3, for Linear χ^2 tests, *df* = 1, for Deviance tests, *df* = 2, for Cramer's *V*, *df** = 1. HLM2 *n* = 68, HMLM *n* = 78.
* indicates a significant *p* value.

that in instances of smaller sample sizes, the Linear-by-Linear chi square may be used, which returned a significant *p* value in Comparison 3 (Linear χ^2 (1, *n*=68) =6.5, *p* = 0.011).

In general, the results indicated the visual analysis ratings and HLM significance classifications were associated in some way. The Linear-by-Linear chi squares indicated the relationships were linear, and each Deviation from Linear chi square statistic confirmed this

Table 12.

Ordinal Visual Analysis All Aspects Ratings versus HLM Level and Trend Dichotomies

χ^2 Test	HLM2		HMLM	
Pearson χ^2	9.11	(0.028)*	35.16	(0.001)*
Linear χ^2	8.80	(0.003)*	33.63	(0.001)*
χ^2 Deviation	0.31	(0.05)	1.53	(0.05)
V	.37		.67	

Note: *p* values are indicated by parenthesis. For all Pearson χ^2 tests, *df* = 3, for Linear χ^2 tests, *df* = 1, for Deviance tests, *df* = 2, for Cramer's *V*, *df** = 1. HLM2 *n* = 68, HMLM *n* = 78.
* indicates a significant *p* value.

relationship. Overall, as raters gave higher Likert scale values, the probability increased that HLM would result in a significant *p*-value.

Although the relationship between HLM and visual analysis was linear in every comparison, the degree of this association differed by the factor and model analyzed. In general, both HLM models showed higher degrees of association when compared to visual analysis ratings of level (*V* = .57 for HLM2 and *V* = .71 for HMLM). The association of the variables dropped considerably for HLM2 for the other visual analysis factors (ranging from .25 to .37), but only decreased significantly for HMLM when comparing classifications based on trend (*V* = .34). HMLM consistently showed higher degrees of association than HLM2, as HMLM had three "high" and one "moderate" associations, and HLM2 had one "high," two "moderate," and one "low" associations, as based on the interpretation method described previously from Gravetter & Wallnau (2007).

Visual depictions of HML2 and HMLM classification accuracy for the chi square tests can be found in Figures 6-13. The figures depict the percentage of graphs analyzed as significant

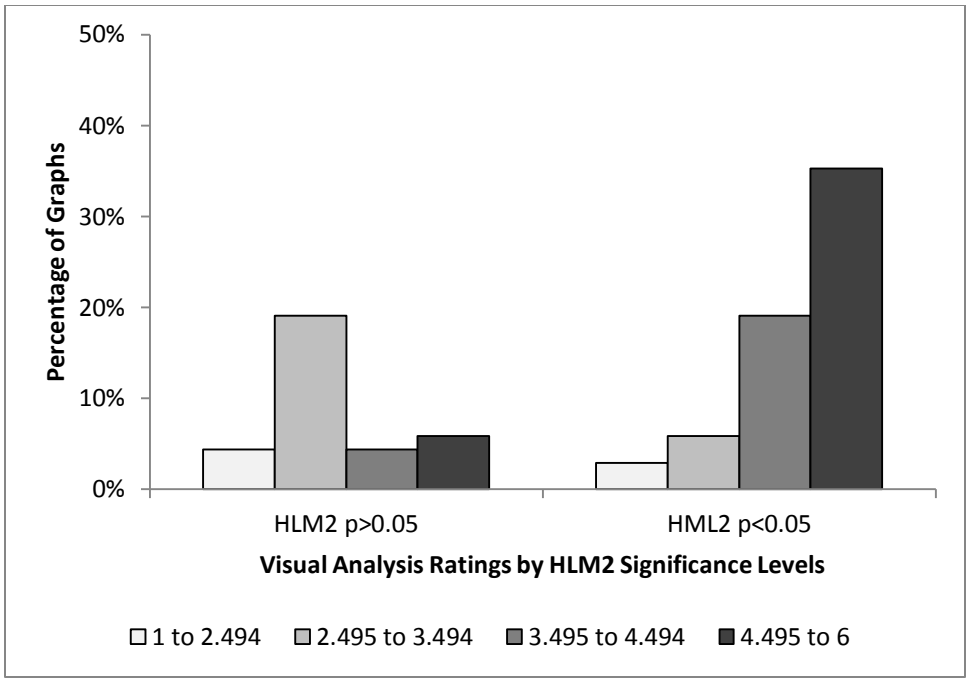

Figure 6. Comparison 1 for HLM2. The figure represents the distribution of visual analysis ratings as a function of HLM2 hypothesis test results.

or non-significant by HLM as a function of their visual analysis classification. An ideal graph would show all graphs rated between 1 and 3.494 a classified as non-significant by HLM and all graphs between 3.495 and 6 classified as significant by HLM.

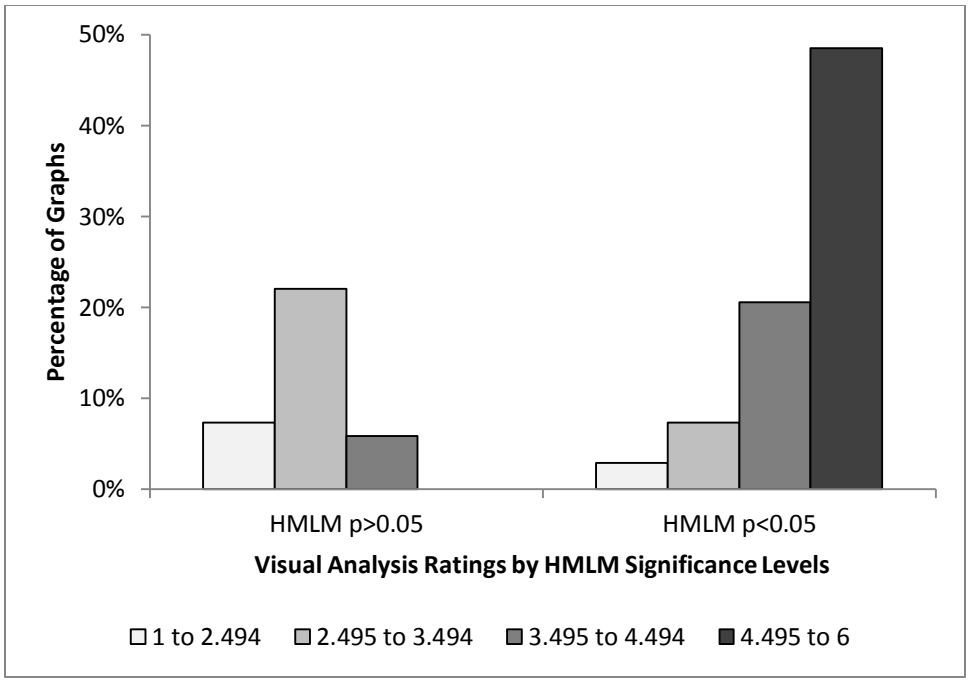

Figure 7. Comparison 1 for HMLM. The figure represents the distribution of visual analysis ratings as a function of HMLM hypothesis test results.

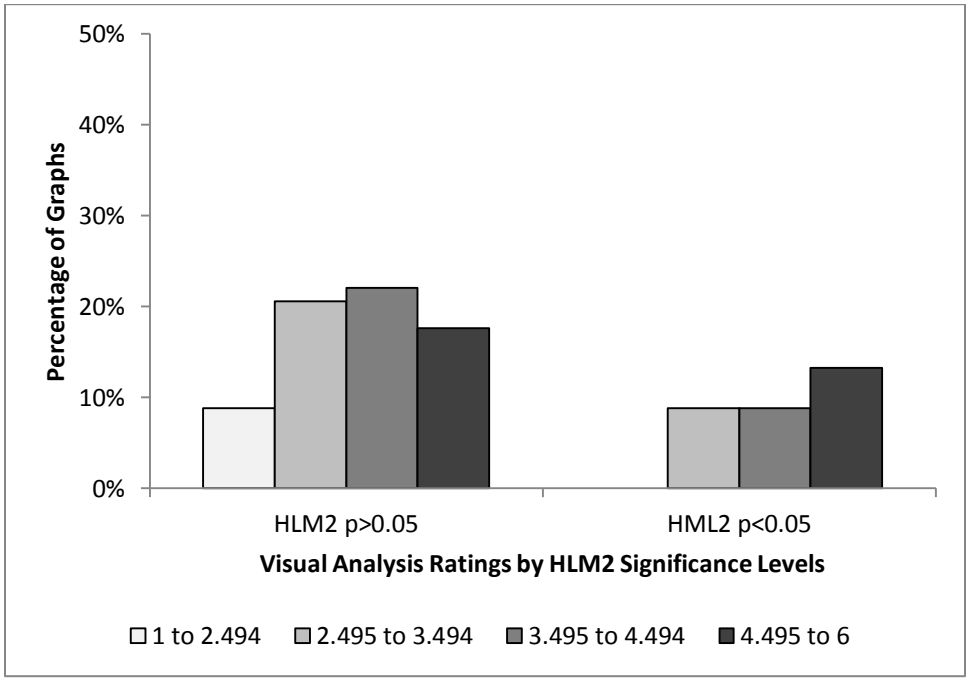

Figure 8. Comparison 2 for HLM2. The figure represents the distribution of visual analysis ratings as a function of HLM2 hypothesis test results.

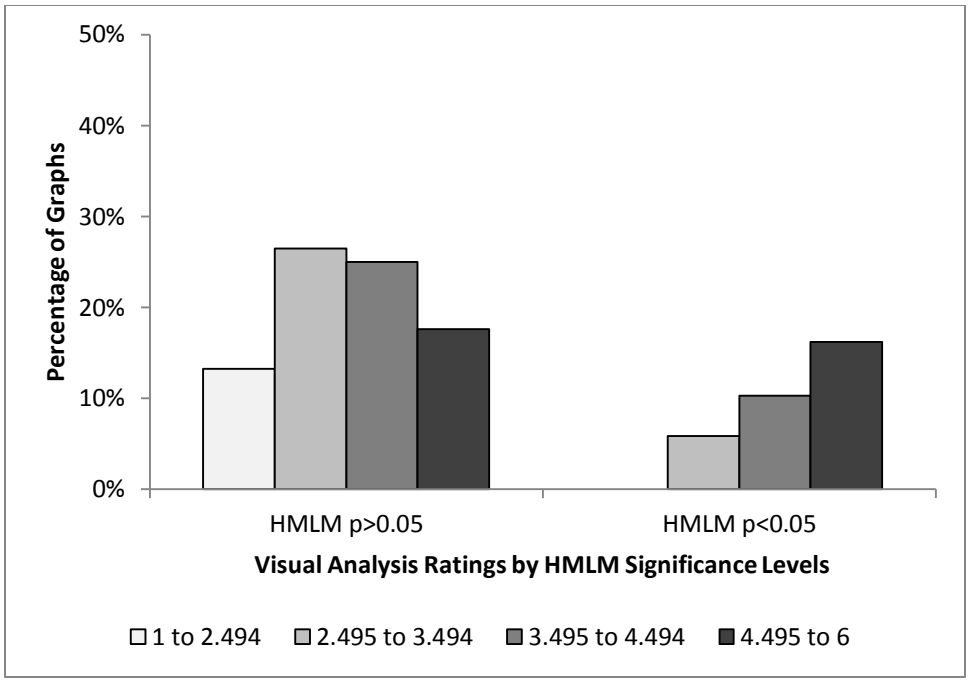

Figure 9. Comparison 2 for HMLM. The figure represents the distribution of visual analysis ratings as a function of HMLM hypothesis test results.

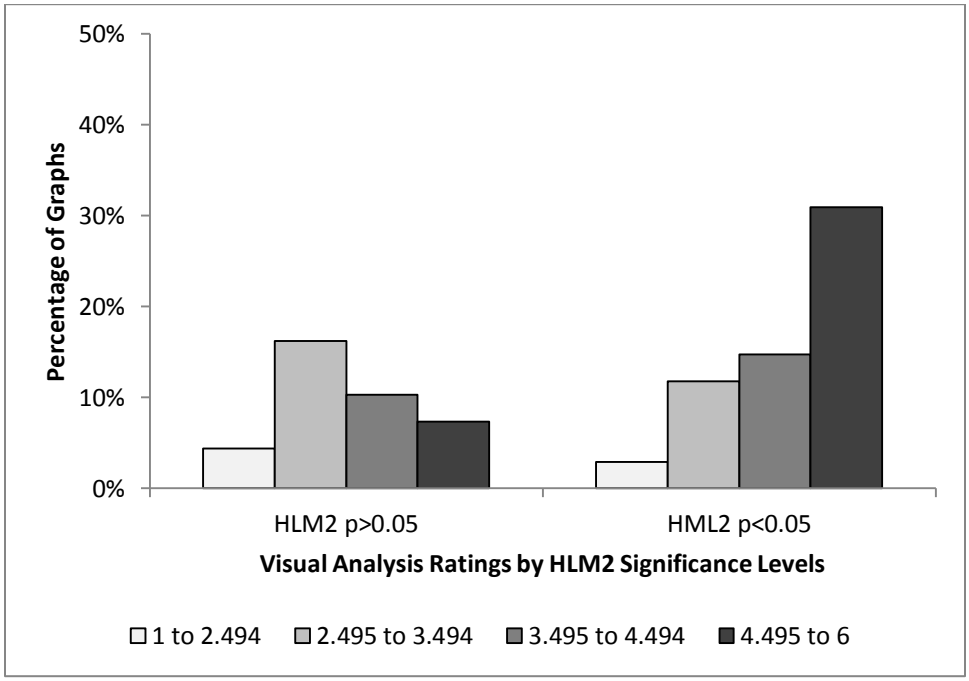

Figure 10. Comparison 3 for HLM2. The figure represents the distribution of visual analysis ratings as a function of HLM2 hypothesis test results.

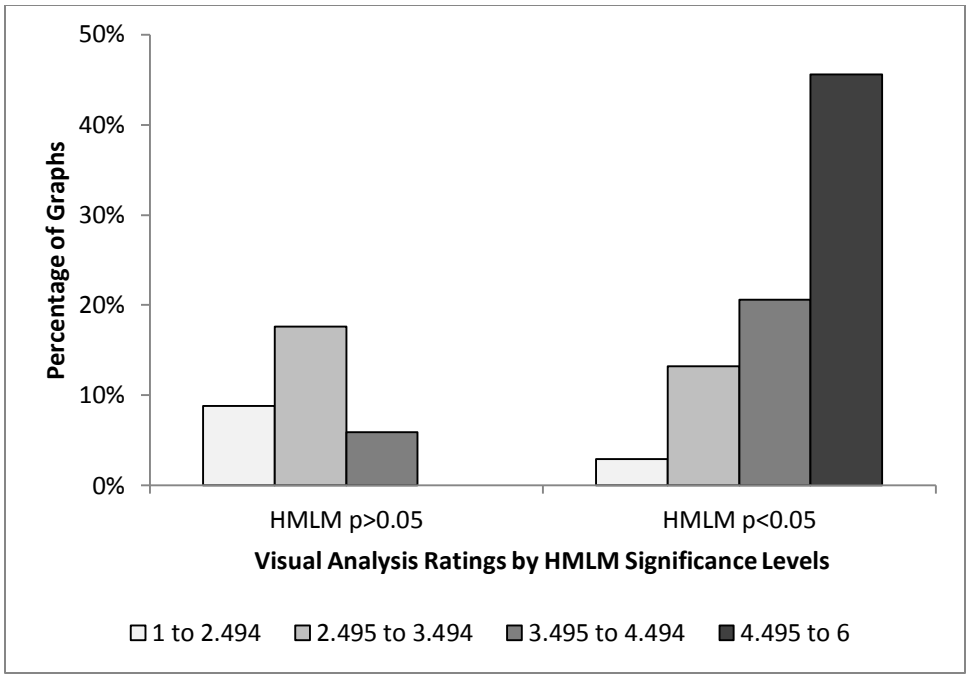

Figure 11. Comparison 3 for HMLM. The figure represents the distribution of visual analysis ratings as a function of HMLM hypothesis test results.

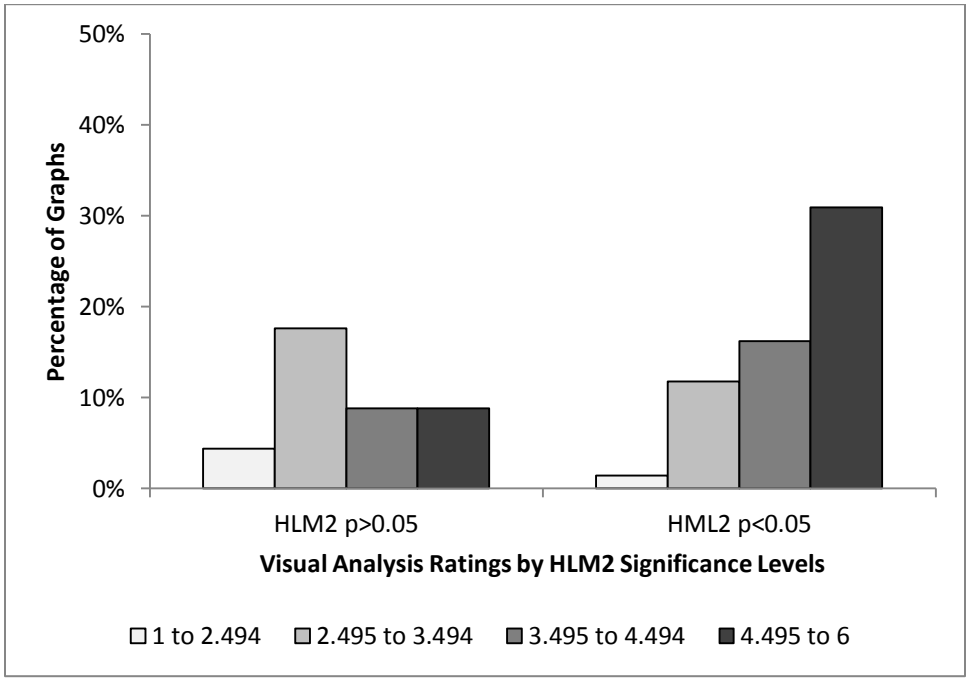

Figure 12. Comparison 4 for HLM2. The figure represents the distribution of visual analysis ratings as a function of HLM2 hypothesis test results.

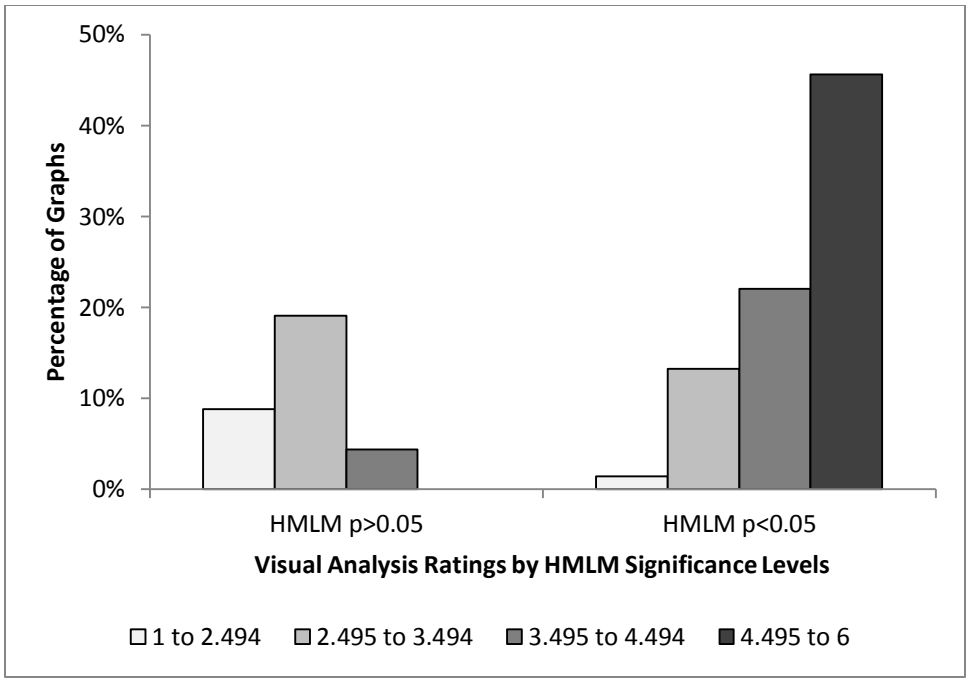

Figure 13. Comparison 4 for HMLM. The figure represents the distribution of visual analysis ratings as a function of HMLM hypothesis test results.

DISCUSSION

HLM was most accurate when determining the effect of condition on level, as both the HLM2 and HMLM models displayed acceptable levels of Overall Accuracy, Sensitivity, and Specificity. In addition, both models had strong positive linear relationships with visual analysis in this comparison. These results are similar to the previous study, which also found HLM to be most accurate when classifying changes based on level (Godbold, 2008). When comparing the HLM2 and HMLM models across all visual analysis factors, HMLM appears better suited to analyzing single-subject data. In each comparison, HMLM had higher levels of accuracy than HLM2 and stronger positive linear relationships. This difference may be due to the longitudinal nature of the HMLM model, which is better suited to analyzing different measurement occasions than HLM2. Strong conclusions about differences cannot be drawn, however, as there is no direct statistical comparison between the two models. Additionally, HLM2 had a smaller sample size than HMLM, which could have affected its power in Comparison 2.

HLM was least accurate in analyzing condition effects based on trend, which is again similar to the previous study. This decrease in accuracy was most likely related to incorrectly identifying graphs as non-significant. HLM identified more graphs as non-significant than significant based on trend (an average of 29% of graphs were significant for trend for HLM2 and HMLM), which is opposite of how it classified results in the other comparisons (66.5% significant for level, 64.5% for level and trend). In contrast, the percentage of graphs rated "significant" by participants remained constant across all visual analysis factors: 61.25% for trend, 65% for level, and 63.75% for level and trend and all aspects.

This decrease could lead to HLM being described as inherently more conservative when judging trend changes than visual analysis, or this difference could have been more connected to

how raters judged trend changes. Several raters commented they either had difficulty judging the effects of trend without also judging level or they could not provide an accurate rating because they did not factor trend changes into their ratings at all. Participants could have justifiably decreased their certainty level and subsequent rating of trend in these instances, but this choice may not have been obvious. In fact, 95% of trend ratings were within one Likert scale value of the corresponding level rating, indicating participants judged level and trend similarly, and giving more support to the idea that the decrease in HLM accuracy may be more a function of participant difficulty with how to rate trend than a problem with the statistic.

Differences in individual model accuracy also affect any conclusions about overall HLM accuracy for trend. If HLM were considerably more conservative when judging trend, it would also likely be more conservative when judging the effects of both level and trend simultaneously(assuming any changes in level would not be so large as to overshadow a lack of change in trend); however, classification results based on both level and trend were mixed. HMLM classified changes in both level and trend with nearly the same accuracy as it classified changes in level; chi square tests indicated a strong linear relationship between HLM and visual analysis for this comparison. The HLM2 model showed a decrease in accuracy for the level and trend comparison, though not to the same degree as the trend comparison. The contrasting results means a conclusion cannot be made as to which factor had the largest effect: problems with participant ratings, a potential "conservativeness" of HLM in regards to trend, or differences in how the models analyzed data.

However, HLM showed much higher accuracy compared to visual analysis than previously tested statistics and retained this accuracy for both significant and non-significant effects. In addition to supporting the viability of HLM as a statistical aid, the current study was

able to overcome many of the limitations of the previous. Raters were qualified to judge single-subject graphs as much as could be determined by their self-reported certifications, degrees, professional responsibilities, and professional experience using single-subject designs. As compared to the previous study, however, they were from a wider range of disciplines and professions, and were perhaps more reflective of the larger community using single-case designs. Any statistical aid should be useful to and usable by this larger community, and the accuracy of HLM as compared to ratings from a more diverse group supports it application across a variety of fields and research questions.

Another support to HLM's wide applicability is the various designs it was able to model successfully. In addition to the designs modeled in the previous study, AB designs and functional analyses were accommodated by the HMLM model, and both models were able to analyze multiple baseline graphs with the addition of the $Baseline_i$ predictor.

The ceiling effects found with the previous study were alleviated as well. The previous study used a 5-point Likert scale and average ratings were divided into categories similar to the ones used currently. In the previous study, however, no graphs received an average rating falling within the highest category. In the current study, all categories were represented.

Current study results also indicated an increase in rater agreement. Pearson *r* calculations of agreement from the previous study were .46 for level, .46 for trend, and .43 for both level and trend (all $p < 0.05$). Current average levels of agreement using ICCs were .83 for level, .78 for trend, and .84 for level and trend. Although the metrics are different, both are judged on a .0 to 1.0 scale, with scores closer to 1.0 indicating agreements that are more consistent; higher levels of agreement appear to have been reached in the current study. In addition to increased rater agreement, the range of Likert scale values used to rate each graph decreased. In the previous

study, 89% of graphs were given Likert scale values across the entire 5-point scale; in the current study, 28% percent of graphs had ratings across the entire 6-point scale; 33% were rated using a range of five Likert values. These results indicate that the current visual analysis ratings can be considered valid representations of accurate visual analysis, and that higher agreement can be reached by providing more structured comparisons and questions; additional training or overly rigid response requirements is not necessarily required.

Two components were added to the current study design to answers questions from the previous study. The first was any impact on HLM accuracy due to the requirement in that both level and trend had to be significant separately to be considered significant together. This requirement limited the comparison of HLM analyses to visual analysis ratings, as raters considered both level and trend simultaneously whereas HLM did not. The current study included both the results of HLM when considering level and trend simultaneously as well as when both had separate, significant *p* values. As discussed previously, accuracy was mixed when HLM considered both simultaneously (HLM2 Overall Accuracy = .66; HMLM Overall Accuracy = .80), though both models showed positive linear relationships with visual analysis (V = .32 for HLM2, a moderate coefficient, and V = .62 for HMLM, a high coefficient). When using the criterion that both level and trend had to have significant *p* values, however, Overall Accuracy decreased significantly (average Overall Accuracy = .57 for HLM2 and .56 for HMLM for all visual analysis factors). The linear relationship between visual analysis and HLM remained significant, but the association between the variables dropped from high for HMLM (V = .62) to moderate (V = .36); the association between variables for HLM2 remained at a moderate level. These results indicate that analyzing the effects of level and trend simultaneously is an improvement over the methods of the previous study and is more analogous

to how visual analysts rate effects using both factors

The second component added was a survey question asking participants to rate change based on all aspects of the graph. This question was intended to determine if the questions about level and trend encompassed all of the graph characteristics participants used in making their ratings. Different ratings were expected for the all aspects question as compared to the level and trend question because participants would be able to factor in variability, a key component to visually analyzing graphs, as well as any idiosyncratic graph characteristics like presentation style. Very few differences were found in ratings between the level and trend and all aspects questions, however. When dividing the ratings into the "significant" and "non-significant" dichotomies for contingency probabilities, no differences were found in the ratings (Pearson r = 1.0, $p < 0.001$). Differences did arise when the ratings were split into the categories used for the chi square tests but were minimal, $r = .95$, $p < 0.001$, and were based on a change in classification for just two graphs. Therefore, it appears the questions about level and trend were sufficient to capture most of the factors affecting raters' decisions, at least within the context of the survey.

Limitations

The current study indicates HLM may be a useful statistical aid. This utility would lie in adding a measure of statistical significance to visual analysis decisions, but would not provide a measure of clinical significance or effect. For example, HLM could analyze data from two treatments and providence p values for each, but if both were significant, HLM would not quantitatively determine which had a larger treatment effect. Development of an effect size based on HLM modeling, however, would be appropriate given the accuracy of HLM in the current study, as it would provide users with a measure of treatment effect.

Other limitations of HLM in this study include the potential lower power for HLM2 in Comparison 2 and the inability to use the Binomial distribution when it was indicated. Therefore, results of the HLM2 analysis should be interpreted with caution. Additionally, the use of $Time_{ti}$, $Condition_i$, and $Baseline_i$ as predictors was tested using a random subset of graphs, and though the predictors were found to account for substantial amounts of variance in those graphs, the use of these predictors may not have been appropriate for every graph. Another consideration is that the models chosen were appropriate for a large number of graphs but were not used to model changing criterion designs, so the utility of HLM with this type of single-case data remains unknown.

Another limitation of HLM modeling in general is the amount of response effort required of users. Data must be extracted from graphs (if original data points are unknown) and phases and time points identified; variables must be coded in a statistical package and then imported into the HLM7S program. The HLM7S program leaves many decisions up to the user – not necessarily a detriment as it makes the program flexible – but a feature that requires substantial knowledge before using the program and interpreting the output. The information available about the program and HLM in general has increased over the last few years, but is still left to the user to seek out. General users of HLM should anticipate a learning curve for determining appropriate data analysis techniques.

The survey also had limitations. The effect of using a Likert scale with no midpoint may have skewed ratings; however, results are mixed. The results of some studies show ratings are positively skewed without a midpoint (e.g., Garland, 1991) and the results of other studies show they are negatively skewed (e.g., Dawes, 2001). Any influence on the current survey is unknown. Additionally, although the graphs within each graph set were randomly ordered, they

were not randomly ordered for each participant. Therefore, some order effects may be present in the rating data.

Future Directions

Directions for future research include more advanced HLM modeling of single-case designs. The current study used two conditions targeted for comparison even when graphs had multiple condition types. Using a more advanced model may be more analogous to visual analysis ratings, because even though raters were directed toward a particular phase comparison, they were still exposed to other phases in the graph. More advanced HLM models could account for multiple phase types and retain the ability to pinpoint individual condition effects.

More research should also be conducted into the accuracy of HLM2 models versus HMLM models. The current results indicated HMLM may be the more accurate model but there was no direct statistical comparison of model accuracy. Another potential line of research is the use of 2-level versus 3-level models for multiple baseline graphs. The 2-level model used here is based on the idea that baselines and conditions are "equal level" predictors, but nesting baseline in condition, or vice versa, may provide a more accurate representation of multiple-baseline single-subject designs.

Additionally, the diversity of the rater sample is conducive to comparing HLM accuracy against visual analysis classifications by different groups of raters, such as raters who are board certified versus those with without certification, or raters who have over five or ten years of professional single-subject design experience versus those with less. If HLM is found to be accurate even when raters have lower levels of experience, it would further demonstrate its wide applicability as a statistical aid.

Implications

Results from the current study demonstrated HLM might be a viable statistical aid to visual analysis. In general, HLM correctly classified both positive and negative results at acceptable levels as shown by contingency probabilities and these results were established for a wide variety of single-case designs. Chi square tests accounting for the ordinal nature of the ratings confirmed a positive linear relationship between visual analysis and HLM, and Cramer's *V* coefficients demonstrated the relationship between HLM and visual analysis ratings was strong.

REFERENCES

Agresti, A. (1989). Tutorial on modeling ordered categorical response data. *Psychological Bulletin, 105*, 290-301.

Agresti, A. (1996). *An Introduction to Categorical Data Analysis*. New York, NY: John Wiley & Sons, Inc.

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.

Beardsell, I., Bell, S., Rumbold, H., & Robinson, S. (2009). *MCEM Part A: MCQs*. United Kingdom: Royal Society of Medicine Books.

Bormann, I. (2010). DigitizeIt© (Version 1.5.8b). Retrieved from http://www.digitizeit.de/

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.

Connor-Linton, J. (Producer). (2003). Chi square tutorial. [PDF] Retrieved from garnetthenley.com/ChiSquareLec.pdf

Crewson, P. (2012). Applied Statistics Handbook *Coefficients for Measuring Association* Retrieved from http://www.acastat.com/Statbook/chisqassoc.htm

Dawes, J. (2001). The impact of mentioning a scale mid-point in administering a customer satisfaction questionnaire via telephone. *Australasian Journal of Market Research, 9*, 11-18.

Decoster, J. (Producer). (2002). Hierarchical Linear Modeling (HLM). [Powerpoint Presentation]

Deeks, J. (Producer). (2006). Statistical methods for analysis of diagnostic accuracy studies. [Powerpoint Presentation] Retrieved from www.cebm.net/index.aspx?o=1109

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.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175-191.

Fisch, G. S. (1998). Visual inspection of the data revisited: Do the eyes still have it? *The Behavior Analyst, 21*, 111-123.

Garland, R. (1991). The mid-point on a rating scale: Is it desirable? *Marketing Bulletin, 2*, 66-70.

Garson, G. D. (2011). Reliability Analysis, from http://faculty.chass.ncsu.edu/garson/PA765/reliab.htm#rater

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.

Godbold, E. S. (2008). *Hierarchical linear modeling against the "gold standard" of visual analysis in single-subject design.* M.A., Louisiana State University, Louisiana. Retrieved from http://etd.lsu.edu/cgi-bin/ETD-browse/browse/ Available from Electronic Thesis and Dissertation Library

Gravetter, F. J., & Wallnau, L. B. (2007). *Statistics for the behavioral sciences.* United States: Thomson Wadsworth.

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.

Higgins, J., & Green, S. (Eds.). (2011). *Cochrane handbook for systematic reviews of interventions* (March 2011 ed.): The Cochrane Colloboration.

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children, 71*, 165-179.

Howell, D. C. (2010). *Statistical Methods for Psychology* (Seventh ed.). Belmont, CA: Cengage Wadsworth.

International, S. S. (Producer). (2012a). Centering. [PDF] Retrieved from http://www.ssicentral.com/hlm/help7/faq/FAQ_Centering.pdf

International, S. S. (Producer). (2012b). Multicollinearity. [PDF] Retrieved from www.ssicentral.com/hlm/help6/faq/Multicollinearity.pdf

Introduction to multilevel modeling using HLM. (2012). *Statistical Computing Seminar*, from http://www.ats.ucla.edu/stat/hlm/seminars/hlm6/mlm_hlm6_seminar.htm

Johnston, J. M., & Pennypacker, H. S. (1993). *Strategies and tactics of behavioral research* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

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.

Kirkman, T. W. (1996). Statistics to use: Contingency tables with sparsely populate cells, from http://www.physics.csbsju.edu/stats/contingency.problem.html

Knapp, T. J. (1983). Behavior analysts' visual appraisal of behavior change in graphic display. *Behavioral Assessment, 5*, 155-164.

Kyse, E. N., Rindskopf, D. M., & Shadish, W. R. (in submission). Analyzing data from single-case designs using multilevel models: A primer. *Psychological Methods*.

Manolov, R., Solanas, A., & Leiva, D. (2010). Comparing "visual" effect size indices for single-case designs. *Methodology, 6*, 49-58.

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.

Miller, B. (Producer). (2003, March 31, 2012). Single Subject Research Design. [Powerpoint Presentation] Retrieved from http://www.vchri.ca/i/pdf/SingleSubjectResearch.pdf

Nagler, E., Rindskopf, D., & Shadish, W. (2008). Analyzing data from small *n* designs using multilevel models: A procedural handbook (pp. 107). Retrieved from http://faculty.ucmerced.edu/wshadish/Nagler%20Rindskopf%20Shadish%20SCD%20Manual%2011%2020%202008%20b.pdf

Normand, M. P., & Bailey, J. S. (2006). The effects of celeration lines on visual data analysis. *Behavior Modification, 30*, 295-314.

Olive, M. L., & Smith, B. W. (2005). Effect size calculations and single subject designs. *Educational Psychology, 25*, 313-324.

Osborne, J. W. (2000). Advantages of hierarchical linear modeling. *Practical Assessment, Research, & Evaluation, 7*(1).

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.

Parker, R. I., & Hagan-Burke, S. (2007). Median-based overlap analysis for single case data: A second study. *Behavior Modification, 31*, 919-936.

Parker, R. I., Vannest, K. J., & Brown, L. (2009). The Improvement Rate Difference for single-case research. *Exceptional Children, 75*, 135-150.

Parsonson, B. S., & Baer, D. M. (1992). The visual analysis of data, and current research into the stimuli controlling it. 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.

Ployhart, R. E., & Vandenber, R. J. (2010). Longitudinal research: The theory, design, and analysis of change. *Journal of Management, 36*, 94-120.

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., & Congdon, R. (2011). HLM for Windows (Version 7.0): Scientific Software International. Retrieved from http://www.ssicentral.com/hlm/student.html

Romberg, A. (Producer). (2009). Intraclass correlation coefficients: Reliability and more... [Powerpoint Presentation] Retrieved from http://dionysus.psych.wisc.edu/coursewebsites/Psy710/Discussion/ICC/ICC.ppt

Sarkisian, N. (Producer). (2007a). HLM Model Building Strategies. [Word Document] Retrieved from www.sarkisian.net/sc708/model_strategies.pdf

Sarkisian, N. (Producer). (2007b). Two-level HLM models. [Word Document] Retrieved from http://www.sarkisian.net/sc705/september13.doc

Sidman, M. (1960). *Tactics of scientific research*. Oxford, England: Basic Books.

Sparks, S. D. (2012). Ed. Dept. promotes single-case design research for special ed. Retrieved from http://blogs.edweek.org/edweek/inside-school-research/2012/01/ed_dept_promotes_single-case_d.html

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.

Swaminathan, H., Horner, R. H., Rogers, H. J., & Sugai, G. (2012). *Effect size measure and analysis of single subject designs*. Paper presented at the Society for Research on Educational Effectiveness, Washington, DC. https://www.sree.org/conferences/2012s/program/downloads/abstracts/541_4.pdf

Uekawa, K. (Producer). (2012). Doing HLM by SAS(r) PROC MIXED. [Word Documnt] Retrieved from www.estat.us/sas/PROCMIXED.doc

Van den Noortgate, W., & Onghena, P. (2007). The aggregation of single-case results using hierarchical linear models. *The Behavior Analyst Today, 8*, 196-208.

Waddell, D. E., Nassar, S. L., & Gustafson, S. A. (2011). Single-case design in psychophysiological research: Part II: Statistical analytic approaches. *Journal of Neuropathy, 15*, 160-169.

Wendt, O. (2009). *Calculating effect sizes for single-subject experimental desings: An overview and comparison*. Paper presented at the Campbell Collaboration Colloguium, Olso, Norway. http://www.campbellcollaboration.org/artman2/uploads/1/Wendt_calculating_effect_sizes.pdf

Weng, L. (2004). Impact of the number of response categories and anchor labels on Coefficient Alpha and Test-Retest Reliability. *Educational and Psychological Measurement, 64*, 956-972.

Wuensch, K. L. (Producer). (2007). Inter-Rater Agreement. [Word Document] Retrieved from http://core.ecu.edu/psyc/wuenschk/docs30/InterRater.doc

VITA

Elizabeth Godbold Nelson 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 and earned her master's degree in 2008. She completed her predoctoral internship at the May Institute in Randolph, Massachusetts, and is currently working toward licensure.