Analyzing Data from Single Case Design Studies:

A Demonstration and Comparison of Methods

by

Eden Nagler Kyse

A dissertation submitted to the Graduate Faculty in Educational Psychology in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York

2010

This manuscript has been read and accepted for the
Graduate Faculty in Educational Psychology in satisfaction of the
dissertation requirement for the degree of Doctor of Philosophy.

David Rindskopf, Ph.D.

_____ / _____ / _____
Date

_____
Chair of Examining Committee

Mario Kelly, Ph.D.

_____ / _____ / _____
Date

_____
Executive Officer

David Rindskopf, Ph.D.
Keith Markus, Ph.D.
Jay Verkuilen, Ph.D.

Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK

Abstract

ANALYZING DATA FROM SINGLE CASE DESIGN STUDIES:

A DEMONSTRATION AND COMPARISON OF METHODS

by

Eden Nagler Kyse

Adviser: David Rindskopf, Ph.D.

Data from single case and small-*N* interrupted time series (ITS) design studies offer rich information, not available from group comparison designs, about the effects of an intervention on individuals. Several methods for analyzing and synthesizing these kinds of data have been proposed to date, though many are limited or flawed. A more sophisticated statistical method, using multilevel modeling techniques, overcomes many of the limitations of the earlier approaches. This claim is supported by a comparative discussion and demonstration of several methods with two reversal design datasets. Procedures and estimates are explained, interpreted, and compared. Potential solutions for accommodating technical complexities of the data are discussed.

Acknowledgements

Through their support, many people have helped me to conquer this dissertation. To those who encouraged me when I needed it, put things in perspectives when I needed it, let me blow off steam when I needed it, or otherwise just put up with it… Thank you.

I would like express my endless gratitude to the following extraordinary contributors:

To my mentor and advisor, Dr. David Rindskopf, for your unparalleled teaching and your indispensable guidance… Thank you. For helping me narrow in on a topic, for providing me with remarkable opportunities to explore it, and for your ongoing encouragement… Thank you.

To my committee members, Dr. Keith Markus and Dr. Jay Verkuilen, for your responsiveness, for your thoughtful insight, and for encouraging me to continually improve upon the utility of this dissertation… Thank you.

To my readers, Dr. Lou Primavera and Howard Everson, for your valuable time and reflections on the details of this dissertation... Thank you.

To my exceptional family and friends, all of whom may not have understood exactly what it was I was writing about, but who cheered me on even so from near and far… Thank you.

To my parents, for showing me how to work hard to accomplish your goals, for instilling in me the confidence that I could achieve anything, and for trying so hard not to ask too many times how my dissertation was coming along… Thank you.

And to my husband, Tim, for your love, patience, and encouragement, for believing that the finish line would come, and for always knowing when and how to make me laugh... Thank you.

Table of Contents

List of Tables

List of Figures

Analyzing Data from Single Case Design Studies

As the drive for accountability in education thrives, the need to document the effectiveness of educational practices grows beside it. From local school districts to the U.S. Department of Education, schools and programs are being pushed to substantiate student growth and to use evidence-based practices. The No Child Left Behind Act of 2001 included the promotion of rigorous scientific research methods to reliably assess the effectiveness of school-based programs (U.S. Department of Education, 2002). More recently, in 2007, the New York City Department of Education (NYCDOE) began an effort to measure and report on the effectiveness of each of its 1,600-plus public schools through its school progress reports and quality reviews (New York City Department of Education, 2007). And currently states are competing for at least $4 billion in federal education funds in the "Race to the Top" grant competition, which includes as a major component an initiative toward tracking individual student progress and supporting data-driven instructional planning by teachers and principals (U.S. Department of Education, 2010). The accountability trend is not limited to the domain of education. Public and private funders want to see a return on their investments. In a time of economic stress, funded programs across a variety of industries are being pressured to justify these investments to stakeholders through rigorous evaluation.

Program evaluation can utilize various research designs, each of which involves the collection and analysis of information to answer defined questions about the initiative under study. Some designs provide for stronger evidence for an intervention's effectiveness than do others (i.e., rigorous evaluation). The U.S. Department of Education's What Works Clearinghouse (WWC), under the Institute of Education Sciences (IES), has developed a list of eligible designs that meet methodological criteria for the strength of evidence that a study

provides (U.S. Department of Education, 2008). These eligible designs include randomized controlled trials (RCT), quasi-experimental designs with well-matched comparison groups, regression discontinuity designs, and single subject or single case designs.

Single case research design involves repeated documentation of individuals' behavior over time, across phases: before, during, and/or after an intervention is implemented. In this way, single case designs are also sometimes known as "short-phase interrupted time series (ITS) designs." Single case design research can include data on more than one subject. Whether it involves examination of one or more subjects, this type of research can provide for systematic and detailed analysis of individual behavior patterns not available with group design research, by providing data on individual behavior patterns rather than just single number estimates of performance. Sometimes, single case designs are more practical than group designs; at other times, they make more clinical sense. Over the past two decades, single case research design methodologies have become increasingly prevalent across a range of fields such as behavior modification, clinical psychology, neuropsychology, special education research, addiction research, econometrics and communicative disorder treatment (Kratochwill, 1992; Van den Noortgate & Onghena, 2007).

Explorations conducted via single case design can often provide relevant evidence for measuring program impacts, meeting accountability requirements for the programs they track. However, data from these designs are rarely synthesized to reveal overall effects, determinants of effects, or generalizability of results (Van den Noortgate & Onghena, 2003b). Instead, methods for evaluating single case design data often include graphically plotting behaviors across phases and then assessing results by visual inspection methods that can be unreliable and subjective (Richards, Taylor, Ramasamy, & Richards, 1999; Scruggs, Mastropierri, & Casto, 1987).

Several statistical methods, many of which are based on an underlying regression model, have been proposed for analyzing and synthesizing single case design data. Most of these, however, are limited by the fact that they do not take advantage of the depth and precision of single case design data and/or that they are not based on sound statistical principles. In general, the small sample sizes in these types of studies (i.e., small number of individuals and relatively few observations per individual) cannot support the strong conclusive claims often made by single case researchers when misapplying analytical methods intended for between-group studies. There is a demonstrated need for more cautious claims and for the use of more appropriate methods developed and/or adapted specifically for these types of data.

A newer quantitative method that utilizes multilevel modeling, or hierarchical linear modeling (HLM), avoids many of the limitations of earlier methods by design (Van den Noortgate & Onghena, 2003a, 2003b). Multilevel modeling accommodates for the nested structure of these data – observations within individuals – allowing estimation of overall effects across individuals and exploration and explanation of the heterogeneity of these effects (Raudenbush & Bryk, 1992). The multilevel modeling approach to analysis of single case design data possesses strengths in its accommodation of technical complexities and in its potential flexibility in modeling effects.

Not surprisingly, authors of each method put forward claim their own to be the most suitable approach. In comparison to the previous methods proposed, the multilevel modeling approach seems to better handle the technicalities of single case design data, allowing for stronger and more valid claims about program effectiveness. However, no comprehensive comparison has been conducted (by a third party) to assess the differential performance of these

methods in appropriately expressing the intricate data patterns available from single case research.

Increasingly, multilevel modeling is being accepted by researchers as the most effective and appropriate approach for analyzing these types of data. The aim of this research is to review the development of the multilevel modeling approach in the broader context of the other proposed methods for analyzing and synthesizing single case design data, including a comparison of the strengths and weaknesses of each technique to that of the multilevel modeling method. After a review of the literature on these proposed methods, a demonstration of a selection of competing methods will be conducted on two different datasets (one simulated, one published) to explore the expectation that the multilevel modeling approach is best. Once this has been established, a demonstration of estimation methods will also be conducted – comparing two possible approaches to estimating the random effects of the multilevel model (maximum likelihood empirical Bayes and fully Bayesian).

Background

Single case design research is experimental. It aims to document the causal relationships between independent variables (e.g., time, treatment) and dependent variables. Small-*N* ITS studies intimately follow each member of a small group of individuals over some time period; otherwise abiding by the principles of single case designs. Between- and within-subjects comparisons are used to control for major threats to external and internal validity. Subjects serve as their own controls (i.e., within-subjects comparisons), whereby behavior patterns prior to the introduction of an intervention are compared to patterns during and/or after the intervention is implemented (Horner et al., 2005).

Data from single case or small-*N* ITS design studies include many observations for an individual (or individuals) recorded over time. Dependent variables are measured repeatedly so that patterns of behavior can be clearly identified. The time during which behavior is observed is divided into phases: baseline (A), when natural performance of a behavior of interest is measured over a period of time without any intervention; and treatment (B), when an intervention is introduced and performance of the target behavior is again measured for some period during this intervention. Often, behavior is again observed after the removal of the intervention so that maintenance or retention of treatment effects can also be examined. This allows for subjects' pre-, during, and post-intervention performances to be compared (Richards, Taylor, Ramasamy, & Richards, 1992; Horner et al., 2005; Kratchowill & Levin, 1992; Franklin, Allison, & Gorman, 1996). Variations on this design include treatment-only (A), two-phase (AB), and withdrawal or reversal designs (e.g., ABA, ABAB), as well as more complex designs that involve more than one intervention (e.g., ABCA, ABACA) or variations in sequencing or start times of phases across subjects within one study (e.g., multiple baseline designs) (Richards, Taylor, Ramasamy, & Richards, 1992).

By practice and design, analysis of data from single case design studies often involves technical data considerations that are not necessary to consider when analyzing data from more common group design studies. For instance, dependent variables in single case design studies are often discrete rather than continuous variables; most commonly counts, in particular rates or proportions (i.e., percentages). All statistical analyses require assumptions about the distribution of the data to be analyzed (Agresti, 2002). Therefore, as with any design, it is important to consider the type of dependent variable when analyzing data from single case designs. Rates and

proportions are more suitably modeled by the Poisson or Binomial distributions, respectively, rather than the normal distribution often used for continuous variables.

The Poisson distribution is used to model the frequency of some event in a given period of time (i.e., a rate). For example, a dependent variable that expresses the number of times a student gets out of his/her seat during a 10-minute observation period is a rate and is best modeled by the Poisson. The Binomial distribution is used to model the frequency of some event out of a total known number of possible binary trials (i.e., a proportion or percentage). An outcome variable that expresses a proportion, like the correct number of responses out of 10 total items, is best modeled by the Binomial distribution. Unlike normal distributions, in which the variance is a completely separate parameter from the mean, in Binomial and Poisson distributions the variance is a function of the mean. For the Poisson, the mean and the variance are assumed equal. As the mean increases, so does the variance. Thus counts tend to vary more when their average value is higher. For the Binomial, the variance is expected to be less than the mean (Agresti, 1996).

Assumptions about the relational values of the mean and variance of these distributions are sometimes violated. When count data (including both rates and proportions) exhibit greater variability than would be expected by the Poisson or Binomial distributional models (i.e., when the variance exceeds the mean for either), this is called overdispersion. Overdispersion can be caused by statistical dependence or heterogeneity among subjects (Agresti, 1996, 2002). It is important that analytical methods acknowledge appropriate distributions for the dependent measures being studied and accommodate for any potential overdispersion present. Most software used for multilevel modeling (e.g., *HLM 6.0*) has an option to check for and accommodate overdispersion in its analysis (Raudenbush & Bryk, 2002; Raudenbush et al.,

2004). Many of the other statistical approaches to analyzing and synthesizing single case design data published to date do not suitably accommodate analyses of non-continuous outcome variables, let alone accommodate special cases with evidence of overdispersion. These will be identified below in a discussion of each method.

One of the most important technical considerations for data from single case design studies is the independence of data points. A major assumption of many statistical models is that observations are independent. That is, that the value of one observation does not depend in any way on the value of any other observations. However, in single case design data, independent and normally distributed observations and residuals are somewhat unlikely because sources of variation that operate at one time of observation are likely to produce carryover effects that influence later observations of the same subject (Franklin, Gorman, Beasley & Allison, 1996). When observations are serially dependent, that is when errors at one point are predictive of errors at future points, we say that residuals are autocorrelated. Typically, the closer one observation is to another in a series, the higher the correlation between their residuals. Correlations between the residuals of adjacent scores in a time series are known as lag-1 autocorrelation. Correlations between the residuals of scores that have another observation between them (e.g., observation 1 with 3, observation 2 with 4, 3 with 5) are called lag-2 autocorrelation and so on (Todman & Dugard, 2001).

Huitema (1985) initially argued that autocorrelation does not exist to a great extent in single case data. Several single case design methodologists cited his work as justification for ignoring this potential complication in defending their own approaches (e.g., Center, Skiba, & Casey, 1985; Gentile, Roden, & Klein, 1972; Gingerich, 1983; Scruggs, Mastropieri, & Castro, 1987a). However, many other researchers since have shown that autocorrelation is indeed

prevalent in single case design data and should be considered. These researchers criticized the logic of Huitema's (1985) work and counter-surveyed vast amounts of published datasets to prove the existence and significance of autocorrelation in single case design data. Huitema later acknowledged that autocorrelation is possible, and even likely, in these data and offered suggestions for analysis of data that contain autocorrelation (Huitema & McKean, 2000).

Autocorrelation, positive or negative, is an important data characteristic to consider when designing analyses for single case design data. High positive autocorrelation decreases within-phase variation in data (i.e., if successive observations are more similar than would be predicted by chance, variation will decrease), resulting in misleadingly small standard errors and thus inflated estimates of treatment effect (Jenson et al., 2007). The presence of autocorrelation in the data violates the assumption of independence of observations required for standard (or typical) hypothesis testing methods. The analysis method chosen has great influence over how well the autocorrelated data is handled.

Some methods are better at handling autocorrelated data than others. Statistical tests based on means (e.g., t-tests, F-tests) are invalidated by serial dependence. These distributions are known to be inaccurate for data that have dependent errors (Gorsuch, 1983; Matyas & Greenwood, 1996; Jenson et al., 2007). As discussed above, when residuals are positively autocorrelated, standard errors for regressions, t-tests, and analyses of variance (ANOVAs) will tend to be deflated, resulting in test values that are too large, which increases the likelihood of falsely finding statistical significance (i.e., a Type I error). [When residuals are negatively autocorrelated, standard errors will be inflated, because observations are less similar than would be expected by chance, and resulting test values will be too small (Busk & Marascuilo, 1992; Gorman & Allison, 1996; McCleary & Wayne, 1992; Raudenbush & Bryk, 2002; Richards,

Taylor, Ramasamy, & Richards, 1992; Van den Noortgate & Onghena, 2003b). In these cases, treatment effects will be underestimated.] Clearly, it is important to pay attention to proper handling of autocorrelated data. Even modest magnitudes of autocorrelation can increase the change of a Type I error considerably (Van Belle, 2008) and bias values of t-tests (Franklin, Gorman, Beaslely, & Allison, 1996). Some more sophisticated methods can deal with serially dependent data suitably (e.g., multilevel modeling); others cannot. Attention to autocorrelation will be discussed for each method.

Review of Analytic Methods

Typical methods for analyzing single case design data involve the presentation of raw or graphed data for readers to visually inspect and determine effect of treatment. Quantitative approaches have also been suggested; however, many of the statistical methods proposed to date are unsuitable for analyzing and synthesizing these data because of statistical limitations or lack of consideration of important and unique features of single case design data. Better methodologies are needed to appropriately analyze single case design data while accommodating the technical considerations discussed above. Beyond summarizing the effect of treatment on the average individual, syntheses of data from single case design studies can provide valuable information about how the effects of an intervention vary across cases. In the best circumstances, using the most flexible methods, we can assess whether and how the specific characteristics of individuals or settings can influence treatment effects.

Analysis methods proposed include visual inspection, those related to t- and F-tests (i.e., ANOVA and regression-based methods), those that involve the calculation of effect size(s), those that draw on time series methods, nonparametric methods, and those that rely on multilevel modeling to describe data patterns and effect of treatment. Techniques in each of these

categories are discussed below, followed by a demonstration of some of these methods in an attempt to validate the strengths and weaknesses of each. Considerations will include whether or not each method provides reliable estimates and rich interpretations of data patterns, as well as how well each method accommodates for the technical issues involved in analyzing data from single case research design studies.

Visual Analysis

The earliest and most traditional approach to examining patterns in single case design data is visual analysis. This approach typically includes investigation of the level, trend, and variability of recorded responses by visually inspecting graphs of the data. Data are examined for patterns both within and across phases. For instance, in assessing level, an analyst studies the mean value of the dependent measure in each phase along with any change (or shift) in this mean level across phases. In examining trend, a line of best-fit is imposed upon the data points for each phase and the slope (or rate of change) of these lines is compared across phases. Finally, in assessing within-phase variability, a visual analyst considers the degree to which data points fluctuate around the mean (level) or slope (trend) in each phase in considering the strength of these patterns (DeProspero & Cohen, 1979; Franklin, Gorman, Beasley, & Allison, 1996; Horner et al., 2005; Ottenbacher, 1990; Parsonson & Baer, 1992; Scruggs, Mastropieri, & Casto, 1987a).

Visual inspection may also involve judgment of the immediacy of change following the start or withdrawal of an intervention as well as judgment of the consistency of data patterns shown. Regarding the former, if changes in the measured dependent variable are not immediately apparent following the onset of treatment, inferring a treatment effect may not be warranted. The closer to treatment an effect is apparent, the stronger the inference that can be drawn about the intervention's effect on the target behavior. It may be invalid to attribute

delayed effects to the intervention alone as other unobserved factors may also be at work on the outcome pattern (Horner, et al., 2005). The consistency of data patterns across multiple representations of baseline or treatment phases should also be examined. If performance patterns differ widely from one baseline phase to another baseline phase, or from one treatment phase to another phase in which the same intervention was applied, the strength of any conclusions drawn may be limited (DeProspero & Cohen, 1979; Franklin, Gorman, Beasley, & Allison, 1996; Horner et al., 2005; Ottenbacher, 1990; Scruggs, Mastropieri, & Casto, 1987a).

Visual analysis has many strengths as a method for examining patterns in this type of data. It is a relatively quick and straightforward way of making preliminary judgments about treatment effectiveness. It does not require a great deal of statistical knowledge, only understanding of the time series graphs created from single case design data. Because of this, it is also easy to explain results to research consumers. Readers of results based on visual analysis do not need to have a sophisticated statistical background. Skilled visual analysts can simultaneously take into account level, trend, variability, overlap, delay and consistency to determine functional relationships. Visual analysis has been defended as a conservative approach because, proponents presume, only robust effects are identified by the naked eye (Ottenbacher, 1990; Crosbie, 1993; Brossart, Parker, Olson, & Mahadevan, 2006). However, still important but less obvious effects may not be identified if relying solely on visual assessment of graphed data. Simply, the scale of the axes of the graph may hide important patterns from view if axes scales are too compressed or graphs themselves are not shown large enough to make patterns clear (Robbins, 2005).

There are several other limitations to using visual analysis as a primary approach to analyzing single case design data. Most significantly, visual analysis is purely subjective and has

shown low interrater reliability even among highly trained raters (Center, Skiba & Casey, 1985; Crosbie, 1993; DeProspero & Cohen, 1979; Franklin, Gorman, Beasley & Allison, 1996; Ottenbacher, 1990; Scruggs, Mastropieri, & Casto, 1987a). DeProspero & Cohen (1979) systematically examined this subjectivity by asking trained reviewers to rate a series of graphs for treatment effect and experimental control. In this study, the authors prepared 36 four-phase reversal design (i.e., ABAB) graphs selected to represent a range of features. The variety of graphs presented included those that illustrated an ideal and consistent pattern of mean shift across phases (i.e., mean level is similarly low in both baseline phases and similarly high in both treatment phases) and others in which the degree of mean shift across phases varied from a proportionately small change to a much larger proportionate shift. Within-phase variability ranged from SD=.25 to SD=1 across the graphs. Finally, slope over time varied so that some graphs illustrated no trend at all (i.e., slope = 0 and lines are flat) and others graphs displayed increasing slopes or trends (i.e., slope $\neq$ 0) within phases.

Each of 250 behavior journal reviewers and editors was sent a set of nine of these 36 graphs and asked to rate how well each demonstrated experimental control on a scale of 0 to 100. Of the 250 reviewers, 114 replied. Interrater agreement was analyzed via correlational analyses of all ratings on each of the 36 graphs. The mean agreement (i.e., average interrater agreement correlation coefficient) found across all graphs (r=0.61, SD=0.26) is not particularly high or low. However, once the graphs with the most obvious and ideal patterns (i.e., large mean shift across phases and low within-phase variability) were removed, the authors found a wide variability in the ratings given to each of the remaining graph. A substantial degree of disagreement existed about what constitutes an effect even among these highly trained reviewers. The authors concluded that there does not seem to be any one characteristic of the graphs that determines a

reviewer's rating. Instead, they presume, when processing the graphic information, reviewers weigh the various features of a graph to come to a conclusion about treatment effects and that individuals weigh each feature differently. Because of a lack of standardized protocol for decision making, subjectivity prevails, allowing reviewers to assign different ratings to the same graphic information as they develop and use their own individual algorithms for determining treatment effect (Crosbie, 1993; DeProspero & Cohen, 1979; Franklin, Gorman, Beasley & Allison, 1996).

In addition to the unreliability of such a subjective method, visual analysis has also been criticized for its insensitivity to subtle changes in level or trend across phases (Busse, Kratochwill & Elliot, 1995; DeProspero & Cohen, 1979; Franklin, Gorman, Beasley & Allison, 1996). As discussed above, supporters of the method admit this limitation themselves, when they claim its conservative nature. But depending on the scale of the graph's axes and on study context, changes that seem subtle to the naked eye may in fact represent important and significant impact of practical value.

Visual inspection performs poorly in other areas and is especially discouraged in at least four circumstances: First, if baseline phases are not stable or do not include a sufficient number of data points, visual analysis is not recommended. Baselines reported with human subjects in applied journals are typically too short to draw strong inferences about initial pre-intervention levels of the dependent variable via visual analysis alone (Crosbie, 1993; Brossart, Parker, Olson, & Mahadevan, 2006). Without a reliable understanding of the pattern of behavior before the intervention is introduced, analysts cannot draw solid conclusions about change in the pattern of behavior due to treatment (i.e., treatment effect). Visual analysis is also discouraged as a primary method for assessment of effect when between-phase differences are smaller than

within-phase differences. When between-phase variability (i.e., mean shift) is small or similar to within-phase variability it is difficult to separate out variation due to treatment from variability due to the operation of uncontrolled factors (Franklin, Gorman, Beasley, & Allison, 1996; Horner et al., 2005; Crosbie, 1993). Visual analysis does not permit any statistical control for extraneous environmental factors. Visual inspection alone cannot determine if there are other characteristics of the subject(s) or setting that interact with the dependent variable, creating change irrelevant of the treatment (Brossart, Parker, Olson, & Mahadevan, 2006). Finally, as discussed above, visual inspection is limited by the apparent immediacy of changes in the outcome measure after the start or removal of treatment. If there is delay between the start of the intervention and any visible change in the measured dependent variable, visual analysis cannot validly determine whether or not the intervention itself was effective or if some other factor is responsible for the change (Horner, et al., 2005).

Visual inspection is not immune to the complications of autocorrelation either. As noted by Gorsuch (1983), "The same factors which make the significance test 'see' the data as more or less significant than they actually are may also influence the visual judgments (p. 142)". Autocorrelation has been shown to be positively correlated with Type I error rates from visual judgments (Brossart, Parker, Olson, & Mahadevan, 2006; Matyas & Greenwood, 1996). Matyas & Greenwood (1990) found that graphed data with higher positive autocorrelation produced more false positive identifications of treatment effect than those with lower autocorrelation (Matyas & Greenwood, 1996). This supports Gorsuch's (1983) claim in that autocorrelation produces bias in visual inspection that is analogous to that produced by statistical hypothesis testing.

Beyond its technical weaknesses, visual analysis on its own is also inadequate for synthesizing results across subjects or studies. It does not involve description or quantification of the effectiveness of treatment. Visual inspection does not allow for any expression of the average model for the data or for how subject, setting, or intervention variations affect the average model. Because synthesizing and expressing effects across subjects is the primary goal of the present study, visual analysis is not expected to provide useful and valuable conclusions on its own. However, it is true that visual inspection can provide value to analyses of patterns that should not be ignored. Visual inspection can help to detect features in the data like outliers, treatment effect delay, and any nonlinearity in trend that would need to be identified and then intentionally included in any statistical analysis. That is, visual analysis is useful as a preliminary overview of data patterns before any statistical models are built or run. Thus, it is expected that the use of visual inspection along with more sophisticated statistical methods will provide the most valid and complete assessment of data patterns and treatment effectiveness.

A series of statistical methods are presented below in general categories by type.

Regression-based Methods

The first suggestion for a statistical approach to evaluating treatment effectiveness in single case design data was for the application of ordinary statistics, analysis of variance (ANOVA). Building upon an earlier proposal by Shine & Bower (1971) to apply a one-way ANOVA to a single case (N=1) dataset, Gentile, Roden & Klein (1972) suggested the use of a two-way ANOVA to analyze single case design data from more than one subject (Gentile, Roden, & Klein, 1972; Shine & Bower, 1971). The two-way ANOVA, they offered, could be used to assess whether behavioral changes from phase to phase are large enough to be statistically significant and whether or not these changes are consistent across individuals.

Using Gentile, Roden, & Klein's (1972) method, phase (baseline, treatment) is treated as a between factor in the analysis and the repeated observations on each individual are treated as a within factor. That is, individual observations within phases are treated like subjects within groups and mean observations across phases are treated as analogous to between-group comparisons in a traditional group design application of two-way ANOVA. When synthesizing across more than one graph, individuals become another between factor, so that subject differences can be assessed along with the interaction between treatment and subject factors. Resulting F-statistics provide indicators of effectiveness from three points of view: (1) the between-phases F-statistic is an indicator of whether there are reliable differences between baseline and treatment phases, across all subjects; (2) the between-subjects F-statistic is an indicator of whether subjects performed differently from one another, across all phases; and finally, (3) the interaction F-statistic (phases by subjects) serves as an indicator of whether or not subjects reacted differently to the introduction of the intervention (i.e., is the baseline-treatment difference uniform across subjects?). According to the authors, statistically significant F-statistics suggest significant effects of each kind (Gentile, Roden, & Klien, 1972).

There are several strengths of this ANOVA approach. Firstly, before this method, no other quantifiable criteria had yet been developed, or published, to evaluate whether changes produced by the introduction of a treatment exceed the variability that may have resulted from the operation of uncontrolled factors. This is the first suggestion of a statistical significance test for analysis of these data. As well, the reliability of this method far exceeds that of the visual inspection method used almost exclusively to this point (Gentile, Roden, & Klein, 1972; Crosbie, 1993). While visual analysis is not likely to produce consistent judgments, statistical analysis, this approach being the first of its kind, produces the same results every time (DeProspero &

Cohen, 1979). Thirdly, the ANOVA method also allows the determination of variability of treatment effects across subjects in a systematic way. The F-statistic for the interaction term (phases by subjects) is a clear indicator of whether or not treatment effects vary significantly by subject. And finally, according to the developers of this approach, when assumptions of normality, homogeneity of variance, and independence of observations are met, it has been shown to maintain Type-I error at the desired level so that false positives are less likely (Gentile, Roden, & Klein, 1972; Crosbie, 1993).

However, as discussed above, these assumptions are often violated by the very nature of single case design data. For instance, the ANOVA approach is only truly appropriate when the dependent measure is a normally distributed, continuous variable. As we know, single case designs often employ counts – rates or proportions – as outcome measures, especially in behavioral research. These types of measures are more suitably modeled by the Poisson or Binomial distributions, respectively, rather than by the normal distribution that is used in this approach.

Also discussed above, single case data may also show evidence of dependence between observations. Gentile et al. (1972) argue that since nonadjacent phases (in designs more than 2 phases) must be combined (A1 with A2, B1 with B2) for the ANOVA approach, any non-independence between adjacent observations or phases is irrelevant. The authors contend that combining data like this will tend to make data in treatment phases more similar to data in their adjacent baseline phases than to each other, deflating the F-statistic and making this a more conservative test (Gentile, Roden, Klein, 1972; Scruggs, Mastropieri, & Casto, 1987a). By essentially ignoring serial dependency, treating observations as if they are independent when they are not, error variance and subsequently the F-statistic are artificially inflated or deflated,

depending on the sign of autocorrelation (Busse, Kratochwill, & Elliot, 1995; Crosbie, 1993; Kavale et al., 2000; Todman & Dugard, 2001). And as discussed above, this does have implications for resulting test statistics.

Finally, this approach ignores any effect of trend. By using only the mean level of each phase, Gentile, Roden, & Klein's ANOVA approach does not account for slope, ignoring the systematic changes that take place within and across phases. By excluding trends from consideration in the model, deviations from the mean level in each phase are then treated as errors (Center, Skiba, & Casey, 1985) instead of being used to better describe patterns of the data before and after treatment is introduced.

In response to these limitations, Gorsuch (1983) proposed an ANCOVA-like method for analyzing single case design data that first makes an effort to remove or reduce serial dependency from the data and then tests the developed model using an equation that includes a term for time or trend. Gorsuch (1983) assessed the performance of three different equations for this purpose and found trend analysis to provide the best control of Type I and Type II errors. Trend analysis is a test of the mean difference between phases while controlling for overall trend in the data. It is analogous to an analysis of covariance (ANCOVA) with time as the covariate in an equation regressing treatment on the dependent variable (Parker & Brossart, 2003), equation 1. According to Gorsuch (1983), this inclusion of a time slope should eliminate autocorrelation that arises from trend effects (but not necessary other sources).

$$Y = a + b_t(\text{Time}) + b_x X + u_t \text{ , where} \qquad (1)$$

(1) a is the mean level of the outcome variable during the baseline phase,

(2) b_t is the trend (i.e., influence of time) which does not vary throughout the series,

(3) b_x is the change in mean level from baseline to treatment phase (i.e., treatment effect), and

(4) u_t is the residual error component of this model.

Gorsuch's attention to trend (slope) is a strength of his suggested approaches above and beyond that of the ANOVA method before him. However, by implicitly assuming a constant trend throughout the study (i.e., baseline phase trend is assumed to be equal to treatment phase trend), this method does not test for any change in slope due to treatment (Allison & Gorman, 1993; Center, Skiba, & Casey, 1985; Faith, Allison, & Gorman, 1996). It is not unreasonable to presume that the intervention could lead to an altered slope (i.e., change in trajectory) in data as well as level. Not being able to test for and estimate that shift in slope is a clear drawback of this approach.

Gorsuch's ANCOVA-like regression technique has some additional limitations. Firstly, while Gorsuch's attempt to account for and decrease the effects of autocorrelation is effective at controlling for lag-1 positive autocorrelation that arises from trend effects, it does not attend to other potential sources of autocorrelation. As well, model building is based upon the total number of data points available, regardless of whether there is a strong or weak treatment effect. Finally, this approach estimates effects within single cases only. It does not suggest how to combine single case design data across subjects for meta-analytic purposes which is the focus of the present study. As well, it does not suggest how patterns in baseline phases may differ from each other or how treatment phases may differ from each other in an ABAB or more complex design study.

Effect Size Methods

Several authors have recommended approaches to single case design data analysis that involve the calculation of effect sizes, from simple application of the standard effect size formula originally developed for group design data to more complex approaches that take into account distributional properties of the single case design data and/or include regression techniques. Again, group design methods, however, are not always appropriate for application to single case design data because: (i) they do not account for important aspects of the data, and/or (ii) because they oversimplify data patterns, ignoring the intricacies of single case design data that separate it from group design data. For instance, Gingerich (1984) argued for the application of the Glassian meta-analytic effect size calculation, typically applied to the analysis of group designs. Gingerich (1984) suggested that an effect size could be computed by dividing the difference between the mean levels in the treatment and baseline phase ($\overline{X}_{treatment} - \overline{X}_{baseline}$) by the standard deviation of baseline phase data points ($SD_{baseline}$) (Faith, Allison, & Gorman, 1996; Glass, 1976; Gingerich, 1984). Equation 2 displays this formula:

$$ES = \frac{\overline{X}_{treatment} - \overline{X}_{baseline}}{SD_{baseline}} \quad (2)$$

The formula is straightforward and simple and allows for comparable interpretation of computed effect sizes in a way that is similar (but not identical) to those used in established group meta-analyses. However, the data of discussion are not from a group design study and as such, we must use caution in applying group design analytic methodologies to data from single case design studies. Again, serial dependency brings about complications from autocorrelation, common in single case design data, that are not as of great concern in group design data. Any

non-zero autocorrelation can result in artificially inflated or deflated effect sizes calculated in this manner (Faith, Allison, & Gorman, 1996). Gingerich admits this possibility and although he advises caution when autocorrelation is suspected, he also cites Huitema's preliminary research suggesting that time series datasets do not often contain autocorrelation anyway (Gingerich, 1984). (As discussed, Huitema later rescinded this position.)

Another limitation of Gingerich's (1984) approach is that it ignores any trend, or slope, in the data, which can lead to least two problems. Firstly, it may gloss over important pre-intervention data patterns in the baseline phase, ignoring any preexisting trend. By assuming baseline phase trend lines are flat (slope = 0), this method does not allow for the very real possibility that behavior was already changing before the introduction of the intervention. Secondly, this method may ignore any apparent effect of the treatment on the trend as well, evidenced by a change in slope from baseline to treatment phase (Faith, Allison, & Gorman, 1996). That is, this method also restricts modeling of the treatment phase trend by assuming it is flat (slope = 0) as well. In fact, Gingerich admitted that this technique is not appropriate for use with data that contain trend (Gingerich, 1984). The author admits that using phase means may also be inappropriate if there is wide within-phase variability or an abnormal distribution of data points. These stipulations greatly limit the scope of the data that is appropriate for analysis via this approach.

In general, effect size methods like these are misleading because although they may look similar, they do not produce values comparable to those from group design research. Standard deviations in within-subject studies (i.e., variations in measurements for an individual) are often

much smaller than those from between-subject studies (i.e., variations in measurements across individuals). Therefore, while group design effect sizes usually range from about -1 to +1, single case design effect sizes can often be much larger (since this smaller standard deviation is used as the divisor in the calculation of this effect size). If someone incorrectly compares the two types of effect sizes (group design and single case design), they might falsely conclude a large treatment effect where none existed.

White, Rusch, Kazdin, & Hartmann (1989) recognized some of these limitations and recommended a modification to Gingerich's (1984) formula for computing effect sizes from single case design data. First, they suggested taking into account the variability of data points in the treatment phase(s) and the overall trend in the data by using a pooled standard deviation in the effect size formula denominator. If there is indeed trend in the data, they also recommended further modifications to the formula, as seen in equation 3.

$$ES = \frac{X'_t - X'_b}{SD'}, \text{ where} \qquad (3)$$

(1) X'_t = the predicted level of the observed variable on the last day of the Treatment phase, as predicted by a within-phase regression,

(2) X'_b = the predicted level of the observed variable on the last day of the Baseline phase, as predicted by a within-phase regression and,

(3) $SD' = \sqrt{\mathrm{SD}_{\mathrm{pooled}} * (1 - \mathrm{r}^2)}$, where

(3a) SD_{pooled} is the pooled standard deviation of data points in both phases and,

(3b) r is the correlation between the target measure and the day of observation.

White et al. (1989) claimed this method can be used with data from designs with two or more phases. If there is equivalent performance across phases of the same condition, then White et al. (1989) recommend averaging baseline phases and averaging treatment phases. If, though, there is unequal performance across replicated phases, the authors suggest using data only from the first baseline and the last treatment phases.

This approach improves upon Gingerich's (1984) but shares many of the same limitations. It does control for lag-1 autocorrelation due to trend effects to some degree (by including attention to the correlation between the target measure and time) but it does not attend to other possible types of autocorrelation. It also improves upon Gingerich's (1984) technique by allowing for the important impact of trend in the formula. However, this method is not very reliable. Different effects can lead to equal estimates of effect size as long as the proportional relationship between treatment-control phase differences and variability of data points remain the same (Van den Noortgate & Onghena, 2003a). Other problems include the fact that it cannot accommodate designs any more complex than those with one intervention (e.g., AB, ABA, ABAB) and that outcomes are dependent on the length of the study (like most other approaches of this kind). And like all other one-number summaries of effect, much of the information about the intricate data patterns uniquely illustrated by single case designs is lost when only one number expresses effect. One number alone cannot express the effect of the treatment on specific patterns in the data.

Another attempt to include trend in the assessment of effect along with regression techniques is that of Center, Skiba, & Casey (1985). Building from Gorsuch's work, Center, et al. (1985) proposed a piecewise regression technique that involves the calculation of three separate effect size estimates for changes in level and slope. This proposed regression equation

(equation 4) includes a term for baseline level (b_0), a term for change in level (b_1X) from baseline to treatment, a term for baseline trend (b_2t) and an interaction term to measure the change in slope due to treatment ($b_3X(t-n_a)$), where n_a is the number of observations in the baseline phase:

$$Y_i = b_0 + b_1X + b_2t + b_3X(t-n_a) + e_i \quad (4)$$

To obtain an effect size for each change, the authors proposed computing the model with and without each of the effect terms and then: (1) differencing the R-squared of the 2 models, (2) converting the R-squared difference to an F-statistic, (3) converting the F-statistic to a *d*, and then, (4) using the *d* to calculate one effect size for each parameter. (This process seems unnecessary since a *t* statistic, automatically provided for each parameter, would essentially serve the same function.) If there are more than two phases (AB) in the single case design being analyzed, Center et al. (1985) suggest using data from only the first baseline-treatment pairing and ignoring the remaining replications altogether (Center, Skiba, & Casey, 1985).

As a statistical option for analyzing data from single case designs, Center et al. (1985) claim that this approach is less constraining than some of the other methods discussed. For instance, they assert that it produces more conservative estimates of effect size than does the method proposed by Gentile et al. (1972) because it accounts for and attempts to model trend in the data. In addition, it allows for separate assessments of variance accounted for by changes in level, changes in slope, and the combined effects of level and slope changes.

However, as a claimed effect size method, it is unclear which, if any, of the three separate effect sizes calculated is most comparable to the typical *treatment effect* estimate in group design (Busk & Serlin, 1992; Busse, Kratochwill, & Elliot, 1995; Kavale et al., 2000; Scruggs, Mastropieri, & Casto, 1987a; Van den Noortgate & Onghena, 2003a; White, et al., 1989).

Additionally, without a large number of data points, this method may not produce reliable, unbiased, consistent, results (Busk & Serlin, 1992; Center, Skiba, & Casey, 1985; Scruggs, Mastropieri, & Casto, 1987a; White, et al., 1989). Subjects with more observations are weighed more heavily than those with less. And, it may also underestimate effect (in regards to change in slope) by overestimating baseline trend (Allison & Gorman, 1993; Faith, Allison & Gorman, 1996).

Center et al.'s (1985) piecewise regression technique also cannot handle more complex versions of single case designs, like multiple baseline or removal and reintroduction of treatment (e.g., ABAB) as is evidenced by their suggestion to ignore all but the first AB pairing in the data analysis. And though it does provide quantifiable estimates of effect, it ignores the direction of effect. The lower bound of any identified effect is 0, and sometimes treatment does produce worse results than no treatment at all; or in the case of reducing undesirable behavior, a negative effect is the target (Allison & Gorman, 1993; Faith, Allison & Gorman, 1996). Another limitation of this approach is that, like the ANOVA proposal, regression is based on assumptions of normal distribution and independence of data points. As discussed above, these assumptions are often violated by single case design data. These authors, like those before them, ignore the issue of autocorrelation, citing Huitema's publication about the low incidence of autocorrelation in single case design data overall, which was later discounted.

In response to concerns about underestimating effect by overestimating baseline trends, Allison & Gorman (1993) proposed an adaptation of Center et al.'s (1985) approach that also combined regression techniques with a focus on effect sizes. The technique includes first fitting the baseline data series only to a simple regression equation that includes a term for the baseline intercept (b_0) and for baseline slope over time (b_1t).

$$Y_t = b_0 + b_1 t + e \quad (5)$$

This baseline-only regression equation is then projected into both the baseline and treatment phases and residuals between the actual data and the baseline projections are computed for both conditions. In a third step, a time series is formed from the residualized values and then Center et al.'s (1985) approach is applied to analyze the residual series and compute an effect size. If the treatment is effective, treatment phase data points might deviate greatly in level and/or slope from the projections based on the baseline series alone, as indicated by large residual values under the treatment condition. Allison & Gorman (1993) include a step-by-step algorithm for their approach. This method is appropriate for use with two-phase AB or reversal ABAB designs only. If applying it to a four or more phase design, the authors suggest using data from the first A and last B phases only.

Allison & Gorman's (1993) method effectively responds to concerns about underestimating effect by computing trend based on baseline data only and measuring residuals from those projections. Another strength of the approach is that it assesses treatment effects on both level and slope while controlling for trend in the baseline (Allison & Gorman, 1993). According to its authors, it has also been shown to control for lag-1 autocorrelation due to trend quite effectively (Allison & Gorman, 1993).

However, there are several limitations of the approach. Like most other methods discussed, results of this method are dependent upon the length of the study, or the number of data points available (Van den Noortgate & Onghena, 2003a). Secondly, Allison and Gorman's (1993) approach has not been shown to perform well with negatively autocorrelated data. Finally, this method does not directly address the issue of synthesizing data across subjects, the primary focus of the present study.

In spite of such problems, effect size methods have remained popular in the methodological literature. For example, around the same time that Allison & Gorman (1993) published about this method, Busk & Serlin (1992) also suggested calculating an effect size to assess treatment effect in single case design data. These authors offered three different approaches with varying denominators in the effect size equations. The method chosen, the authors said, should be dependent upon distributional characteristics of the data (Busk & Serlin, 1992; Faith, Allison, & Gorman, 1996; Kavale et al., 2000; Van den Noortgate & Onghena, 2003a).

Busk & Serlin (1992) support Gingerich's (1984) suggestion that Glass's effect size calculation (see equation 3) may be used when no assumptions about the distribution of the data are made. An effect size is calculated for each subject's data and then a binomial sign test (i.e., is it true that 50% of the differences are positive and 50% are negative?) is used to obtain a confidence interval and draw conclusions about the significance of the overall treatment effect.

If instead a more stringent assumption is made that within-phase variances and covariances are equal across phases (i.e., variances and covariances in baseline and treatment phases are the same) then, like White et al. (1989), the authors recommend that a pooled standard deviation should be used to achieve a better estimate of the denominator. An effect size would be calculated for each individual's data using the formula in equation 6 and then a Wilcoxon model (taking into account the symmetry of the values around zero and not just the signs as in the binomial sign test) is used to obtain a confidence interval and draw conclusions about treatment effectiveness.

$$ES = \frac{\overline{X}_{treatment} - \overline{X}_{baseline}}{SD_{pooled}} \qquad (6)$$

Finally, if the data may also be assumed to be normally distributed, then a further alteration to the denominator is recommended. Here, after effect sizes are calculated for each individual's data, a *t*-distribution would be used to get a confidence interval about the overall treatment effect.

$$ES = \left[\frac{\overline{X}_{treatment} - \overline{X}_{baseline}}{\sqrt{MS_{error}}} \right], \text{ where} \qquad (7)$$

$$MS_{error} = \frac{SS_{error}}{df}$$

Use of Busk & Serlin's (1992) methods allows for the assessment of individual as well as overall effect size, something lacking in the other effect size calculation suggestions. And, these methods do take into account distributional characteristics of the data. However, they have some substantial drawbacks as well. Most obviously, these approaches like many others, ignore trend in the data. There is no inclusion of data trend in any of the formulas recommended by Busk & Serlin (1992). This ignores an important piece of the behavior patterns illustrated by single case design data and can lead to the calculation of misleading effect sizes. That is, they may be unreliable. As with some of the other effect size computation methods discussed to this point, different effects can lead to equal estimates of effect size, using these formulas. Thirdly, these methods do not control for all possible types of autocorrelation, a danger already discussed in detail above. And finally, although Busk & Serlin's (1992) methods allow for the calculation of group effect sizes, each assumes that the effect of the treatment is the same for all subjects being observed. In other words, with this set of methods, we could not test the heterogeneity of individual effects (i.e., whether treatment effects are different for different individuals) (Busk & Serlin, 1992; Van den Noortgate & Onghena, 2003b).

In general, effect size methods have been rightfully criticized for their limitation of the description of data patterns down to one-number estimates. One simple estimate of effect size cannot sufficiently describe treatment effect, independent of the length of the experiment and without losing important information about the kind of effect present (Van den Noortgate & Onghena, 2003a). How can one number simultaneous summarize patterns in data level, trend, variability, phase change, etc? More sophisticated, complex time series methods can provide richer outcomes and better descriptions of how target measures may be impacted by the treatment being studied.

Time Series Methods

Time series methods combine several techniques in a regression approach to include the influence of previous observations and of accumulated residual error in the prediction and modeling of repeated observed behaviors. Box-Jenkins' (1976) autoregressive integrated moving average approach (ARIMA) assumes a concept known as stationarity, that time series have constant mean levels and that variability around that mean level is constant throughout the series (Gorman & Allison, 1996). Thus, the ARIMA approach is inappropriate for use with data that contain systematic trend (or slope). Instead, the method suggests transforming observation data (e.g., using differencing methods) before modeling. The resulting model includes terms that express autoregression (i.e., that a given observation can be predicted from past observations), differencing (i.e., data transformation), and moving average components (i.e., the degree to which residual errors linger in the series). There are an infinite number of ARIMA models that might fit a data series. The optimal combination of autoregressive terms, differencing terms, and moving average components is sought. After examining any trends, any uneven variability,

and/or any autocorrelation in the data, an ARIMA analyst would attempt to fit models and assess the fit of each to the data, checking that assumptions are met.

This approach requires a large number of data points in order to achieve reliable estimates (Scruggs, Mastropieri, & Casto, 1987a; Todman & Dugard, 2001; Van den Noortgate & Onghena, 2007). As discussed, applied behavior research typically contains relatively few pre- and few post- intervention observations.

As an alternative, an interrupted time series (ITSE) technique was proposed by Gottman (1981) that combines the ARIMA method with a regression approach. This approach uses autocorrelation information, dummy variables, and trend information to analyze time series data (Gorman & Allison, 1996). ITSE assumes that autocorrelation exists and that each phase has a different intercept and slope. General linear modeling methods are used to determine whether these intercepts and slopes are significantly different from baseline to treatment phases (Crosbie, 1993). A pre-treatment (i.e., baseline) model might look like equation 8a:

$$Y_t = m_1 t + b_1 + \sum_{1}^{k} a_i y_{t-1} + e_t \qquad (8a)$$

and a post-treatment model might look like equation 8b:

$$Y_t = m_2 t + b_2 + \sum_{1}^{k} a_i y_{t-1} + e_t \text{ , where} \qquad (8b)$$

(1) m_1 and m_2 represent baseline and treatment phase slopes, respectively,

(2) b_1 and b_2 express baseline and treatment intercepts respectively, and

(3) a_i's represent autocorrelation terms (of which there may be up to *k* such terms).

Gottman and Williams also developed a computer program to estimate the *a*, *b*, and *m* terms and to assess the differences between pre- and post- slopes and intercepts (Gorman & Allison, 1996).

Crosbie (1993) noted that the ITSE approach is often problematic with short series because it uses a lag-1 autocorrelation term that is biased by small sample sizes. By not removing all positive autocorrelation, ITSE leaves an increased risk of a Type I error (Crosbie, 1993). Crosbie modified the ITSE technique to provide better, less-biased estimates of treatment effects in short series, by adapting the ITSE model to include an improved estimate of first-order autocorrelation (Crosbie, 1993; Gorman & Allison, 1996).

Crosbie's (1993) interrupted time series analytical methods with an improved estimate of autocorrelation (ITSACORR) is simpler and, according to its author, requires little expertise in analysis as a computer program performs all computation and decision making. It is claimed to have acceptable statistical power, even with series as short as 10 observations and to be successful in providing better control of Type I error than the ITSE method it is based on. Although, other authors question if this improvement is sufficient (Busse, Kratochwill, & Elliot, 1995; Kavale et al, 2000).

Even with the computer program, analysis using ITSACORR can be especially complex to interpret and explain. As well, Crosbie (1993) recommends that 10 data points per phase be held as a minimum in order to achieve a more accurate estimate of autocorrelation and better power. If data patterns are complex, this minimum increases greatly. Test statistics may be inflated where data series are as short as $n<20$ and first-lag autocorrelation is high ($r_1>.6$). In addition, there is no mention of using the ITSACORR method for data synthesis across subjects, a primary focus of this paper. Therefore, time series methods, even one as improved as ITSACORR, are not likely to be useful here either.

Nonparametric Methods: Two very different approaches

As discussed throughout the above, ordinary t-tests and ANOVAs and other parametric statistical procedures are often inappropriate for single case design data because they require assumptions that are regularly violated by data from these types of designs. As an alternative, nonparametric methods have been proposed to analyze and synthesize single case data without reliance on the stringent assumptions required by parametric tests. Two very different nonparametric approaches will be discussed: Scruggs, Mastropieri, & Casto's (1987a) percentage of non-overlapping data points and Edgington's (1980) randomization tests.

In 1987, Scruggs, Mastropieri, & Casto (1987a) published an article suggesting one possible nonparametric method that is still creating controversy more than 20 years later. They proposed assessing the percentage of non-overlapping data points (PND) across phases as an indication of treatment effect; a simple calculation combined with visual inspection that does not rely on any distributional assumptions of the data (i.e., nonparametric). Specifically, the number of data points in the treatment phase(s) that exceed the highest baseline data point (in the expected direction) is divided by the total number of data points in the treatment phase(s).

$$\text{PND} = \frac{\text{\# data points in treatment phase which exceed the highest data point in previous baseline phase}}{\text{total \# of data points in the treatment phase}} * 100 \quad (9)$$

This calculation can combine multiple baseline and treatment phases within one individual's graph (by combining the numerator and denominator in each proportion fraction) or across several individuals' graphs.

To its credit, the PND approach is easy to compute and simple to interpret (Faith, Allison, & Gorman, 1996; Kavale et al., 2000; Scruggs, Mastropieri, & Casto, 1987a). According to its authors, PND is systematic yet free of any assumptions of normality, homogeneity of variance,

or independence of data points. The PND approach can also synthesize studies and/or be compared across studies.

However, the drawbacks of using the PND method to assess the effect of treatment in single case design data far outnumber these advantages. In fact, the controversy of the validity of the PND approach has spawned at least eight whole publications (Allison & Gorman, 1994; Salzberg & Strain, 1987; Scruggs & Mastropieri, 1994, 1998; Scruggs, Mastropieri & Castro, 1987a, 1987b, 1987c; White et al., 1989) and mention in at least as many more (e.g., Allison & Gorman, 1993; Busk & Serlin, 1992; Busse, Kratchowill & Elliot, 1995; Faith, Allison, & Gorman, 1996; Kavale et al., 2000; Shadish & Rindskopf, 2007; Strain, Kohler, & Gresham, 1998).

The most severe limitation of the PND approach is that it relies only on the single, most extreme data point in each phase. By ignoring the overall variability of data points across the duration of the study, PND disregards a glaring benefit of single case designs as illustrations of behavior patterns, not just endpoints (Allison & Gorman, 1994; Salzberg, Strain, & Baer, 1998; White, 1987). The movement of just one data point in either phase can effect a 100% change in the PND estimate. The movement of all but that one extreme data point in each phase can result in a 0% change in the PND. Whether the data points in the treatment phase barely exceed the highest point in the baseline phase or far exceed the highest point in the baseline phase makes no difference to the PND computed from the data. Suffice it to say, the PND approach does not provide a reliable indicator of treatment effect (Allison & Gorman, 1994; Salzberg, Strain, & Baer, 1998; White, 1987).

In that it does not take into account all data points, it is clear that PND also ignores any trend in the data, again forsaking the ability to examine patterns over time provided by single

case designs (Allison & Gorman, 1993; Jenson et al., 2007; Kavale et al., 2000; Scruggs, Mastropieri, & Casto, 1987a; White, 1987). As a solution, Scruggs et al. (1987a) recommend excluding data from PND analysis if it includes trend. However, trends are meaningful to the examination and expression of what is going on in the data, of what behavior patterns look like both before and after an intervention is applied and of how those patterns relate to one another. By ignoring any data with a non-zero slope, we would be unacceptably biasing the outcomes of the synthesis.

The authors themselves admit that PND is inappropriate in a number of other common circumstances as well. For instance, PND should not be applied when there are floor or ceiling effects in the data, when there are all or most zero-baseline data, or when the design is complex (Scruggs, Mastropieri, & Casto, 1987a). Because of the issues discussed above, other authors have added that the PND method is especially inappropriate when there are outliers in the baseline phase(s) or if treatment has a (potential intentional) detrimental effect (Allison & Gorman, 1993; Faith, Allison, & Gorman, 1996). Additional criticism has included the fact that the PND method ignores the issue of autocorrelation. Serial dependency in the data should not be overlooked even when data with baseline trends are ignored (White, 1987).

Scruggs et al. (1987a) have responded to the many critiques of the PND approach through several additional publications (Scruggs & Mastropieri, 1994, 1998; Scruggs, Mastropieri, & Casto, 1987b, 1987c) offering weak defenses to many of the criticisms received. Instead, they argue irrelevantly that the validity of the PND method is evidenced by its abundant publication in research journals. Throughout, they fail to sufficiently justify their most significant criticism: that PND reduces the most relevant and rich information in single case designs (e.g., variability, slope) down to a one-number summary of performance across phases

calculated from the relative positioning of two extreme data points, ignoring the complexity of real data instead of trying to model or express it.

An alternative non-parametric method shares virtually no similarities with the PND approach beyond this classification. Unlike PND, randomization tests attend to the entirety of the dataset, not just the most extreme points, and involve more sophisticated computation. Edgington (1980) recommended at least three types of randomization tests that have been discussed in the literature; of these, permutation tests are most prominent. Each of the three types involves randomly dividing the data and assessing the distribution of statistics computed from the multitude of data divisions. Permutation tests involve: (i) computing some test statistic (e.g., t, F, etc.) to measure phase differences on an original set of data, (ii) reordering and dividing the data repeatedly for all possible permutations (orders) of the data, (iii) recomputing the test statistic for each data division, and finally (iv) comparing the proportion of these computed values that are as large or larger than the original statistic's value [in (i)] to determine significance (Edgington, 1980, 1992; Edgington & Ongehna, 2007; Gorman & Allison, 1996). If too many permutations are possible from a set of data, the analyst may also choose to randomly sample a selection of data rearrangements from which to calculate the proportion of significance (Edgington, 1992).

Randomization tests maintain several strengths. The logic of the randomization test is not extremely sophisticated. It can likely be understood by those with limited mathematical backgrounds. Also, available computer software can automatically divide and calculate data to determine statistical significance. As well, even though it is not assumption-free, it is somewhat more flexible in its assumptions than the parametric methods discussed abobve.

However, randomization tests are not without limitations either. Some of the more restrictive distributional assumptions of parametric methods like ordinary t-tests, ANOVAs and regression-based methods may not be applicable to randomization tests but randomization tests do maintain some restrictions. Though more flexible, these restrictions still force a structure and constraint on the data that may be difficult to meet. For instance, as a nonparametric procedure, randomization tests do not require data to come from a normal distribution. However, the two groups (or in this case, phases) from which data are to be compared, must be similarly distributed in shape and variance (i.e., homogeneity of variance). Otherwise, conclusions drawn may be invalid (Gorman & Allison, 1996; Todman & Dugard, 2001).

Random assignment is another assumption of parametric tests that this nonparametric alternative is required to meet, though in a different way. In lieu of random sampling, a requirement for parametric tests of group design data, single case data should come from studies in which treatments were randomly assigned to observation points, in order for randomization tests to produce valid results. In other words, to be wholly valid, start and end times for the different phases (i.e., baseline or treatment) should be assigned at random, regardless of stability of baseline or treatment phase trends. Regardless of whether assignment to treatments is random or not though, the values of the test will remain the same (Gorman & Allison, 1996; Todman & Dugard, 2001). That is, findings from randomization tests should not differ whether or not phases are randomly assigned to data points; but the conclusions that are drawn from these findings will hold greater validity if treatment is randomly assigned to observation times.

Finally, like all hypothesis tests, randomization tests require some evidence of independence of observations. Although randomization tests have been promoted as an alternative to dealing with autocorrelated errors common in typical single case design data, it

might not be true that this assumption can be totally ignored (Edgington, 1980, 1992; Kavale, et al., 2000; Richards, Taylor, Ramasamy, & Richards, 1992). Randomization tests assume no carryover effect. If there is a carryover effect from one data point to the next, scores in baseline trials that follow treatment trials would measure closer in value to the preceding treatment trial. This would artificially minimize the difference between phases (Matyas & Greenwood, 1996). As with the approaches discussed above, any positive autocorrelation in the data will incur underestimation of values for determining statistical significance and any negative autocorrelation will incur overestimation of values for determining statistical significance (Gorman & Allison, 1996). Edgington (1980) claims that autocorrelation is probably less of a problem in randomization tests than it is in parametric analyses, especially if treatments are randomly assigned to observation times (Edgington, 1980; Todman & Dugard, 2001), and he makes no provisions for assessing or dealing with serial dependency using randomization tests to analyze this type of data (Matyas & Greenwood, 1996).

This approach serves only as a significance test not a modeling technique. It does not provide estimation or description of patterns in the data, of what characteristics affect what outcomes in what ways (magnitudes, directions). It provides only a simple non-directional test of the difference between phases. As well, it does not value trend in any way. Like some of the earlier methods discussed above, randomization tests do not consider or estimate slope in the data. Lastly, since this method only tests the significance of effects and does not make any estimates of effects, it is clearly then unable to synthesize across subjects in small-*N* designs to determine overall treatment effectiveness.

Multilevel modeling methods

Multilevel modeling offers solutions to several of the limitations of the analytical methods discussed to this point. By considering various levels of data, multilevel modeling accommodates the often nested structure of educational and psychological data. Frequently, multilevel modeling is used to analyze data from students that are nested within classrooms and/or classrooms that are nested within schools. These design specifications are also relevant to the single case design data of interest in this paper. For these types of data, it is possible and appropriate to consider observations as nested within individuals.

Multilevel modeling approaches distinguish the levels in the data structure (e.g., observations, individuals) by formally expressing each level with its own statistical model. Each sub-model includes the expression of the relationships occurring at that level along with any residual variability. The characteristics of the units from one level are used as predictors in describing the coefficients of the equation for the level just below (Raudenbush & Bryk, 2002; Raudenbush, Bryk, Cheong, Congdon, & du Toit, 2004; Schreiber & Griffin, 2004; Van den Noortgate & Onghena, 2003b). In the case of synthesizing results within a small-*N* A-B design study, data are available on two levels: observations on individuals (level-1) with predictors at the observation level (such as time and phase) and at the individual or subject level (level-2) (subject characteristics such as age and gender).

The equations displayed in Figure 1 express the levels of the hierarchical model for an ABAB single case design study. At level-1, an observation (Y_i) for a subject (j) can be predicted from a baseline intercept (i.e., level during the first baseline – A – phase) for that individual (P_{0j}); a baseline slope (P_{1j}); the effect of treatment on the intercepts and time slope (P_{3j} and P_{4j}); an effect of order on the intercept and time slope (i.e., any observed difference between intercepts

and slopes during the first AB pair and intercepts and slopes during the second AB pair) (P_{3j} and P_{5j}); the effect of treatment on the order effect (P_{6j}); and finally, a three-way interaction between time, treatment and order (P_{7j}). At level-2, each of the level-1 coefficients can be predicted from a mean estimate (*B*'s) and random effect (*R*'s). Subject characteristics could also be included in these equations, allowing valuable explanation of variation in average estimates.

Figure 1. Equations for Full Multilevel Model of ABAB Study Data

Level-1 (Observations)

$$Y_{ij} = P_{0j} + P_{1j}*(time) + P_{2j}*(trt) + P_{3j}*(order) + P_{4j}*(time*trt) + P_{5j}*(time*order) + P_{6j}*(trt*order) + P_{7j}*(time*trt*order) + e_{ij}$$

Level-2 (Subjects)

$$P_{0j} = B_{00} + R_{0j}$$
$$P_{1j} = B_{10} + R_{1j}$$
$$P_{2j} = B_{20} + R_{2j}$$
$$P_{3j} = B_{30} + R_{3j}$$
$$P_{4j} = B_{40} + R_{4j}$$
$$P_{5j} = B_{50} + R_{5j}$$
$$P_{6j} = B_{60} + R_{6j}$$
$$P_{7j} = B_{70} + R_{7j}$$

The goal is to model the target behavior (dependent measure, *Y*) of each individual and then connect these within the study to look for commonalities and differences across subjects. It is possible to estimate average effects across individuals and to explore any variation from these average effects by using subject-level (Level-2) characteristics to explain that variation. This method also attends to the technical considerations partially or wholly ignored by the other methods.

The multilevel modeling approach to synthesizing single case design data is flexible, adapted according to the data available. It allows treatment effects to vary across individuals, and allows the explanation of heterogeneity of treatment effects by subject characteristics. This

method does not require a large number of data points or a large sample of subjects. Because information is used efficiently to estimate parameters, it can be used with data that has only a small number of observations available for each subject and still obtain reliable estimates. User-friendly software is available to run analyses, test parameters, and accommodate for limitations of the data such as overdispersion and autocorrelation (Van den Noorgate & Onghena, 2003a, 2003b). If appropriate selections are made, findings can provide richer explanation of patterns in the data, not possible using the methods listed above, that have important implications. For instance, being able to explore how treatment effects may differ across subjects or settings can help to guide clinicians as to where and with whom particular treatments are best implemented (Light & Pillemer, 1984; Van den Noortgate & Onghena, 2003b).

As with all of the methods described above, the power of this approach in re-analyzing data is limited by the breadth of information reported by a study's original author. Reports of studies may not give enough information about subjects or study characteristics to serve as potential explainers of variation in intercepts or slopes. Without such information, an analyst may not be able to explain enough of this variation. Failure to explain substantial proportion of between-subject or between-study variance can hamper interpretation and practical utility of results. Finally, the flexibility of the multilevel modeling approach leads to several choices in the model-building process. Although there is some protocol, it is essentially up to the analyst to make these decisions at each turning point. Thus, different models and conclusions (some appropriate, some not) may be drawn from one dataset.

Despite these issues, this approach overcomes many of the limitations of the other approaches discussed. For instance, several methods provide only one number expressions of the rich patterns existing in the data pre- and post-intervention (e.g., Busk & Serlin, 1992;

Gingerich, 1984; Scruggs, Mastropieri, & Casto, 1987a; White et al., 1989). This method instead preserves and utilizes this rich information, expressing the functional relationships present in the data. In contrast to other methods that either ignore trend altogether or assume it remains constant throughout the study (e.g., Busk & Serlin, 1992; Gingerich, 1984; Gorsuch, 1983; Scruggs, Mastropieri, & Casto, 1987a), the multilevel modeling method includes attention to baseline trends and to how these trends change across phases. Trend is modeled instead of ignored or oversimplified.

Instead of disregarding autocorrelation completely (e.g., Gentile, Roden, & Klein, 1972; Scruggs, Mastropieri, & Castro, 1987a; Center, Skiba, & Casey, 1985; Gingerich, 1983) or only partially controlling for only some types of autocorrelation (Gorsuch, 1983; White et al, 1989; Busk & Serlin, 1992; Allison & Gorman, 1993), the multilevel modeling approach offers an opportunity to assess and deal with autocorrelated residuals. However, the analyst must seek it out; must test and measure it intentionally, so that it may be modeled. If autocorrelation exists, and if serial dependency is not modeled, its presence in the time series data can produce biased significance tests of random effects. In other words, it might look like there are reliable differences among subjects that warrant the inclusion of subject-level factors in the model, when, in fact, differences are unreliable. If variation among subjects is unreliable, then attempts to explain that variation by adding variables to level-2 equations are not likely to be successful (Jenson et al., 2007). It is important that autocorrelation is considered and accommodated for in use of this and any approach.

Unlike some other approaches, the multilevel modeling approach is not limited to use with data with equal number and spacing of observations (e.g., Gorsuch, 1983), continuous outcome variables (e.g., Center, Skiba, & Casey, 1985), simple phase structures (e.g., Allison &

Gorman, 1993), or random assignment of treatments to times (e.g., Edgington, 1980). Multilevel modeling is a flexible method that can handle unequal numbers and spacing of observations, count outcome variables (like rates and proportions), complex designs, and fewer data points than otherwise possible. It can also test for and accommodate for overdispersion, discussed earlier, which is common in categorical outcome data. And most importantly, the multilevel modeling approach can combine single case data across subjects or studies, offering opportunities to explore and explain heterogeneity of effects across replications of the intervention, the goal of the present study.

Estimation Methods: ML vs. EB

In multilevel modeling, model parameters can be estimated using any of several methods. Different estimation methods can lead to somewhat different estimates of effects, so the choice of which estimation method to use can have implications for interpretation of the results. For example, *HLM 6.0* program utilizes an inference approach that combines maximum likelihood (ML) and empirical Bayes (EB) estimation, whereas the *WinBUGS* computer program, also capable of conducting multilevel modeling analyses, employs fully-Bayesian estimation..

The estimation method used by *HLM 6.0* makes conditional estimates of fixed effects, based on estimates of the random effects. This works well for large samples, but when the number of Level-2 units (i.e., sample size) is small, like the data of interest in this paper, this method may underestimate uncertainty due to its assumption that the random effects are known (when they are not), which inflate significance test values for fixed effects (i.e., standard errors are deflated making test statistics inflated and significance levels deflated). When sample sizes are small, there may not be sufficient information to produce stable estimates of these effects

with the certainty implied by the ML or EB approach (Afshartous & de Leeuw, 2005; Brown & Draper, 2006; Raudenbush & Bryk, 2002).

Fully Bayesian estimation, on the other hand, approaches multilevel modeling inference by acknowledging greater uncertainty in estimates of the variance components, and may therefore be more appropriate for use with small samples. Using fully Bayesian estimates, fixed effects would be similar or identical to those estimated by the EB approaches of *HLM 6.0*, but standard errors would be larger. These larger standard errors more accurately represent the greater uncertainty in estimates which produce smaller significance test values than does the EB approach (Afshartous & de Leeuw, 2005; Browne & Draper, 2006; Raudenbush & Bryk, 2002).

Especially when sample sizes are small, it is important to consider which estimation method is best. These estimation methods will be compared in order to examine differences in interpreted findings. The analytic methods discussed in the above section represent the most prominent approaches proposed to date for analyzing short-phase ITS data. From the least sophisticated visual inspection approach to the most sophisticated multilevel modeling approach, each method has strengths and limitations for appropriately handling these data and for providing informative findings. Table 1 provides a summary of the discussion above, noting which methods possess which strengths in analyzing single case design data. The following section will outline the way in which the effectiveness of these methods will be demonstrated and compared.

Table 1. Evaluation of Methods – Summary Table

	Reliable	Tests Significance of Effect	Quantifies Functional Relationships	Handles Count Outcome Data	Handles Autocorrelation	Includes Attention to Trend	Handles Complex Designs	Allows for Synthesizing Across Subjects	Allows Explanation of Heterogeneity of Effects Across Subjects
Visual Analysis				✓		✓	✓	✓	
Gentile, Roden & Klein's (1972) **ANOVA Method**	✓	✓						✓	
Gorsuch's (1983) **ANCOVA-like Regression Method**	✓	✓	✓			✓	---		
Gingerich 's (1984) **ES Calculation**		✓						✓	
White, Rusch, Kazdin & Hartmann's (1989) **ES Calculation**		✓	---			---		---	
Center, Skiba & Casey's (1985) **Piecewise Regression Method**		✓	✓			✓		---	
Allison & Gorman's (1993) **Regression-based ES Calculation**		✓	✓		---	✓		---	
Busk & Serlin's (1992) **ES Calculation**		✓	✓		---	✓		---	
Scruggs, Mastropieri & Castro's (1987a) **PND Comparison**				✓			---	✓	
Edgington's (1980) **Randomization Tests**	✓	✓		✓	---				
Crosbie's (1993) **ITSACORR Time Series Method**	✓	✓	✓		✓	✓	---		
Van den Noortgate & Onghena's (2003a,b) **Multilevel Modeling Method**	✓	✓	✓	✓	✓	✓	✓	✓	✓

Methods

In the remainder of this paper, several of the methods discussed are demonstrated and their performance in analyzing and synthesizing the results of two single case reversal design (ABAB) studies are compared. In making the claim that the multilevel modeling approach is most appropriate for these analyses, each method is first reframed in relation to the multilevel model. This way, stable comparisons can be made across models and conclusions can be drawn about the relative strengths and weaknesses of each approach. As mentioned above, many of the earlier methods are based upon an underlying linear regression model. As such, these can naturally be framed as special cases of the mulitlevel model. The *HLM 6.0* program was run with modified (i.e., restricted) models that imitate each of the less sophisticated methods before it. The models for each demonstration are expressed in Table 2. Interpretations of the estimates from each of these models are compared, including a discussion of the extent to which interpretation may be limited by missing parameters (e.g., parameters that allow slope to change across phases).

Table 2. HLM Equations for Demonstrated Methods

Method	*Level-1*	*Level-2*
a) **ANOVA Method** (Gentile, Roden & Klein, 1972)	$Y_{ij} = P_{0j} + P_{1j}*(trt)$	$P_{0j} = B_{00} + R_{0j}$ $P_{1j} = B_{10} + R_{1j}$
b) **ANCOVA-like Method** (Gorsuch, 1983)	$Y_{ij} = P_{0j} + P_{1j}*(time) + P_{2j}*(trt)$	$P_{0j} = B_{00} + R_{0j}$ $P_{1j} = B_{10} + R_{1j}$ $P_{2j} = B_{20} + R_{2j}$
c) **Piecewise Regression Method** (Center, Skiba, & Casey, 1985)	$Y_{ij} = P_{0j} + P_{1j}*(time) + P_{2j}*(trt) + P_{3j}*(time*trt)$	$P_{0j} = B_{00} + R_{0j}$ $P_{1j} = B_{10} + R_{1j}$ $P_{2j} = B_{20} + R_{2j}$ $P_{3j} = B_{30} + R_{3j}$
d) **Effect Size Methods** (general)	$Y_{ij} = P_{0j} + P_{1j}*(trt)$	$P_{0j} = B_{00} + R_{0j}$ $P_{1j} = B_{10} + R_{1j}$
	$\delta = \frac{B_{10}}{\sigma}$ or $\frac{B_{10}}{\sqrt{Tau_{00}}}$	
e) **ITSACORR Method** (Crosbie, 1993)	$Y_{ij} = P_{0j} + P_{1j}*(time) + P_{2j}*(trt) + P_{3j}*(time*trt) + autocorrelation\ term$	$P_{0j} = B_{00} + R_{0j}$ $P_{1j} = B_{10} + R_{1j}$ $P_{2j} = B_{20} + R_{2j}$ $P_{3j} = B_{30} + R_{3j}$
f) **Multilevel Modeling Method** (Van den Noortgage & Onghena, 2003a, 2003b)	$Y_{ij} = P_{0j} + P_{1j}*(time) + P_{2j}*(trt) + P_{3j}*(order) + P_{4j}*(time*trt) + P_{5j}*(time*order) + P_{6j}*(trt*order) + P_{7j}*(time*trt*order)$	$P_{0j} = B_{00} + R_{0j}$ $P_{1j} = B_{10} + R_{1j}$ $P_{2j} = B_{20} + R_{2j}$ $P_{3j} = B_{30} + R_{3j}$ $P_{4j} = B_{40} + R_{4j}$ $P_{5j} = B_{50} + R_{5j}$ $P_{6j} = B_{60} + R_{6j}$ $P_{7j} = B_{70} + R_{7j}$

The statistical methods that do not fit into this framework (e.g., PND, randomization tests) were included in the discussion above for the sake of completeness (i.e., because they are discussed in the literature), but they are not demonstrated as part of this study since there is no stable way of comparing their results to the multilevel modeling method. In actuality, these methods are not appropriate anyway since they do not allow for the modeling and/or estimation of effects.

Clearly, visual analysis does not fall into a modeling framework, either. However, it is understood that visual analysis can provide for useful information to support statistical alternatives for analyzing and synthesizing single case design data. In particular, visual inspection is especially helpful in detecting outliers, nonlinearity, and delay of treatment effect before or after statistical analyses are conducted.

The demonstrations of the selected approaches were conducted on two single case reversal design (ABAB) datasets. The first dataset was simulated to provide a clear and effective illustration of the ability of each of the six illustrated methods to handle the aspects discussed above. Based on the trends observed in single case design study data and on the capacity of the HLM program, the simulated dataset was built using the specifications listed in Table 3 and the model coefficients specified by the values listed in Table 4.

Table 3. Simulated Data Specifications

Characteristic	Specification
***N* subjects** (Level-2 units)	8 subjects
***N* phases** (design)	4 phases (ABAB)
***N* observations** (Level-1 data points)	5 observations/phase (20 data points total)
Dependent Variable type	continuous (0-100)

Table 4. Simulated Data Model Coefficients

Model Coefficients	Range
B_{00}: Baseline Intercept (Level)	70 to 90
B_{10}: Baseline Slope (time)	-2 *(fixed)*
B_{20}: Treatment Effect: Level (trt)	-15 to -35
B_{30}: Order Effect: Level (order)	4 to 6
B_{40}: Treatment Effect: Slope (time*trt)	0 *(fixed)*
B_{50}: Order Effect: Slope (time*order)	0 *(fixed)*
B_{60}: Order Effect: Treatment Effect (trt*order)	-5 to -15
B_{70}: Order Effect on Treatment Effect: Slope (time*trt*order)	0 *(fixed)*

Additionally, in order to allow for testing of the effects of autocorrelation on the estimates, a moderate autocorrelation (0.25) was intentionally included in these data and allowed to vary.

The second dataset used for illustration was taken from a published study, so that the demonstrated methods' ability to handle the complexities of real-world data could be compared. This dataset, from Taylor & Alber's (2003) study of the effectiveness of a peer-mediated instruction program on spelling test scores, included observations of four students over 26 weeks, in an ABAB design. Since raw data were not readily available for this dataset, published graphs were digitally scanned and raw data were extracted using the UnGraph© software.

The performance of each method was assessed for its ability to express the functional relationships between independent variables and outcome measures (e.g., accounting for trend, quantifying effect); and to explain any differences between individuals studied (i.e., heterogeneity of effect). In other words, the ability of each method to optimally identify and express overall effects, determinants of effects, and generalizability of results was compared. The ability of each to appropriately attend to the technical complexities of single case design

research data (e.g., autocorrelation, discrete outcome variables) was also examined. It was expected that a combination of visual analysis and multilevel modeling would most suitably handle all aspects of the data, optimize the richness of information available in the data, and provide the most useful results.

Once this was established, an exploration of estimation methods was also conducted. In particular, two approaches to estimating multilevel model parameters – one of which is not designed for small sample data but is implemented in the most commonly used program (specifically, *HLM 6.0*) and the other which is more complex and less often used but is more appropriate for small samples (used in the *WinBUGs* program) – were compared. Multilevel models were run in both *HLM 6.0*, utilizing EB estimation, and in *WinBUGS*, using fully-Bayesian estimation. Parameters estimated by each program are compared for differences in possible interpretation.

The resulting analyses provide for a thoughtful and stable comparison of several recommended methods for analyzing single case design data; providing evidence for the methodological selections made for future research.

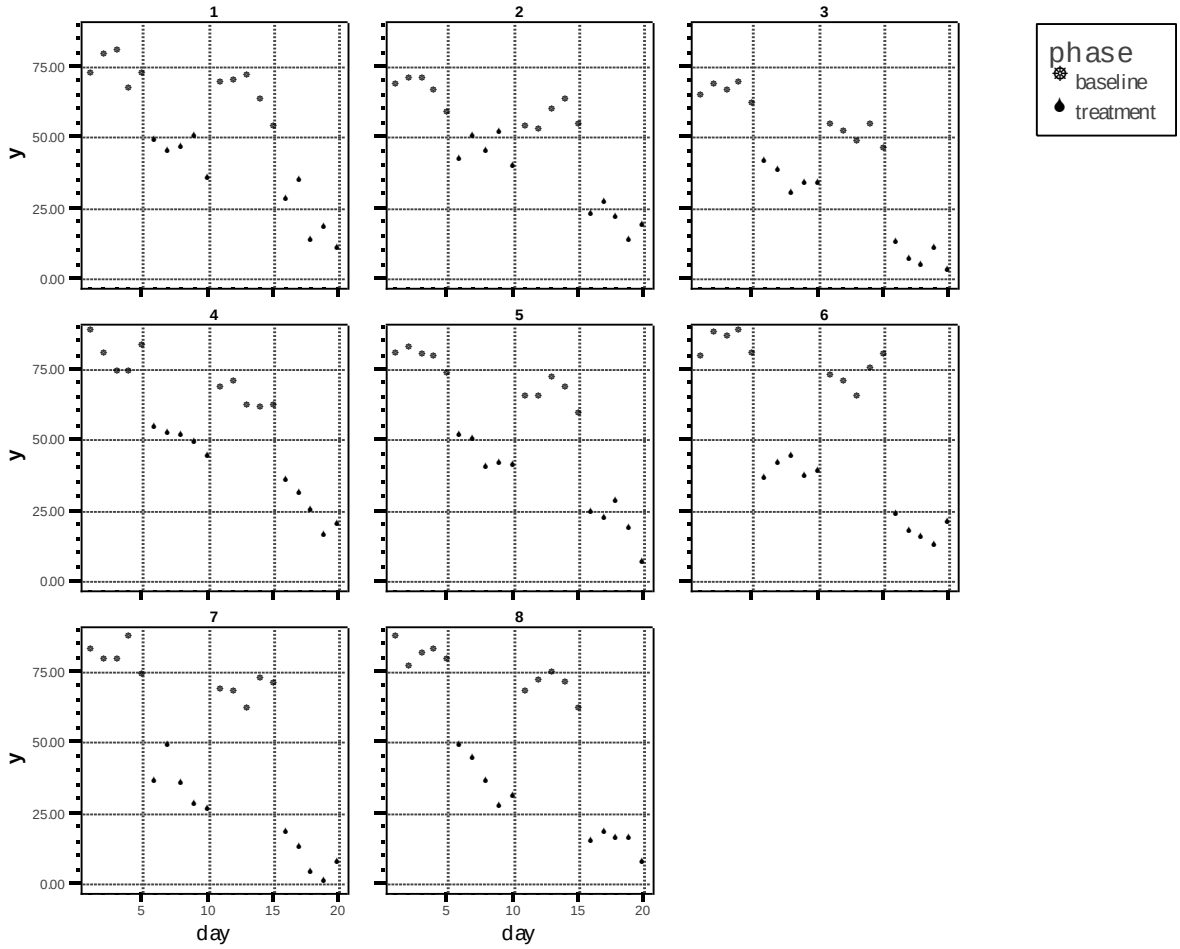
Results

Demonstration 1: Simulated Data

For the first set of analyses, a dataset was simulated to provide a clear and effective illustration of the each method's ability to handle the aspects discussed above. Two SPSS data files were constructed for use in the *HLM 6.0* program. Using the specifications presented in Tables 3 and 4, a level-1 dataset was created containing data on the repeated observations of eight students' scores on some simulated test.[1] It is presumed that lower values on this outcome measure are more desirable than higher values. Variables in the level-1 data file included: subject identifier; time (days 1-20); phase identifier (baseline or treatment); phase pair order (first AB pair or second AB pair); and all possible two- and three-way interactions between the time, phase, and order variables. Variables were coded and centered for improved model interpretation as needed (e.g., time was centered at 5 so that the final day of the first baseline phase was the 0 point). Figure 2 displays the level-1 data graphed as it might have appeared in a published study.

[1] Syntax for building this dataset is available in the appendix.

Figure 2. Simulated Data: Graphed by Subject



A set of level-2 characteristics was also created so that heterogeneity of effect could be explored by subject characteristic in the multilevel model illustration. The level-2 dataset contained simulated background data on each of the eight subjects, including two variables – one to be entered on the intercept and another to be entered on the phase effect – as presented in Table 5. These variables were also centered, at the rounded grand mean, for improved interpretation. In *HLM 6.0*, an MDM file was built using the level-1 and -2 files described above to allow for model building and exploration.

Table 5. Simulated Data: Subject Characteristics (L2 Variables)

Subject ID	L2_Intercept		L2_Phase	
	raw	mean centered *(at 3)*	raw	mean centered *(at 4)*
1	3	0	6	2
2	1	-2	7	3
3	0	-3	5	1
4	3	0	5	1
5	4	1	3	-1
6	6	3	1	-3
7	4	1	2	-2
8	5	2	2	-2

In analyzing these data, the graphs in Figure 2 were first examined visually for apparent patterns, similarities and differences between phases and subjects. Visual inspection of these graphs reveals an obvious downward trend in the data, for all subjects. As observations progress from day 1 to day 20, values generally decrease. It is also apparent that, for most subjects, this negative slope is more severe during treatment phases than during baseline phases. Also, for most subjects, a treatment effect on level is visible; that is, an evident shift in mean level between baseline and treatment phases. Finally, it seems as if that shift may vary across subjects. In fact, it could be determined that both starting points (i.e., intercepts) and treatment effects may vary across subjects. Baseline levels look higher from some subjects and lower for others and baseline-treatment shifts look larger for some subjects and smaller for others.

As proposed, these data were examined under several suggested methods for the statistical analysis of single case design data. These competing models were framed as special cases of the multilevel model and estimated using the *HLM 6.0* program, restricting terms as fit to duplicate original published methods, for consistent comparison across models. It is important to emphasize that using hierarchical models in this way is not exactly the same as duplicating these alternative methods. *HLM 6.0* accommodates technicalities of the data not assumed by all

original models (e.g., calculating the statistical significance of between-subjects variance). Also, many of these methods were originally developed for N=1 studies, not to synthesize results across multiple subjects. Table 6 displays the equations and estimates (fixed effects and variance components) for each model illustrated. Table 7 shows translations of the estimates for clearer comparison across models.

Table 6. Simulated Data: HLM Estimates by Simulated Model

Method	a) **ANOVA Method**	b) **ANCOVA-like Method**	c) **Piecewise Regression Method**	d) **Effect Size Methods**	e) **ITSACORR Method**	f) **Multilevel Modeling Method**
Level 1 Equations	$Y_{ij} = P_{0j} + P_{1j}*(trt)$	$Y_{ij} = P_{0j} + P_{1j}*(time) + P_{2j}*(trt)$	$Y_{ij} = P_{0j} + P_{1j}*(time) + P_{2j}*(trt) + P_{3j}*(time*trt)$	$Y_{ij} = P_{0j} + P_{1j}*(trt)$ --- $\delta = \frac{B_{10}}{\sigma}$ or $\frac{B_{10}}{\sqrt{\tau_{00}}}$	$Y_{ij} = P_{0j} + P_{1j}*(time) + P_{2j}*(trt) + P_{3j}*(time*trt) + autocorrelation\ term$	$Y_{ij} = P_{0j} + P_{1j}*(time) + P_{2j}*(trt) + P_{3j}*(order) + P_{4j}*(time*trt) + P_{5j}*(time*order) + P_{6j}*(trt*order) + P_{7j}*(time*trt*order)$ ***Simplified to:*** $Y_{ij} = P_{0j} + P_{1j}*(time) + P_{2j}*(trt) + P_{3j}*(time*trt)$
Level 2 Equations	$P_{0j} = B_{00} + R_{0j}$ $P_{1j} = B_{10} + R_{1j}$	$P_{0j} = B_{00} + R_{0j}$ $P_{1j} = B_{10} + R_{1j}$ $P_{2j} = B_{20} + R_{2j}$	$P_{0j} = B_{00} + R_{0j}$ $P_{1j} = B_{10} + R_{1j}$ $P_{2j} = B_{20} + R_{2j}$ $P_{3j} = B_{30} + R_{3j}$	$P_{0j} = B_{00} + R_{0j}$ $P_{1j} = B_{10} + R_{1j}$	$P_{0j} = B_{00} + R_{0j}$ $P_{1j} = B_{10}$ $P_{2j} = B_{20} + R_{2j}$ $P_{3j} = B_{30}$	$P_{0j} = B_{00} + R_{0j}$ $P_{4j} = B_{40} + R_{4j}$ $P_{1j} = B_{10} + R_{1j}$ $P_{5j} = B_{50} + R_{5j}$ $P_{2j} = B_{20} + R_{2j}$ $P_{6j} = B_{60} + R_{6j}$ $P_{3j} = B_{30} + R_{3j}$ $P_{7j} = B_{70} + R_{7j}$ ***Simplified to:*** $P_{0j} = B_{00} + B_{01} + R_{0j}$ $P_{1j} = B_{10} + R_{1j}$ $P_{2j} = B_{20} + B_{21} + R_{2j}$ $P_{3j} = B_{30} + R_{3j}$
Basline Level	B00 = 69.570**	B00 = 75.090**	B00 = 73.182**	B00 = 69.570**	B00 = 73.182**	B00 = 72.307** ; B01 = 3.498**
Baseline Slope		B10 = - 1.840**	B10 = - 1.204**		B10 = - 1.204**	B10 = - 1.204**
Effect of Trt on Level	B10 = - 40.919**	B20 = - 31.720**	B20 = - 24.725**	B10 = - 40.919**	B20 = - 24.725**	B20 = - 24.252* ; B21 = 3.783**
Effect of Trt on Slope			B30 = - 1.272**		B30 = - 1.272**	B30 = - 1.272**
SD: Baseline Level	$\tau_{00}^{1/2}$ = 5.813**	$\tau_{00}^{1/2}$ = 6.410**	$\tau_{00}^{1/2}$ = 6.586**	$\tau_{00}^{1/2}$ = 5.813**		$\tau_{00}^{1/2}$ = 1.291
SD: Baseline Slope		$\tau_{11}^{1/2}$ = 0.076	$\tau_{11}^{1/2}$ = 0.051			$\tau_{11}^{1/2}$ = 0.138
SD: Effect of Trt on Level	$\tau_{11}^{1/2}$ = 6.800**	$\tau_{22}^{1/2}$ = 8.381**	$\tau_{22}^{1/2}$ = 8.583**	$\tau_{11}^{1/2}$ = 6.800**		$\tau_{22}^{1/2}$ = 2.147
SD: Effect of Trt on Slope			$\tau_{33}^{1/2}$ = 0.090			$\tau_{33}^{1/2}$ = 0.133
SD: Within-Subjects (σ)	σ = 11.603	σ = 5.760	σ = 4.587	σ = 11.603		σ = 4.539
				$\frac{B_{10}}{\sigma} = \frac{-40.919}{11.603}$ ***d* = - 3.527** ***or*** $\frac{B_{10}}{\sqrt{\tau_{00}}} = \frac{-40.919}{5.813}$ ***d* = - 7.039**	**rho** = 0.127 *(SE = 0.092)*	

*p<.05, **p<.01

Table 7. Simulated Data: Translated Estimates

Method	a) ANOVA Method	b) ANCOVA-like Method	c) Piecewise Regression Method	d) Effect Size Methods	e) ITSACORR Method	f) Multilevel Modeling Method
Baseline Phase Level	69.570	75.090	73.182	d = -3.527 or -7.039	73.182	72.307 + 3.498*B01
Baseline Phase Slope		- 1.840	- 1.204		- 1.204	-1.204
Treatment Phase Level	28.651	43.370	48.457		48.457	48.055 + 3.783*B21
Treatment Phase Slope		- 1.840	- 2.476		- 2.476	-2.476

Model A – Analysis of Variance (ANOVA)

Gentile, Roden, and Klein (1972) suggested the use of a two-way ANOVA model to analyze single case design data from N>1 datasets. This model includes terms to test for the significance of phase differences, subject differences, and heterogeneity of phase effects across subjects (i.e., phase-by-subject interaction). In order to duplicate this method in HLM, a multilevel model was built as follows:

$$\text{Level-1: } Y_{ij} = P_{0j} + P_{1j}*(trt) + e_{ij} \quad (10)$$

$$\text{Level-2: } P_{0j} = B_{00} + R_{0j}$$
$$P_{1j} = B_{10} + R_{1j}$$

The original authors of this method did not propose a way to quantify the size of these effects, only to determine their statistical significance via F-tests. By running this model in *HLM 6.0*, estimates of magnitude of effect are produced. Thus, this simulation offers more information than the method it is attempting to demonstrate.

According to the estimates produced for this model, on average, subjects scored at or around 70 in the baseline phase and 29 in the treatment phase. However, between-subjects variance estimates (shown in Table 6) indicate that using this limited model to describe these data leaves behind a substantial and statistically significant amount of between-person variation unexplained ($\sqrt{\tau_{00}}$ =5.813, $\sqrt{\tau_{11}}$ =6.800). That is, simply estimating mean levels in each phase

does not well enough explain the patterns evident in these data. Visual inspection suggested that including a term for the effect of time (i.e., slope for days) may better suit these data. The next model does just that.

Model B – Analysis of Covariance (ANCOVA)

Adding on to Gentile, Roden, & Klien's (1972) ANOVA model, Gorsuch (1983) suggested the inclusion of a slope (time effect) term in his ANCOVA-like model for the analysis of single case design data. According to Gorsuch (1983), this model tests for mean differences between phases while controlling for overall trend in the data. In order to duplicate this method in *HLM 6.0*, a model was built as follows:

$$\text{Level-1: } Y_{ij} = P_{0j} + P_{1j}*(time) + P_{2j}*(trt) + e_{ij} \quad (11)$$

$$\text{Level-2: } P_{0j} = B_{00} + R_{0j}$$
$$P_{1j} = B_{10} + R_{1j}$$
$$P_{2j} = B_{20} + R_{2j}$$

Estimates produced by this model indicate a statistically significant negative trend in the data (on average, about 1.8 points per day or 9.2 points over the course of a 5-day phase). The ending baseline mean level (due to centering the time variable at the end of the first baseline phase) is estimated by this model to be 75 points and the treatment phase mean level is estimated at 43 (a phase effect of 32 points). Between-subjects variance estimates indicate a statistically significant and substantial amount of variation remains in the baseline and phase effect estimates ($\sqrt{\tau_{00}}$ =6.410, $\sqrt{\tau_{22}}$ =8.381) while no substantial variance remains on the trend (or time slope) estimate ($\sqrt{\tau_{11}}$ =0.076). The determination that there is negative trend of almost 2 points per day present in the data seems to hold for all subjects. The significant coefficient for trend in the data also supports the argument that the ANOVA model above (A) was not able to sufficiently describe data patterns, since it did not attend to slope in the data in any way. The baseline and

phase mean level estimates, on the other hand, may still vary across subjects. It is possible that adding additional terms to the model – to allow slope to vary by phase, perhaps – could again improve the model's fit to these data.

Model C – Piecewise Regression

Building from Gorsuch's (1983) work, Center, Skiba and Casey (1985) proposed a piecewise regression technique that involves the calculation of three separate effect size estimates for changes in level and slope. This proposed regression equation (equation 4 above) includes a term for baseline level, a term for change in level from baseline to treatment, a term for baseline trend (slope) and an interaction term to measure the change in slope due to treatment. In order to duplicate this method in *HLM 6.0*, a model was built as follows:

$$\text{Level-1: } Y_{ij} = P_{0j} + P_{1j}*(time) + P_{2j}*(trt) + P_{3j}*(time*trt) + e_{ij} \quad (12)$$

$$\text{Level-2: } P_{0j} = B_{00} + R_{0j}$$
$$P_{1j} = B_{10} + R_{1j}$$
$$P_{2j} = B_{20} + R_{2j}$$
$$P_{3j} = B_{30} + R_{3j}$$

Estimates produced by this model indicate mean ending baseline levels of 73 and treatment phase levels of 48, on average. Estimates of trend in both baseline and treatment phases indicate that significant negative slope and a significant change in slope across phases exists in these data. Scores are estimated to drop, on average, 1.2 points per day in the baseline phase (or 6 points across a 5-day baseline phase) and 2.5 points per day in the treatment phase (or 12.5 points across a 5-day treatment phase). Even with this more complex model, between-subjects variance estimates indicate significant between-person variation remains on the baseline and phase effect estimates ($\sqrt{\tau_{00}}$ =6.586, $\sqrt{\tau_{22}}$ =8.583), while trend effects seem to be sufficiently explained for all subjects.

Estimates provided by this model indicate that not only is a significant trend present in the data, but that this trend is significantly impacted by the intervention and changes in degree from baseline to treatment phase; something that was detected in the visual inspection described above. The ANCOVA model (B) could not capture non-constant trend in the data; it could only impose a constant trend throughout the data collection period. This Piecewise Regression model allows trend to be modeled differently for each phase. Estimates indicate that the slope does indeed change significantly between phases. However, even this more complex model was not suitable for these data, as substantial and statistically significant between-subject variability still remains on estimates of mean baseline and treatment phase levels. Other models must still be considered.

Model D – Effect Sizes

Several effect size methods for the analysis of single case design data were described. Overall, these methods suggest that treatment effects can be assessed by dividing estimated phase effects (i.e., baseline-treatment phase difference) by some expression of between- or within-subjects standard deviation estimates. In order to examine these types of assessments using *HLM 6.0*, the phase effect computed for the ANOVA model (Model A) was considered as the numerator. This estimate ($B_{10=}$-40.92) was then divided by two different estimates of standard deviation. In one model, the within-subjects standard deviation estimate from the ANOVA model (σ=11.603) was used as the denominator:

$$\delta = \frac{B_{10}}{\sigma} = \frac{-40.92}{11.603} \qquad (13)$$

This effect size, using within-subjects standard deviation as the denominator, is calculated to be -3.527.

Alternatively, an effect size was also computed using the between-subjects (person-to-person) standard deviation from the phase effect estimate ($\sqrt{\tau_{oo}}$ =5.813) as the denominator:

$$\delta = \frac{B_{10}}{\sqrt{\tau_{00}}} = \frac{-40.92}{5.813} \quad (14)$$

This estimate, using between-subjects standard deviation as the denominator, is calculated as -7.039.

In the discussion of this group of methods above, it was mentioned that effect size methods like these can be misleading because they produce values that are not comparable to those from group design research. In particular, within-subjects estimates of variation (e.g., σ, $\sqrt{\tau_{oo}}$) are often much smaller than those from between-subjects studies. In this case, the estimate of within-subjects variation (here, the standard deviation) is actually larger than that of between-subjects variation. Thus, the effect size estimate using the within-subjects standard deviation estimate as the divisor is actually smaller than the one using between-subjects variation. However, effect size estimates using either standard deviation are far out of range for typical group design estimates. While group design effect sizes usually range from about -1 to +1, single case design effect sizes can often be much larger, and both of these estimates (-3.527 and -7.039) are out of this range.

As well, these one-number estimates of effect are superficial descriptors of what impact the treatment had on observed behavior. These estimates cannot answer key questions like: How did the intervention affect behavior over time? Where did observed subjects start out? Where did they end up? How did they get there? Time series data uniquely allows for exploration of these questions. The effect size estimates do not tell us anything about data patterns, a missed opportunity to take advantage of the richness of single case design data.

Model E – ITSACORR (Time Series)

In contrast to the superficial effect size method demonstrated above (D), time series methods allow for exploration of intricate patterns in single case design data. In general, time series methods combine several techniques in a regression-like approach to include the influence of previous observations and of accumulated residual error in the prediction and modeling of repeated observed behaviors. As discussed, time series methods often require a large number of data points in order to achieve reliable estimates. However, Crosbie (1993) proposed a time series model known as ITSACORR to be applied to data from single case studies with series as short as 10 observations. This researcher also developed a computer program by the same name (ITSACORR) to perform all computations and decision making about significance levels. However, this program is no longer available[2]. It should be noted that Crosbie originally proposed this method as a way to analyze data from a single case (N=1) study, not to synthesize effects across several cases (e.g., N=8). Thus, the original ITSACORR software (which would have to be run once for each subject) would not produce data in a form comparable to the other methods.

In following with the plan for direct comparisons using *HLM 6.0* program estimation, an alternative time series model was built and run using the hierarchical multivariate linear model (HMLM) option in *HLM 6.0* which allows for a model with "first-order auto-regressive level-1 random errors and random intercepts and/or slopes at level 2" (Raudenbush et al, 2004, p. 141). In order to duplicate the time series method in HMLM, a model was built as follows:

$$\text{Level-1: } Y_{ij} = P_{0j} + P_{1j}*(time) + P_{2j}*(trt) + P_{3j}*(time*trt) + e_{ij} \quad (15)$$

$$\text{Level-2: } P_{0j} = B_{00} + R_{0j}$$
$$P_{1j} = B_{10} + R_{1j}$$
$$P_{2j} = B_{20} + R_{2j}$$

[2] In fact, Crosbie "officially retired" ITSACORR in 2006 (Crosbie, 2006).

$$P_{3j} = B_{30} + R_{3j}$$

This model as specified would not run (i.e., estimates could not be calculated). In order to solve this problem, random effects on the two time slope terms had to be constrained to zero. (Note that these estimates had not been determined to vary across subjects in any of the models run previously.) The modified model is presented in equation 16.

$$\text{Level-1: } Y_{ij} = P_{0j} + P_{1j}*(time) + P_{2j}*(trt) + P_{3j}*(time*trt) + e_{ij} \quad (16)$$

$$\text{Level-2: } P_{0j} = B_{00} + R_{0j}$$
$$P_{1j} = B_{10}$$
$$P_{2j} = B_{20} + R_{2j}$$
$$P_{3j} = B_{30}$$

According to the estimates produced for this model by HMLM, although autocorrelation was set at a moderate level (0.25) during data simulation, the actual autocorrelation in the simulated data is extremely small (0.127). This amount of autocorrelation was found to be statistically non-significant (SE = 0.092) and a model comparison presented in the output verifies no significant difference between the model with the autoregressive term and one without an autoregressive term (Chi-square = 1.965, p=.157). The comparison showed no substantial difference in the fixed effects estimated by each model either. Table 8 presents the estimates produced by each model as well as model comparison statistics. It can be concluded that no real autocorrelation exists in these data and that the model without the autoregressive term is a sufficient and more parsimonious model of these data. Had we not tested for the presence of autocorrelation here, we would not have known that it was unnecessary to accommodate this complexity of small sample data.

Table 8. Simulated Data: Checking Autocorrelation (ITSACORR)

	No Autocorrelation (No AR)	**Autocorrelation (AR)**
Baseline Level*: Estimate (p-value)*	B00 = 73.182 *(p<.001**)*	B00 = 73.153 *(p<.001**)*
Baseline (Time) Slope*: Estimate (p-value)*	B10 = -1.204 *(p<.001**)*	B10 = -1.209 *(p<.001**)*
Effect of Trt on Level*: Estimate (p-value)*	B20 = -24.725 *(p<.001**)*	B20 = -24.694 *(p<.001**)*
Effect of Trt on Time Slope*: Estimate (p-value)*	B30 = -1.272 *(p<.001**)*	B30 = -1.263 *(p<.001**)*
Rho*: Estimate (SE)*		*rho* = 0.127 *(0.092)*
Sigma-Squared*: Estimate (SE)*	σ^2 = 20.959 *(2.470)*	σ^2 = 21.495 *(2.675)*
Deviance*: Estimate*	*985.991*	*984.026*
Model Comparison	Chi-square = 1.965 ***(p=.157)***	
(Translated) Baseline Phase Level	**73.182**	**73.153**
(Translated) Baseline Phase Slope	**- 1.204**	**- 1.209**
(Translated) Treatment Phase Level	**48.457**	**48.459**
(Translated) Treatment Phase Slope	**- 2.476**	**- 2.472**

Using estimates from the model without the autoregression term, this method estimates the average level on day 5 (last day of baseline phase) to be about 73 points and treatment phase means to be about 48 points on average. Scores are estimated to fall 1.2 points per day in the baseline phase and a stronger 2.5 point loss per day in treatment phases. Estimates of between-subjects variance (i.e., random effects) were not produced by this model. However, given evidence in the above illustrations that between-subjects variation does exist, in particular on estimates of baseline level and phase effect, additional analyses should be run. A two-level model, allowing for subject characteristics to explain variation in fixed effects could be run using *HLM 6.0*.

Model F – Multilevel Modeling Method

As has been demonstrated by the approximate duplications of other methods via *HLM 6.0*, the true multilevel modeling method distinguishes levels in the data structure (e.g., observations, individuals) by formally expressing each level with its own statistical model: observations on individuals (level-1) with predictors at the observation level (e.g., time and

phase) and at the individual or subject level (level-2) (i.e., simulated subject characteristics described above).

The multilevel modeling approach starts out by entering all available observation-level terms into the level-1 model and then parsing out non-significant variables until only significant contributors to the model are left. For these data, a full level-1 model was first run, as specified in Table 6, including all possible main effects, and two- and three-way interaction effects. The term with the p-value furthest from significance (i.e., highest) was dropped, following protocol for keeping main effects in the model when interactions including those terms still remain. One by one, terms were dropped until a model that included only significantly contributing level-1 terms remained. This level-1 model resembled the one run for the Piecewise Regression illustration (Model *C*) as only the intercept, time effect (slope), treatment effect, and time-by-treatment interaction terms remained. Running this model, expressed by the equations below, produced output that supported earlier evidence that subjects vary on intercepts (i.e., starting levels) and phase or treatment effect. To explore whether these level-1 differences could be accounted for, subject characteristics, simulated for this purpose, were included in the level-2 model.

$$\text{Level-1: } Y_{ij} = P_{0j} + P_{1j}*(time) + P_{2j}*(trt) + P_{4j}*(time*trt) + e_i \qquad (17)$$

$$\text{Level-2: } P_{0j} = B_{00} + B_{01} + R_{0j}$$
$$P_{1j} = B_{10} + R_{1j}$$
$$P_{2j} = B_{20} + B_{21} + R_{2j}$$
$$P_{3j} = B_{30} + R_{3j}$$

As shown in Tables 6 and 7, this model estimates that, on average, subjects end the baseline phase at about 72 points, losing about 1.2 points per day in that phase. However, subjects who have higher values on the simulated level-2 variable for the intercept are expected to have higher baseline levels by about 3.5 points for every point above average on the level-2

characteristic (or 3.5 points lower for every point below average). This model also estimates that shifts in level from baseline to treatment phase can be explained by subject characteristics. Scores for subjects who are average on the level-2 characteristic simulated for these data fall about 24 points as they move from baseline phase to treatment phase. Subjects who are higher than average on the level-2 characteristic decrease less, almost 4 points less in baseline-treatment shift for every one unit above average they fall on the level-2 variable. Slopes during the treatment phase are estimated at about -2.5 points per day for all subjects.

Once these level-2 characteristics were added to the model, all estimates of between-subjects variance become non-significant. That is, all estimated fixed effects are found to fit sufficiently for all subjects and as a whole, sufficiently explain patterns in the data. By adding subject characteristics to the model, all remaining variance between subjects on the estimate of intercept and treatment effect dissipated, something none of the previous approaches could achieve.

Estimation Methods: Maximum Likelihood Empirical Bayes vs. Fully Bayesian

As proposed,, an exploration of estimation methods was also conducted. The estimation approach utilized by *HLM 6.0* combines maximum likelihood and empirical Bayes estimation (ML-EB) to make conditional estimates of fixed effects based on estimates of the random effects. As discussed above, when sample size is small, as it is here, there may not be sufficient information to produce stable estimates of effects with the certainty implied by the ML- EB approach. Fully Bayesian estimation, on the other hand, used by *WinBUGS*, approaches multilevel modeling inference by acknowledging greater uncertainty in estimates of the variance components and may therefore be more appropriate for use with small samples.

Multilevel models were run both in *HLM 6.0*, utilizing ML-EB estimation, and in *WinBUGS*, using fully-Bayesian estimation.[3] Parameters estimated by each program are presented in Table 9.

Table 9. Simulated Data: Comparing Estimation Methods (ML-EB vs. Fully Bayesian)

	Maximum Likelihood Empirical Bayes (ML-EB) Estimates	**Fully Bayesian Estimates**
	Estimate *(SE)*	Estimate *(SE)*
Baseline Level	B00 = 72.307 *(0.747)* B01 = 3.498 *(0.310)*	mu0 = 72.490 *(0.848)* mu1 = 3.361 *(0.471)*
Baseline (Time) Slope	B10 = - 1.204 *(0.109)*	mu2 = - 1.206 *(0.120)*
Effect of Treatment on Level	B20 = - 24.252 *(1.339)* B21 = 3.783 *(0.553)*	mu3 = - 23.710 *(1.709)* mu4 = 3.771 *(0.768)*
Effect of Treatment on Time Slope	B30 = -1.272 *(0.146)*	mu5 = -1.275 *(0.161)*
SD: Baseline Level	sd(u0) = 1.291	sd0 = 0.669 / sd1= 0.340
SD: Baseline Slope	sd(u1) = 0.138	sd2 = 0.130
SD: Effect of Treatment on Level	sd(u2) = 2.147	sd3 = 1.515 / sd4 = 0.661
SD: Effect of Treatment on Time Slope	sd(u3) = 0.133	sd5 = 0.171
SD: Within-Subjects	σ = 4.539	sd(y) = 4.628
(Translated) Baseline Phase Level	**72.307 + 3.498*B01**	**72.490 + 3.361*B01**
(Translated) Baseline Phase Slope	**-1.204**	**-1.206**
(Translated) Treatment Phase Level	**48.055 + 3.783*B21**	**48.780 + 3.771*B21**
(Translated) Treatment Phase Slope	**-2.476**	**-2.481**

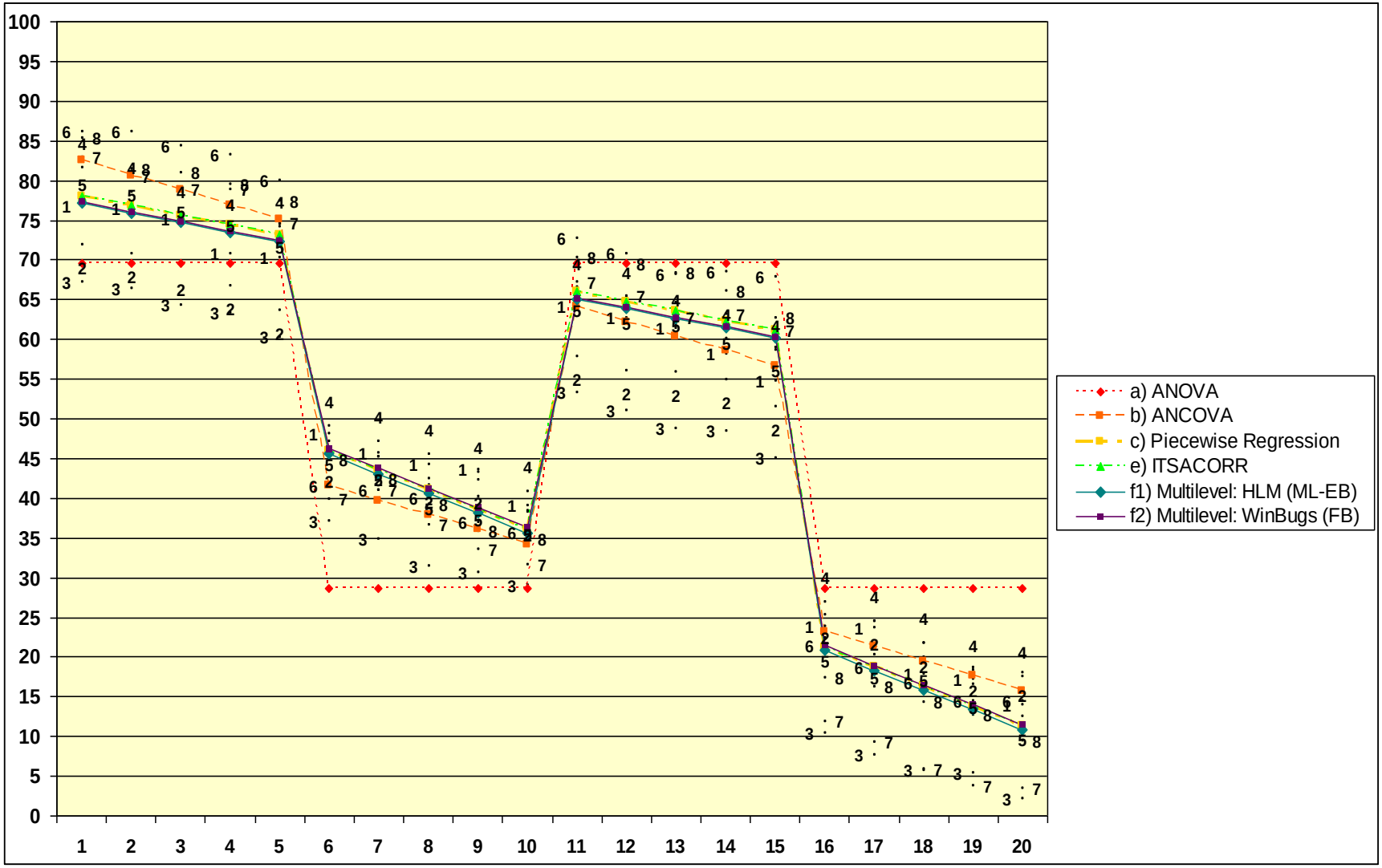
The model run using *WinBUGS*, using fully Bayesian estimation methods, calculates an average ending baseline level of about 72 and an average treatment phase level of 49, with a loss of 1.2 points per day in the baseline phase and 2.5 points per day in the treatment phase. These estimates of fixed effects are very similar to those calculated by *HLM 6.0* using the ML-EB estimation approach. The standard errors of the estimates, however, are slightly larger as estimated by the *WinBUGS* program than by *HLM 6.0* because the fully Bayesian approach acknowledges greater uncertainty in estimation of effects. This acknowledgement is more

[3] *WinBUGS* code for this analysis is available in the appendix.

realistic about what is unknown, which is particularly appropriate with small sample data like these. This fully Bayesian estimated model is the most appropriate for these data.

All estimated models are plotted over the original simulated data in Figure 3.

Figure 3. Simulated Data: Model Estimates and Raw Data



Demonstration 2: Published Data

In addition to the demonstration analyzing data from a simulated data set, data from a published study were also analyzed. This second demonstration allows for application of these methods to real data collected in real school environments. As discussed above, researchers often carry out short-phase ITS design studies in real classroom environments to measure the impact of particular programs on students' behaviors and/or skills. Taylor and Alber (2003) conducted such a study to test the effectiveness of a peer-mediated instruction program – Classwide Peer Tutoring (CWPT) – on the spelling skills of first graders over 26 weeks. Study participants were four first graders (two male, two female) with unspecified learning disabilities. Other reported background characteristics included age (ranging from six to seven years), baseline word reading scores (ranging from 10 to 50), and number of months receiving special education services before the start of data collection. During baseline phases (weeks 1-5 and 18-21), the teacher introduced new words to the class on the Monday of each week; students completed practice activities on Tuesday, Wednesday, and Thursday; and a 10-word spelling dictation test was administered on each Friday. During the treatment (i.e., CWPT) condition (weeks 6-17 and 22-26), students learned and practiced new words in tutoring pairs throughout the week with the teacher available for assistance as needed before the spelling test on Friday.

Graphs of outcome data (number correct out of 10 total words) by subject are presented in Figure 4 as they were published in the original study. In order to determine whether CWPT was more effective at improving students' spelling skills than was non-CWPT teaching, the study's authors visually examined the graph and computed simple mean scores for each student in each phase. Taylor and Alber (2003) declared that because treatment phase means were greater than baseline phase means, that the CWPT treatment should be considered effective.

Figure 4. Taylor & Alber (2003): Published Graphs

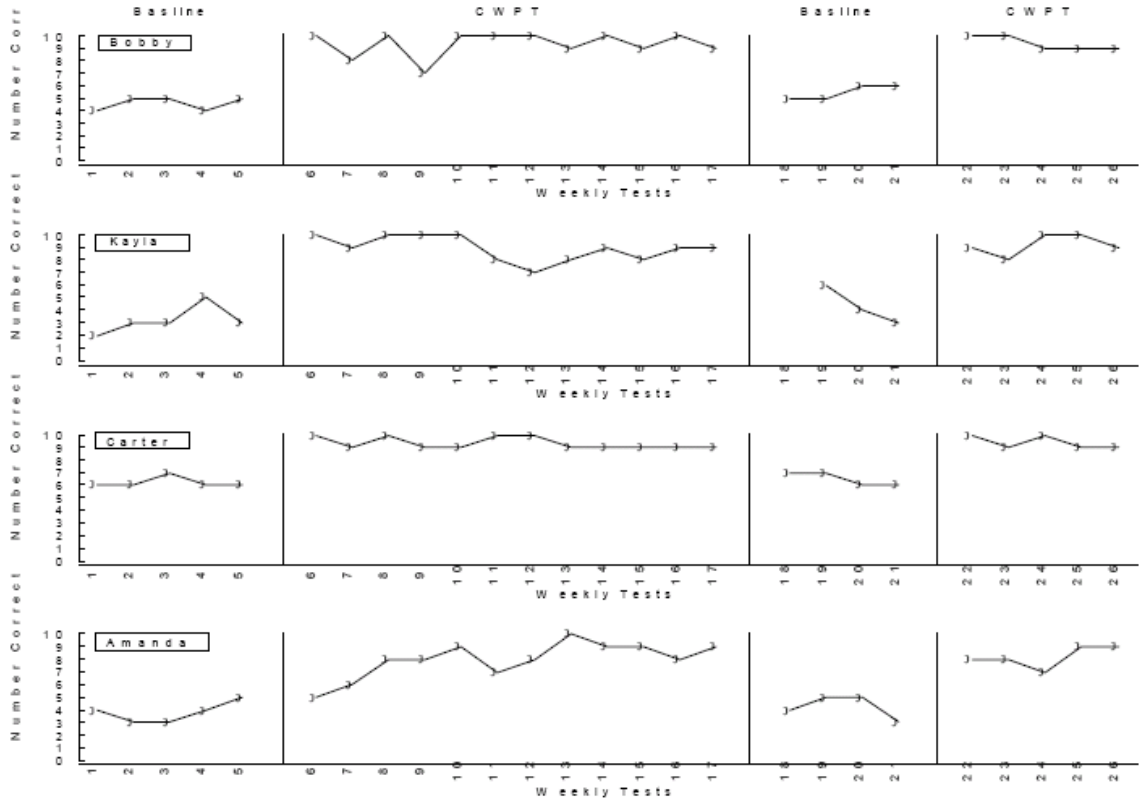


Figure 1. Number of words spelled correctly on weekly tests by Bobby, Kayla, Carter, and Amanda in the baseline and CWPT conditions.

Though visual inspection and simple mean calculation are not invalid methods of examining outcome data, they should be viewed as merely preliminary approaches for determining treatment effect. Researchers should also conduct more in-depth statistical analysis to examine whether within- and between-participant differences are large enough to be considered statistically significant and whether these differences can be attributed to other existing conditions or characteristics besides the treatment program under study.

These data were re-examined using several suggested methods for the statistical analysis of single case design data. These competing models were framed as special cases of the multilevel model and estimated using the *HLM 6.0* program for comparison across models.

In order to make data available for analysis, the published graphs (shown in Figure 4) were first digitally scanned (via flatbed scanner), saved as a *.bmp* file, and imported into the

UnGraph© *5* program. In *UnGraph*© *5*, the coordinate system was defined for each subject's graphed data and individual data points were extracted to SPSS files for further cleaning and labeling. Individual subject data files were merged together to create two SPSS data files for use in the *HLM 6.0* program. The level-1 file contained data on the repeated observations of each student's spelling test scores (i.e., the cleaned UnGraphed data). File preparation entailed rounding scanned test scores to whole numbers and spot checking approximately 50% of all scanned data points against the published figures; adding indicator variables (e.g., for subject, week number, phase type, pair order); and computing interaction terms by multiplying the terms involved in each interaction. Variables in this level-1 file included: subject identifier (1-4); number of items correct (0-10); total number of test items (or "trials") (10); week number (1-26); phase identifier (baseline or treatment; 0, 1); phase pair order (first AB pair or second AB pair; 0, 1); and all possible two- and three-way interactions between the week, phase, and order variables. Variables were also centered for improved model interpretation as needed (e.g., week was centered so that the final week of the first baseline phase was the 0 point). The level-2 file (utilized only for the multilevel model illustration) contained background data on the four students studied, including all known subject characteristics (i.e., gender, age, pre-test scores, number of months receiving special education services). These data were gathered directly from the published study. These variables were also centered for improved interpretation (i.e., so that mean pre-test scores were expressed as 0). In *HLM 6.0*, an MDM file was built with the level-1 and 2 files described above.

Models were run to simulate each of the competing methods proposed for comparison, restricting terms as needed to duplicate original published methods. Models A through E were run using the raw count of items correct as the dependent variable in the model, as these

proposed methods would do. Model F, the only true multilevel modeling method, was run using the Binomial distribution which is more appropriate for modeling proportions (in this case, number of items correct out of 10 total items). This model makes estimates on a log odds scale that can be translated back to counts (as illustrated below). Table 10 displays the equations and estimates (fixed effects and variance components) for each model. Table 11 shows translations of the estimates back to raw counts for clearer comparison across models.

Table 10. Taylor & Alber (2003): HLM Estimates by Simulated Method

Method	a) ANOVA Method	b) ANCOVA-like Method	c) Piecewise Regression Method	d) Effect Size Methods	e) ITSACORR Method	f) Multilevel Modeling Method *(Run as Binomial with 10 trials; with and without Overdispersion)*	
Level 1 Equations	$Y_{ij} = P_{0j} + P_{1j}*(trt)$	$Y_{ij} = P_{0j} + P_{1j}*(time) + P_{2j}*(trt)$	$Y_{ij} = P_{0j} + P_{1j}*(time) + P_{2j}*(trt) + P_{3j}*(time*trt)$	$Y_{ij} = P_{0j} + P_{1j}*(trt)$ ---------- $\delta = \frac{B_{10}}{\sigma}$ or $\frac{B_{10}}{\sqrt{\tau_{00}}}$	$Y_{ij} = P_{0j} + P_{1j}*(time) + P_{2j}*(trt) + P_{3j}*(time*trt) +$ *autocorrelation term*	$\log[\Phi/(1-\Phi)] = P_{0j} + P_{1j}*(time) + P_{2j}*(trt) + P_{3j}*(order) + P_{4j}*(time*trt) + P_{5j}*(time*order) + P_{6j}*(trt*order) + P_{7j}*(time*trt*order)$ ***Simplified to:*** $\log[\Phi/(1-\Phi)] = P_{0j} + P_{1j}*(trt)$	
Level 2 Equations	$P_{0j} = B_{00} + R_{0j}$ $P_{1j} = B_{10} + R_{1j}$	$P_{0j} = B_{00} + R_{0j}$ $P_{1j} = B_{10} + R_{1j}$ $P_{2j} = B_{20} + R_{2j}$	$P_{0j} = B_{00} + R_{0j}$ $P_{1j} = B_{10} + R_{1j}$ $P_{2j} = B_{20} + R_{2j}$ $P_{3j} = B_{30} + R_{3j}$	$P_{0j} = B_{00} + R_{0j}$ $P_{1j} = B_{10} + R_{1j}$	$P_{0j} = B_{00} + R_{0j}$ $P_{1j} = B_{10}$ $P_{2j} = B_{20} + R_{2j}$ $P_{3j} = B_{30}$	$P_{0j} = B_{00} + R_{0j}$ $P_{1j} = B_{10} + R_{1j}$ $P_{2j} = B_{20} + R_{2j}$ $P_{3j} = B_{30} + R_{3j}$ $P_{4j} = B_{40} + R_{4j}$ $P_{5j} = B_{50} + R_{5j}$ $P_{6j} = B_{60} + R_{6j}$ $P_{7j} = B_{70} + R_{7j}$ ***Simplified to:*** $P_{0j} = B_{00} + R_{0j}$ $P_{1j} = B_{10} + R_{1j}$	
Basline Level	B00 = 4.743**	B00 = 4.603**	B00 = 4.532**	B00 = 4.743**	B00 = 4.546**	B00 = -0.103	B00 = -0.101
Baseline Slope		B10 = 0.028	B10 = 0.043		B10 = 0.042**		
Effect of Trt on Level	B10 = 4.198**	B20 = 4.054**	B20 = 4.248**	B10 = 4.198**	B20 = 4.270**	B10 = 2.325**	B10 = 2.318**
Effect of Trt on Level			B30 = -0.027		B30 = -0.026**		
SD: Baseline Level	$\tau_{00}^{1/2}$= 1.169**	$\tau_{00}^{1/2}$= 1.248**	$\tau_{00}^{1/2}$= 1.234**	$\tau_{00}^{1/2}$= 1.169**		$\tau_{00}^{1/2}$= 0.457**	$\tau_{00}^{1/2}$= 0.439**
SD: Baseline Slope		$\tau_{11}^{1/2}$= 0.018	$\tau_{11}^{1/2}$= 0.021				
SD: Effect of Trt on Level	$\tau_{11}^{1/2}$= 0.889**	$\tau_{22}^{1/2}$= 0.829**	$\tau_{22}^{1/2}$= 1.036**	$\tau_{11}^{1/2}$= 0.889**		$\tau_{11}^{1/2}$= 0.365*	$\tau_{11}^{1/2}$= 0.323*
SD: Effect of Trt on Slope			$\tau_{33}^{1/2}$= 0.044				
SD: Within-Subjects (σ)	σ = 0.910	σ = 0.883	σ = 0.867	σ = 0.910		σ = 0.832	
				$\frac{B_{10}}{\sigma} = \frac{4.198}{0.910}$ ***d* = 4.613** *or* $\frac{B_{10}}{\sqrt{\tau_{00}}} = \frac{4.198}{1.169}$ ***d* = 3.591**	**rho** = 0.093 *(SE = 0.037)*		

*p<.05, **p<.01

Table 11. Taylor & Alber (2003): Translated Estimates

Method	a) ANOVA Method	b) ANCOVA-like Method	c) Piecewise Regression Method	d) Effect Size Methods	e) ITSACORR Method	f) Multilevel Modeling Method	
Baseline Phase Level	4.743	4.603	4.532	d = 4.613 or 3.591	4.546	4.743	4.748
Baseline Phase Slope		0.028	0.043		0.042		
Treatment Phase Level	8.941	8.657	8.780		8.816	9.022	9.018
Treatment Phase Slope		0.028	0.016		0.015		

Model A – Analysis of Variance (ANOVA)

The multilevel model built to simulate the two-way ANOVA model suggested by Gentile, Roden, and Klein (1972) in *HLM 6.0* is presented above in equation 10. According to the estimates produced for this model, on average, subjects scored 4.7 of 10 items correct in the baseline phase and 8.9 of 10 correct in the CWPT (treatment) phase, a treatment effect of about 4.2. Using this simple model, variance estimates (shown in Table 10) indicate that a substantial amount of between-person variation remains unexplained ($\sqrt{\tau_{00}}$ = 1.169, $\sqrt{\tau_{11}}$ = 0.889). That is, this model may not sufficiently explain patterns of data for all subjects.

Model B – Analysis of Covariance (ANCOVA)

The multilevel model built to simulate Gorsuch's (1983) ANCOVA-like method in *HLM 6.0* is presented above in equation 11. Estimates produced by this model indicate a very slight positive trend in the data. However, this effect did not reach statistical significance, indicating there is no substantial trend in these data. The average baseline level is estimated by this model as 4.6 of 10 items and the CWPT phase mean level is estimated as 8.7 of 10, a treatment effect of 4.1 items. Between-subjects variance estimates indicate that even after adding a term for slope to the model, a significant amount of variance remains in the baseline and phase effect estimates ($\sqrt{\tau_{00}}$ = 1.248, $\sqrt{tau_{22}}$ = 0.829) while no real variance remains on the trend (or time slope)

estimate ($\sqrt{\tau_{11}}$ = 0.018). That is, the determination that there is no substantial consistent trend in the data seems to hold for all subjects. The next model includes a term to test whether a trend in the data varies by phase.

Model C – Piecewise Regression

The multilevel model built to simulate the piecewise regression technique in *HLM 6.0* is presented above in equation 12. Estimates produced by this model indicate mean ending baseline levels of 4.5 items correct and mean treatment phase levels of 8.8 items correct of 10, a treatment effect of about 4.2. Estimates of trend in both baseline and treatment phases show no significant slope (or change in slope) exists in these data. Consistent with the models described above, between-subjects variance estimates indicate significant between-person variation remains on the baseline and phase effect estimates ($\sqrt{\tau_{00}}$ = 1.234, $\sqrt{\tau_{22}}$ = 1.036), while estimates of flat trend effects seem to be sufficiently explained for all subjects ($\sqrt{\tau_{11}}$ = 0.021, $\sqrt{\tau_{33}}$ = 0.044).

Model D – Effect Sizes

In order to examine effect size assessments again using *HLM 6.0*, the phase effect computed for the ANOVA model (Model *A*) was considered as the numerator. This estimate (B_{00} = 4.743) was then divided by two estimates of standard deviation also produced by the ANOVA model: the within-subjects standard deviation estimate (σ=0.91) and the between-subjects standard deviation for the baseline estimate ($\sqrt{\tau_{00}}$ =1.169).

$$\delta = \frac{B_{10}}{\sigma} = \frac{4.743}{0.91} = 4.613 \quad (18)$$

$$\delta = \frac{B_{10}}{\sqrt{\tau_{00}}} = \frac{4.743}{1.169} = 3.591 \quad (19)$$

This effect size using the within-subjects standard deviation is calculated to be 4.613. The effect size using the between-subjects standard deviation on the baseline level is calculated as 3.591. Again, effect size methods like these can be misleading because they produce values that are not comparable to those from group design research. In particular, within-subjects estimates of variation are often much smaller than those from between-subjects studies. While group design effect sizes usually range from about -1 to +1, single case design effect sizes can often be much larger, and both of these estimates (4.613 and 3.591) are out of this range.

Model E – ITSACORR (Time Series)

In order to duplicate time series methods in HMLM, a model was built as specified in equation 15. Again, the full specified model would not run. Two major steps were taken to solve this problem. First, realizing that the data set was too small for the program to estimate accurately, data on the original four subjects were duplicated to create a dataset of 35 subjects instead. In addition, as in the demonstration above, random effects on the two time slope terms were constrained to zero. (Note that again these estimates had not been determined to vary in any of the models run previously.) The model specified in equation 16 above was run in HMLM.

Estimates produced by HMLM indicate that the autocorrelation in the data is extremely small (0.093). Although the output comparing the first-order autoregressive model with the model without an autoregressive term indicates that even this small amount of autocorrelation is statistically significant (Chi-square = 6.428, p=.011), this is likely just an artifact of the data duplication. The data had to be duplicated to represent 35 subjects when there are really only four in the published data set. As well, a comparison of estimates across the models with and without autoregressive terms included shows no real difference in fixed effects. Either way,

subjects are estimated to score about 4.5 items correct in the baseline phase and about 8.8 in the treatment phase, a treatment effect of about 4.3 items. Very small slopes and changes in slopes were estimated to be statistically significant, likely also an artifact of the artificially inflated sample size. Table 12 presents this comparison. (Note that the HMLM models with autocorrelation could not be run using the Binomial distribution; thus, these estimates are provided on the raw count scale.) It may be concluded that no real autocorrelation exists in these data and that the model without the autoregressive term is sufficient and a more parsimonious model of these data. Again, had we not tested for the presence of autocorrelation in this way, we would not have know it was unnecessary to accommodate this complexity of small sample data.

Table 12. Taylor & Alber (2003): Checking Autocorrelation (ITSACORR)

	No Autocorrelation (No AR)	**Autocorrelation (AR)**
Baseline Level*: Estimate (p-value)*	B00 = 4.550 *(p<.001)***	B00 = 4.546 *(p<.001)***
Effect of Time*: Estimate (p-value)*	B10 = 0.042 *(p<.001)***	B10 = 0.042 *(p<.001)***
Effect of Trt on Level*: Estimate (p-value)*	B20 = 4.272 *(p<.001)***	B20 = 4.270 *(p<.001)***
Effect of Trt on Time Slope*: Estimate (p-value)*	B30 = -0.028 *(p=.001)***	B30 = -0.026 *(p=.003)***
Rho*: Estimate (SE)*		rho = 0.093 *(0.037)*
Sigma-Squared*: Estimate (SE)*	σ^2 = 0.763 *(0.037)*	σ^2 = 0.773 *(0.039)*
Deviance*: Estimate*	*2447.698*	*2441.271*
Model Comparison	Chi-square =6.428 ***(p=.011)***	
(Translated) Baseline Phase Level	**4.550**	**4.546**
(Translated) Baseline Phase Slope	**0.042**	**0.042**
(Translated) Treatment Phase Level	**8.822**	**8.816**
(Translated) Treatment Phase Slope	**0.014**	**0.016**

Model F – Multilevel Modeling Method

The multilevel modeling approach, for which the *HLM* program was designed, was also conducted for this comparison. In addition to the strengths demonstrated above, rather than

treating all types of outcome variables as continuous as the previous methods do, *HLM 6.0* allows for special accommodation of discrete outcome variables like a proportion, as in this study. Models A through E were run using the raw count of items correct as the dependent variable. This model was run using the Binomial distribution. An outcome variable that expresses a proportion, like the correct number of responses out of 10 total trials, is best modeled by the Binomial distribution. As mentioned above, this model makes estimations on a log odds scale that can be translated back to counts by: (1) exponentiating the log odds to compute the odds, and then (2) dividing the odds by (1 + odds) to compute the probability of a successful trial. The Binomial distribution is used to model the frequency of some event out of a total known number of possible binary trials (i.e., a proportion or percentage). It is important that analytical methods acknowledge appropriate distributions for the dependent measures being studied and accommodate for any potential overdispersion present. Overdispersion occurs when count data (including proportions) exhibit greater variability than would be expected by the Binomial distribution model. Therefore, a model was first run to account for potential overdispersion.

The expressions in equation 20 form the first model run in the program:

$$\begin{aligned} \text{Level-1: } & log[\Phi/(1-\Phi)] = P_{0j} + P_{1j}*(time) + P_{2j}*(trt) + P_{3j}*(order) \\ & + P_{4j}*(time*trt) + P_{5j}*(time*order) + P_{6j}*(trt*order) \\ & + P_{7j}*(time*trt*order) + e_{ij} \end{aligned} \tag{20}$$

$$\begin{aligned} \text{Level-2: } & P_{0j} = B_{00} + R_{0j} \qquad P_{4j} = B_{40} + R_{4j} \\ & P_{1j} = B_{10} + R_{1j} \qquad P_{5j} = B_{50} + R_{5j} \\ & P_{2j} = B_{20} + R_{2j} \qquad P_{6j} = B_{60} + R_{6j} \\ & P_{3j} = B_{30} + R_{3j} \qquad P_{7j} = B_{70} + R_{7j} \end{aligned}$$

where, Φ is the probability of a successful trial.

As described in the above demonstration, the multilevel modeling approach starts out by entering all available observation-level terms (including all possible main effects, two- and three-

way interactions) into the level-1 model (as in equation 18) and then parsing out non-significant variables until only significant contributors to the model are left. The final model that remained for these data simplified to the model run to simulate the ANOVA method (Model *A*). This model is expressed in equation 21:

$$\text{Level-1: } log[\Phi/(1-\Phi)] = P_{0j} + P_{1j}*(trt) + e_{ij} \quad (21)$$

$$\text{Level-2: } P_{0j} = B_{00} + R_{0j}$$
$$P_{1j} = B_{10} + R_{1j}$$

This model estimates that the average log odds ratio for all subjects is -0.103 (B00). The expected probability of a student getting an item correct in the baseline phase is then 0.474, or about 4.7 items out of 10 correct. [exp(-0.103) = 0.902; 0.902/(1+0.902) = 0.902/1.902 = 0.474 = Φ; 0.474*10 items = 4.74] The rate of change in log odds as subjects switch from baseline to treatment phase is 2.325 (B10). The expected probability of a student getting an item correct in the treatment phase is then 0.902, or about 9 items out of 10 correct, a treatment effect of about 4.3. [exp(-0.103+2.325) = exp(2.222) = 9.226; 9.226/(1+9.226) = 9.226/10.226 = 0.902 = Φ; 0.902*10 items = 9.02] A statistically significant amount of between-subject variance remains unexplained by this model, however level-2 data (i.e., subject-level characteristics) made available by the authors (e.g., age, pre-test scores, length of time receiving special education services) were not varied enough to explain differences in effects between subjects. Thus, no available level-2 variables fit sufficiently into the model of these data.

According to Raudenbush, et al. (2004), if data are overdispersed, the within-subjects variance (σ^2) estimate will be greater than 1.0. The within-subjects variance estimated by this model was less than 1.0 (σ=0.832, σ^2=0.692) which suggested that these data were not overdispersed and there was no reason to accommodate for this complication. To verify that no

overdispersion existed, the model was run again without overdispersion. Table 13 presents estimates from the two models (with and without overdispersion accounted for) side by side.

Table 13. Taylor & Alber (2003): Checking Overdispersion

	Overdispersion	**No Overdispersion**
Basline Level*: Estimate (SE)*	B00 = -0.103 *(0.246)*	B00 = -0.101 *(0.245)*
Effect of Trt on Level*: Estimate (SE)*	B10 = 2.325 *(0.232)*	B10 = 2.318 *(0.235)*
SD: Baseline Level	$\tau_{00}^{1/2}$= 0.457	$\tau_{00}^{1/2}$= 0.439
SD: Effect of Trt on Level	$\tau_{11}^{1/2}$= 0.365	$\tau_{11}^{1/2}$= 0.323
SD: Within-Subjects (σ)	σ = **0.832**	
Likelihood Function	-133.2979	-135.8931
(Translated) Baseline Phase Level	**4.743**	**4.748**
(Translated) Treatment Phase Level	**9.022**	**9.018**

The estimates produced by the simpler model, without an accommodation for overdispersion, do not differ substantially from those produced by the accommodated model. In keeping with the desire for the simplest well-fitting model, the estimates provided by the simpler model (which translate to the same 4.7 items correct in the baseline phase and 9.0 items correct in the treatment phase) are taken as the most suitable estimates possible for these data. Had this comparison not be conducted, we would not have known that it was unnecessary to accommodate for this complexity of single case design count outcome data. It should be emphasized that between-subjects variance estimates indicate that this model does not fit the data sufficiently for all subjects. However, any information that was made available by the authors of this published study could not explain these differences. Had this model been run using the raw count of items correct as Models A through E were run, estimates would have been identical to that of Model A (ANOVA), which would have slightly underestimated the treatment effect, since the simplified version of the multilevel model includes terms only for baseline level and a phase effect.

Estimation Methods: Maximum Likelihood Empirical Bayes vs. Fully Bayesian

Again, now that the multilevel modeling method without an accommodation for overdispersion has been established as the most appropriate model for these data so far, an exploration of estimation methods was also conducted. Multilevel models were run both in *HLM 6.0*, utilizing ML-EB estimation, and in *WinBUGS*, using fully-Bayesian estimation.[4] Parameters estimated by each program are presented in Table 14.

Table 14. Taylor & Alber (2003): Comparing Estimation Methods (ML-EB vs. Fully Bayesian)

	Maximum Likelihood Empirical Bayes (ML-EB) Estimates	**Fully Bayesian Estimates**
	Estimate *(SE)*	Estimate *(SE)*
Baseline Level	B00 = -0.101 *(0.245)*	B00 = -0.099 *(0.388)*
Effect of Treatment on Level	B10 = 2.318 *(0.235)*	B10 = 2.343 *(0.366)*
SD: Baseline Level	u0 = 0.439	u0 = 0.483
SD: Effect of Treatment on Level	u1 = 0.323	u1 = 0.362
(Translated) Baseline Phase Level	**4.748**	**4.753**
(Translated) Treatment Phase Level	**9.018**	**9.041**

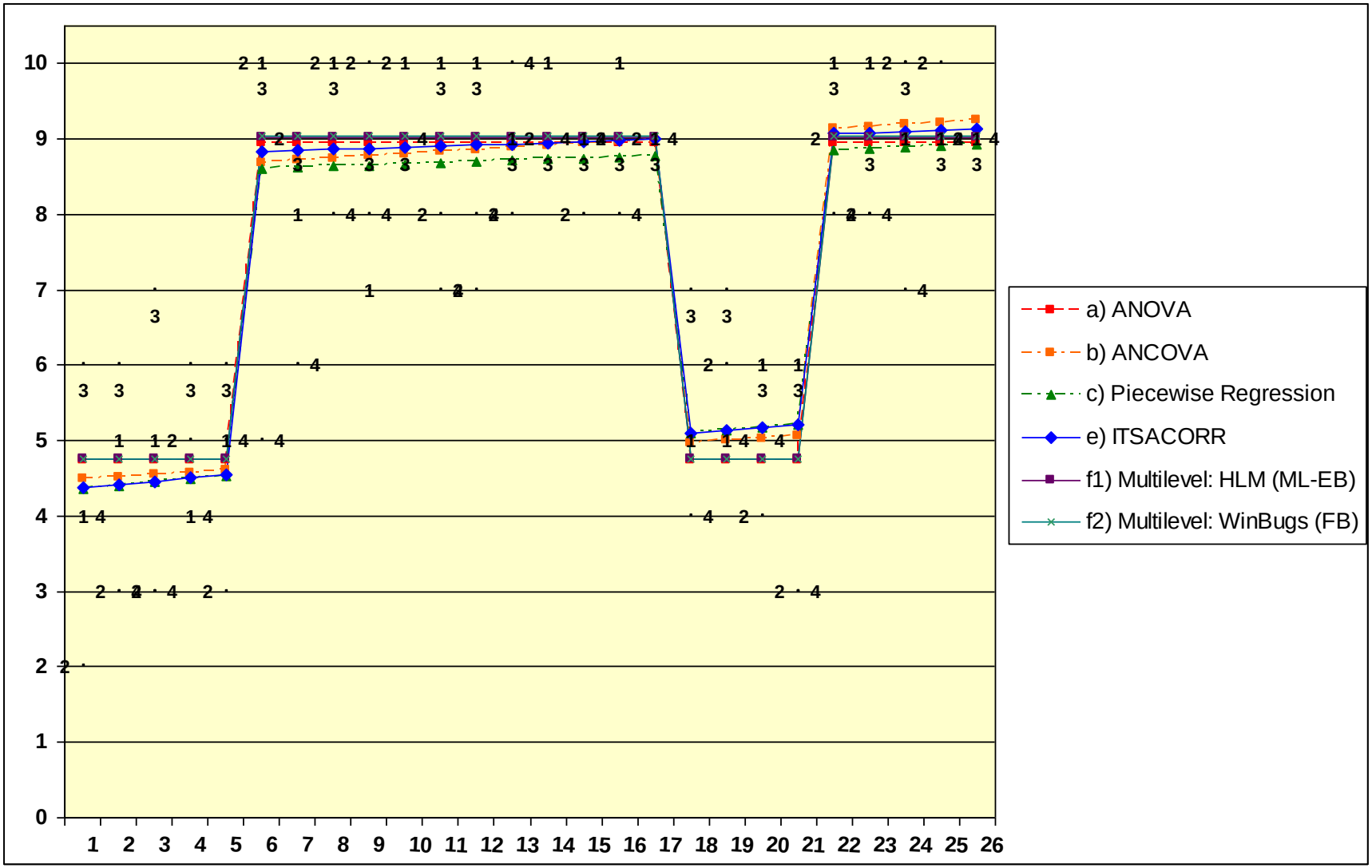
The model run using *WinBUGS*, which incorporates fully Bayesian estimation methods, estimates the average log odds ratio for all subjects at baseline as -0.099 (B00). The expected probability of a student getting an item correct in the baseline phase is then 0.475, or about 4.75 items out of 10 correct. [exp(-0.099) = 0.906; 0.906/(1+0.906), 0.906/1.906 = 0.475 = Φ; 0.475*10 items = 4.75] The change in log odds as subjects switch from baseline to treatment phase is 2.343 (B10). The expected probability of a student getting an item correct in the treatment phase is then 0.904, or about 9 items out of 10 correct, a treatment effect of 4.25. [exp(-0.099+2.343) = exp(2.244) = 9.431; 9.431/(1+9.431) = 9.431/10.431 = 0.904 = p; 0.904*10 items = 9.04] These estimates of fixed effects are very similar to those calculated by *HLM 6.0* using the ML-EB estimation approach. The standard errors of the estimates, however, are slightly larger as estimated by the *WinBUGS* program than by *HLM 6.0* because the fully

[4] *WinBUGS* code for this analysis is available in the appendix.

Bayesian approach acknowledges greater uncertainty in estimation of effects. Again, since this acknowledgement is more realistic about what is unknown, which is particularly appropriate with small sample ITS data like these, this fully Bayesian estimated model is found to be the most appropriate model for these data produced in this demonstration.

All estimated models are plotted over the original study data in Figure 5.

Figure 5. Taylor & Alber (2003): Model Estimates and Raw Data



Discussion

As schools and programs are pushed to utilize practices with strong and sound evidence of success, researchers have intensified efforts toward the use of more rigorous designs for evaluating program impacts. One such technique is the use of single case research designs. In recent years, single case research designs have become increasingly prevalent across a variety of fields, including education. Single case design research is experimental; it aims to document causal relationships between independent variables and dependent variables. Between- and within-subjects comparisons are used to control for major threats to external and internal validity. It is possible to use data from these designs to conduct systematic analyses to rigorously assess treatment effectiveness, however, data from single case designs are often underutilized or misused. When single case design data have been analyzed statistically, they have often been done so with inappropriate methods which are limited by their inability to take advantage of the unique depth and precision of this type of data or by the fact that they are not based on sound statistical principles.

Single case designs involve repeated documentation of individuals' behavior over time across phases (before, during, and/or after an intervention is implemented), providing rich information about individual behavior patterns that are not available from group design research methodologies, which provide only single number estimates of performance. Because of this unique richness, data from single case research designs require distinct approaches to analysis and interpretation, different from those developed specifically for group design research data. In addition, analysis of data from single case design studies often involves technical data considerations that are not necessary to consider when analyzing data from more common group design studies. These technical considerations include accommodation of discrete rather than

continuous outcome variables (e.g., counts) and of serial dependence of data points. If these technical complexities are ignored, statistical assumptions may be violated and estimates, standard errors, or both may be biased. Appropriate attention to these potential complexities particular to single case design data is essential in order to take full and appropriate advantage of the rich detail available from these designs.

Sound synthesis of these data can provide information about overall effects of an intervention, but beyond summarizing the treatment effect on the average individual, syntheses of data from single case design studies can provide valuable information about how the effects of an intervention vary across cases. In the best circumstances, using the most sound and flexible methods, we can systematically assess whether and how the specific characteristics of individuals or settings can influence treatment effects.

Throughout this paper, a number of methods proposed for analyzing and synthesizing data from single case research design studies have been discussed and demonstrated. These methods include visual inspection, those related to t- and F-tests (i.e., ANOVA and regression-based methods), those that involve the calculation of effect sizes, those that draw on time series methods, nonparametric methods, and those that rely on multilevel modeling methods to describe data patterns and treatment effects. That is, how the effectiveness of an intervention can be predicted by characteristics of subjects or settings.

Of these, it was expected that the multilevel modeling method would provide for the most appropriate, rich, and rigorous analysis and interpretation. Increasingly, multilevel modeling is being accepted by researchers as the most effective and appropriate approach for analyzing these types of data. In order to allow for stable comparisons of each to the multilevel model, each demonstrated method was first reframed as a special case of the multilevel model. This was

possible since many of the methods reviewed were based on an underlying regression model. The statistical methods that did not fit into this framework (e.g., nonparametric methods) were included in the review of methods for the sake of completeness, but not in the demonstration since there was no concrete way of comparing their results to the multilevel modeling method and since they do not allow for modeling or appropriate estimation of magnitude of effects. It is important to emphasize that using hierarchical models in this way is not exactly the same as duplicating these alternative methods. Multilevel modeling software (e.g., *HLM 6.0*) attends to technicalities of the data not assumed by all original models. In addition, many of the reviewed methods were originally developed for N=1 studies, not to synthesize results across multiple subjects.

The selected methods were demonstrated on two reversal design (ABAB or baseline-treatment-baseline-treatment) datasets. The first dataset was simulated to allow for a clear and effective illustration of the ability of each of the demonstrated methods to handle the most common features of single case research design data as they were discussed in the review. The second dataset was taken from a published study so that each method's ability to handle the complexities of real-world data could be tested. For each dataset, variables were coded and centered as needed for improved model interpretation (e.g., time was centered at the final observation of the first baseline phase).

Before any models were run through the multilevel modeling software, preliminary exploration was conducted via visual inspection to identify apparent patterns, similarities and differences between phases and subjects. Visual inspection was expected to provide for quick and easy preliminary analyses of behavior patterns, allowing for simultaneous evaluation of

multiple data features (e.g., level, trend, variability, delay, consistency) in order to guide and confirm findings from more rigorous statistical analyses.

A number of limitations of using visual analysis as the sole method for assessing impacts were discussed, including that it is a subjective analytic method and that it has been shown to perform poorly in several common circumstances (e.g., datasets with short baseline phases). Also, visual analysis alone does not allow for the expression of any quantitative model of data patterns. However, visual inspection's strength in detecting patterns was expected to support more rigorous, statistical approaches to analyzing single case research design data.

In the first demonstration, using the simulated dataset, several data patterns were identified via preliminary visual inspection: negative slope across all phases and subjects, treatment effects on level (i.e., mean levels drop from baseline to treatment phases) and on trend (i.e., more severe negative slope during treatment phases than during baseline phases), and possible differences between subjects in pre-intervention (baseline) levels and the treatment effect on level. Demonstrations of statistical analyses aimed to verify these patterns and to illustrate the strengths and weaknesses of the models discussed.

- The first method, the ANOVA model (A) confirmed a statistically significant treatment effect on level. This model estimated that all subjects start out at around 70 points (69.57) in the baseline phase and, on average, drop to about 29 points (28.65) in the treatment phase. According to estimated variance components of this model, these estimates of baseline level and treatment effect did not sufficiently explain data patterns for all subjects. Though slope was detected visually, this model could not account for or estimate any slope in the data. As was suspected, the ANOVA method was not able to capture the patterns expressed in these data.

- The ANCOVA method (B) builds upon the ANOVA method (A) by adding a term to estimate slope in the data. This model estimated that subjects have mean baseline levels of about 75 points (75.09) and mean treatment levels of about 43 points (43.37), an average treatment effect on level of about 32 points; and that across all phases, scores slope negatively at an average rate of -1.8 points per observation, or -9.0 points over the course of each 5-observation phase. This model determined these fixed effects to be statistically significant in magnitude. However, not all effects were found to fit for all subjects. According to estimated variance components, estimates of baseline level and treatment effect on level did not sufficiently explain data patterns for all subjects. As expected, though the ANCOVA model was able to estimate slope in the data, it did not include a term to identify whether slope effects changed from phase to phase, while preliminary visual inspection suggested that this might be the case. These results support the prediction that a more complex model might better express these data.
- The Piecewise Regression model (C) builds upon the ANCOVA model (B) by adding a term to allow slope to vary across phases. This model estimated that subjects start out with mean baseline levels of about 73 points (73.18) which slope negatively at an average rate of about -1.2 points per observation, or about - 6.0 points per phase. Treatment levels were estimated to drop to about 48 points (48.46) and to slope more severely at an average rate of about -2.5 points per observation, or about -12.5 points per phase. Similar to the previous models, variance estimates for the ANCOVA model indicated that significant between-person variation remains on the estimates of baseline and treatment levels but that slopes seem to be sufficiently explained for all subjects. This model performed better and was able to provide estimates consistent with conclusions drawn from preliminary visual inspection (e.g.,

treatment effect on level and slope and differences between subjects), but it was not enough to explain data patterns across all subjects.

- Due to their prominence in the literature, effect size methods (D) were also estimated for these data. These methods express treatment effects in terms of between- or within-subjects standard deviations. Here, the phase effect computed for the ANOVA model (A) was utilized as the numerator (-40.92), which was divided by two different estimates of standard deviation: the within-subjects standard deviation estimated for that ANOVA model (11.603) and the between-subjects standard deviation estimated for that ANOVA model (5.813). The resulting effect size estimates (-3.53 and -7.04, respectively) could be misleading because they are not comparable to effect size values more commonly known and estimated for group design research (e.g., Cohen's *d*), which (although they can range more widely) typically range from -1 to +1. Also, these one-number estimates of effect are superficial descriptors that on their own do not reveal much about rich patterns in the data.
- The fifth demonstrated model, ITSACORR (E), is a time series method. Time series methods allow for expression of data patterns in a regression-like approach more similar to the first three models demonstrated (A-C). Beyond the capacity of the first three models however, this model also allows for estimation of autocorrelation, an expression of the extent to which data points are serially dependent. Autocorrelation is one of the technical complexities of single case design data that does not have to be considered for group design research. In single case design data, it is expected that sources of variation that operate at one observation are likely to produce carryover effects that influence later observations of the same subjects (i.e., that errors at one point are predictive of errors at future points). The presence of autocorrelation in data violates the assumption of independence of observations

required for many statistical models. As described above, if residuals are autocorrelated, model estimates may be biased. In this case, although a moderate amount of autocorrelation was intentionally built in to the simulated data (0.25), the ITSACORR model was able to detect only an extremely small, statistically non-significant, amount of autocorrelation in these data (0.127). Additional exploration comparing this model with and without serial dependence accounted for verified the non-significance of this autocorrelation (Chi-square = 1.965, p=.157). However, had the presence of autocorrelation not been tested for, there would have been no way of knowing that it was unnecessary to accommodate this complexity of small sample data. The estimates produced by the accepted ITSACORR model without autocorrelation were identical to those produced for the Piecewise Regression model (C), with baseline and treatment levels estimated at 73 and 48 points and baseline and treatment slopes estimated at -1.2 and -2.5 points per observation, respectively. The ITSACORR model, as demonstrated, did not allow for the estimation of variance components to determine whether these fixed effects fit suitably for all subjects. Given the significant variance remaining between subjects as estimated by the Piecewise Regression model (C), and that fixed effects were estimated identically here, it was concluded that the ITSACORR model (E) may not be the best model to fully describe patterns in these data.

- Finally, the actual multilevel modeling method (F) was run to illustrate how this method is able to accommodate the complexities of these data and explain treatment effects for all subjects as hypothesized. The multilevel model expresses each level in the data structure (e.g., observations and individuals) with its own statistical model. This method involves entering all available observation-level terms into the model and pulling out non-significant fixed effects until only significant contributors to the model remain. When computed with

these data, the simplified level-1 model resembled the Piecewise Regression model (C) as well, leaving terms for the baseline level, slope, and treatment effects on level and on slope. (All effects of phase order or other two- and three-way interactions were deemed non-significant for these data.) However, as discussed above under models (C) and (E), these estimates alone could not sufficiently explain between-subject variation on baseline levels or phase effects.

- Unlike the methods discussed above, however, the multilevel modeling method allows for testing whether between-subjects differences may be accounted for by subject characteristics (level-2 variables simulated for this purpose). And indeed, once subject characteristics were entered into the model appropriately, no significant variation between subjects could be found remaining. This final model estimated that, on average, subjects score about 72 points (72.31) in the baseline phase, losing about 1.2 points per observation. However, subjects with higher values on the simulated level-2 variable loaded onto the intercept were found to have higher mean baseline levels (about 3.5 points higher on intercepts for every 1 point above average on the level-2 characteristic). Mean treatment levels were estimated to be about 48 points (48.06) and mean treatment slopes about -2.5 points per observation. Again, level-2 subject characteristics were able to explain away all remaining significant between-subjects variation, such that subjects with higher than average values on the subject characteristic (simulated for this purpose) had smaller treatment effects, dropping 4 points less from baseline to treatment for every 1 unit above average they fell on the level-2 variable.

The final multilevel model was able to explain data patterns for all subjects and as expected, of all demonstrated methods, best expresses patterns in the data. However, as

discussed above, when sample size is small, as it is here, the ML-EB estimation approach used by the modeling software employed for this demonstration (i.e., *HLM 6.0*) may not be best for producing stable estimates of effect. Instead, the fully Bayesian estimation approach (used by software like *WinBUGS*) may make better estimates by acknowledging uncertainty in estimating variance components and as such may be more appropriate for use with small sample data. Therefore once the multilevel model was established as the most suitable of all models demonstrated, it was again estimated using the fully Bayesian estimation approach for comparison to the ML-EB approach.

The model run using *WinBUGS*, which incorporates fully Bayesian estimation methods, estimated almost identical fixed effects as the ML-EB approach: an average baseline level of about 72 (72.49) and an average treatment phase level of 49 (48.78), with a loss of 1.2 points per day in the baseline phase and 2.5 points per day in the treatment phase, and similar effects of the subject characteristics. The standard errors of the estimates, however, are slightly larger as estimated by the *WinBUGS* program (fully Bayesian) than by *HLM 6.0* (ML-EB) because the fully Bayesian approach acknowledges greater uncertainty in estimation of effects. This acknowledgement, expressed by larger standard errors, is more realistic about what is unknown, which is particularly appropriate with small sample data like these. Thus, the fully Bayesian estimated model was concluded to be the most suitable model for these data produced in this demonstration. Figure 3 illustrates the visual comparison across models.

In order to compare these methods on data collected in a real educational setting, this demonstration set was repeated with a dataset from an existing study. In the published article from which this data for the second demonstration was taken, the study's authors concluded that

their intervention was effective after simply comparing the raw difference in phase mean values, without calculating any statistics to test whether this difference was large enough to be considered real. Visual inspection indicated that treatment effects on level were small and that slopes were relatively flat. Again, each of the six competing methods was demonstrated using these data to test whether these or other effects are statistically significant

- The first method, the ANOVA model (A), was able to test the significance of that small identified treatment effect. It estimated that all subjects start out at around 4.7 items correct in the baseline phase and, on average, increased to about 8.9 points in the treatment phase. The model determined that this treatment effect (average jump of 4.2 points from baseline to treatment) was statistically significant. However, according to estimated variance components of this model, these estimates of baseline level and treatment effect did not sufficiently explain data patterns for all subjects (i.e., variance components that remain are of statistically significant magnitudes). Again, as was suspected, the ANOVA method was not able to capture the patterns expressed in these data.
- The ANCOVA model (B) estimated that subjects have mean baseline levels of about 4.6 items and mean treatment levels of about 8.7 items correct, a statistically significant treatment effect on level of about 4.1 points; and, that across all phases, no significant slope in scores could be detected. However, while the estimate of non-significant slope across phases was deemed well-fitting for all subjects, variance components for estimates of baseline level and treatment effect on level did not sufficiently explain data patterns for all subjects.
- The Piecewise Regression model (C) estimated that subjects started out with mean baseline levels of about 4.5 items correct and that treatment levels to jump to about 8.8 items correct,

a statistically significant treatment effect on level. Estimates of slope and change in slope were not found to be statistically significant. That is, slopes were estimated to be essentially flat in both phase types. Similar to previous models, variance estimates indicated that significant between-person variation remained on the estimates of baseline and treatment levels but that the finding of flat slopes fit data patterns for all subjects.

- Again, effect size methods (D) were estimated for these data. And again, the phase effect computed for the ANOVA model (A) was used as the numerator (4.198) and divided by two different estimates of standard deviation: the within-subjects standard deviation estimated for the ANOVA model (0.910) and the between-subjects standard deviation estimated for the ANOVA model (1.169). As mentioned above, the resulting effect size estimates (4.613 and 3.591) may be misleading because they are not comparable to effect size values more commonly estimated for group design research which typically range from -1 to +1, although they can range more widely.
- The fifth demonstrated model, ITSACORR, a time series method (E), allowed for estimation of autocorrelation in these data. However, because this study provided data on only four subjects, this model would not run through the HLM software as entered. Instead, the data had to be duplicated numerous times to initiate the model run. Once the ITSACORR model was run, the model did detect a statistically significant amount of autocorrelation in these data (0.093). However, this amount is very small and likely just an artifact of the data duplication. A comparison of the fixed effect estimates across the models with and without autoregressive terms included showed no real difference between estimates made by each. Either way, subjects are estimated to score about 4.5 items correct in the baseline phase and about 8.8 items correct in the treatment phase. The model estimated very minor slope (0.04)

and change in slope (-0.03) in these data, likely another artifact of the multiple duplication. The ITSACORR model, as demonstrated, did not allow for the estimation of variance components to determine whether these effects fit suitably for all subjects.

- The multilevel modeling method (F) was run using the Binomial distribution, instead of the normal distribution used for the previous models. The Binomial is more appropriate for modeling proportions (like these, number of items correct out of 10) than is the normal distribution. This type of model makes estimates on a log odds scale that can be translated back to counts by exponentiating and dividing as described and demonstrated. It is important to acknowledge and use appropriate distributions for each outcome variable and to accommodate any potential overdispersion, which can occur when count data exhibit greater variation than would be expected by the Binomial distribution model. This model was run with an accommodation for overdispersion. Using Binomial distribution and accommodating for overdispersion, all available observation-level terms were entered into the model and non-significant fixed effects were pulled out one-by-one until only significant contributors to the model remain. This time, the simplified level-1 model resembled the ANOVA model (A), leaving terms for the baseline level and treatment effect on level only. (All effects of slope, phase order and two- and three-way interactions were deemed non-significant for these data.) Translated estimates expressed an average baseline level of 4.7 items correct and an average treatment level of 9.0 items correct. However, significant between-subject variation remained unexplained by this model; none of the subject-level characteristics (alone or in combination) provided in the study (e.g., age, pre-test scores) were able to account for these differences in effects, likely due to lack of variability.

Overdispersion was also tested for these data. Low within-subjects variance (0.832) estimated in the accommodated model suggested that these data were not overdispersed, so a model was run without accommodating for overdispersion for comparison. Fixed effect estimates for this unaccommodated model corresponded with those of the overdispersion accommodated model with estimates of 4.8 and 9.0 items correct in baseline and treatment phases, respectively. Since the estimates produced by the simpler model do not differ substantially from those produced by the accommodated model and since the within-subjects variance estimate was less than 1.0 in the accommodated model (suggesting no overdispersion), estimates from the simpler model were deemed sufficient. Estimates of between-subjects variance for this unaccommodated model indicated that this model did not fit well for all subjects either, but again, additional data on subject characteristics made available by the study's authors were not able to explain these differences.

In order to again demonstrate ML-EB estimation in comparison to fully Bayesian estimation, this model was also run using *WinBUGS*. This model estimated an average baseline level of about 4.8 items correct and an average treatment phase level of 9.0 items correct. These estimates of fixed effects are identical to those calculated using the ML-EB estimation approach above. The standard errors of the estimates, however, are again slightly larger as estimated by the *WinBUGS* program (fully Bayesian) than by *HLM 6.0* (ML-EB) because as mentioned, the fully Bayesian approach acknowledges greater uncertainty in estimation of effects, and is more realistic about what is unknown, which is particularly appropriate with small-sample data like these. The fully Bayesian estimated model was decided to be most appropriate for these published data. Figure 5 illustrates a visual comparison across models plotted over the data.

Statistical methods developed to test treatment effectiveness in data like these should be flexible to accommodate the possible variations of single case design types (e.g., ABA, ABAB, ABCA, ABACA). Most of the models demonstrated here are limited in their ability to accommodate these assorted variations or the design aspects or technicalities that are inherent to data from single case designs. For example, the ANOVA model (A) does not allow for estimation of the effect of time (i.e., slope) on the target behavior. In the dataset simulated for the first demonstration, significant slope effects were identified (via both visual inspection and more sophisticated statistical models). If the ANOVA method had been used solely to model that data, this time effect would not have been accounted for, which would have oversimplified data patterns and biased estimates of baseline and treatment levels. Since time slope is an essential aspect of the data to be examined, it can be determined that the ANOVA method is not an optimal choice for modeling single case design data.

The slightly more accommodating ANCOVA model (B) does allow for estimation of slope, but only in a way that is constant across all phases. In other words, it does not allow for estimation of possible treatment effect on slope; of how slopes may change from baseline to treatment phases. A statistically significant treatment effect on slope was also found in the simulated dataset. If the ANCOVA model had been used exclusively to analyze these data, this treatment effect on slope could not have been identified or estimated. Since treatment effect on slope is also a key feature of the data patterns to be tested, it can also be concluded that the ANCOVA method is not an optimal choice for modeling single case design data either.

The Piecewise Regression model (C) includes a term to test for the treatment effect on slope, and, in this way, is slightly more flexible in analyzing these types of data. However, this model cannot accommodate complex designs (e.g., developers suggest using data from only the

first AB phase pair to calculate effects) or essential technical aspects of single case design data (e.g., possible overdispersion, small number of data points, explanation of heterogeneity of effects). In both demonstration sets, statistically significant amounts of between-subjects variation were found remaining after modeling the data using the Piecewise Regression method. This method does not allow for explanation of this heterogeneity using other available data on subjects or settings. Although it performs better than the models demonstrated before it, the Piecewise Regression method is also deemed to be inadequate for analyzing data from these design types.

At face value, the effect size method (D) for modeling single case design data performs poorly because it does not take advantage of the richness unique to data from these designs. Most simply, one number estimates of effect size cannot effectively summarize patterns in data level, trend, variability, and phase change. In addition, effect size values calculated from single case design data are often out of range of those more typically computed (and more familiar) from group design data. The within-subjects standard deviations used as the denominator in some effect size equations for single case design data are inherently different from the standard deviations used in calculating effect sizes for group design data. The within-subject standard deviation estimate is often expectedly lower than that of the typical between-subject standard deviation used in group design effect size computation because it represents variation within the data collected for each subject (i.e., less variation in performance is expected within a person than across people). A lower denominator of this type will then naturally produce a higher estimate of effect size. Estimates of between-subjects standard deviation for single case design data may also be lower than those calculated in group design research since subjects studied under single case studies are often more initially homogeneous on the target behavior by design

(i.e., study samples may be selected from an artificially restricted population since subjects are often chosen for study based on their tendency to do too much or too little of the target behavior). Again, these smaller estimates of standard deviation will lead to larger estimates of effect size. In the demonstrations performed here, computed effect sizes ranged from absolute values of 3.5 to 7.0, all seemingly large compared to the typical range of 0 to 1 for group design data. However, this method's least desirable feature is its inability to quantify relationships and express the rich patterns in the data uniquely made available by single case designs.

Time series methods (E) do take advantage of the richness of single case design data. These methods allow for modeling of levels and slopes and treatment effects on levels and slopes. As well, they can estimate and accommodate for possible autocorrelation in the data, one of the technical complexities discussed above. However, ITSACORR, the time series model demonstrated here, is not recommended for use with data with less than 10 observations per phase. This minimum is not realistic of typical single case design studies, especially in educational settings. In fact, both of the demonstrations performed used datasets that violated this minimum (5 points per phase in the first demonstration, and 3-12 points per phase in the second demonstration). And indeed, in both demonstrations, data had to be manipulated in order to run the models. That is, the HLM software model would not run the model with autocorrelation without constraining the model and/or artificially duplicating the data to falsely include more data points or subjects. Therefore, although this method does allow for expression of several important features of the data and although it allows for the testing of autocorrelation in the data, its inability to analyze data from small samples with few observations per phase make it a less than ideal option for analyzing data from singe case design studies in the ways aimed here.

The multilevel modeling method (E) was expected to be the best performing method for analyzing single case design data and, in many ways, it was. The multilevel model is the most flexible of the methods demonstrated. It allows for estimation of those effects modeled by other methods (e.g., level, slope, treatment effect on level and slope) as well as other effects (e.g., pair order). The multilevel modeling method can handle unequal numbers and spacing of observations, count outcome variables (e.g., rates or proportions), complex designs, and fewer data points than otherwise possible. It can also test for and accommodate for overdispersion, which is common in categorical outcome data. Whereas the HLM software was unable to run the model with autocorrelation with few observations per phase in these demonstrations, the multilevel modeling method did not have that problem. Because information for all subjects is efficiently used simultaneously to estimate parameters, the multilevel modeling method can be used with data that has only a small number of observations available for each subject and still obtain reliable estimates (Van den Noortgate & Onghena, 2003a, 2003b). And, of all demonstrated methods, only this true multilevel model can statistically explore heterogeneity of effects across replications of the intervention on different subjects or in different settings. In the first demonstration, all remaining significant between-subject variance could be explained by subject characteristics such that subjects' data patterns could be estimated using level-2 variables (i.e., subject characteristics) to account for and compute deviation from the average model. Being able to explain how treatment effects may differ across subjects or settings of different types can help to guide clinicians as to where and with whom particular interventions are best implemented. It is clear that the multilevel modeling method is the best of those demonstrated for analyzing these types of data in a way that allows for modeling quantifiable relationships between independent and dependent variables.

Additional exploration was conducted to compare maximum likelihood empirical Bayes (ML-EB) estimation with fully Bayesian estimation of effects using the multilevel model. It was expected that since the fully Bayesian approach better acknowledges uncertainty present in models estimated for small-sample data, that it would be more. Although estimates of fixed effects were found to be similar across methods (ML-EB vs. fully Bayesian), estimates of random effects (i.e., variance components) and standard errors of fixed effects were found to be slightly larger in the fully Bayesian estimated models than in the ML-EB estimated models in both demonstrations, illustrating this realistic acknowledgement of greater uncertainty. This indicates that the fully Bayesian estimation approach is a better choice in both demonstrations.

While visual inspection was found to be a useful and efficient way of detecting outliers or obvious patterns in the data, it is in concert with a sophisticated statistical method that it is best used. In the end, evidence from the demonstrated analyses supports the hypothesis for this study. In comparison to all other demonstrated models, a combination of visual inspection and multilevel modeling using fully Bayesian estimation was found to be the most suitable method for analyzing data from single case design studies. Using this method allows one to write a statistical model to summarize the average behavior pattern and treatment effect on that behavior pattern, to test whether there are differences among subjects in various aspects of their behavior, and if so, to test whether those differences can be explained by subject characteristics. Table 15 presents 10 steps for using this method for analyzing data from single case design studies.

Table 15: 10 Steps for Analyzing Data from Single Case Design Studies

1	Visually inspect graphed data for salient patterns and/or expected complications. Make preliminary predictions about findings.
2	Choose the correct distributional form for the outcome variable: ▪ Normal distribution for continuous variables without a floor or ceiling effect; ▪ Poisson distribution for count or rate outcomes without an upper limit; ▪ Binomial distribution for the number of successes out of a fixed number of trials or a proportion.
3	Construct Level-1 file (observations) in SPSS (or other database program) appropriately, including: ▪ Construct variables that represent the design of the study (e.g. in an ABAB study, construct variables to indicate baseline or treatment *phase*, first AB pair or second AB pair *order*, etc). ▪ Construct a variable for time (e.g., session number) and center appropriately (i.e., subtract an appropriate amount to represent a reasonable change point, such as the last session of the first A phase or the first session of the first B phase in an ABAB design). ▪ Compute all two- and three-way interaction terms, as appropriate to the design, by multiplying single effects (e.g., time*phase, phase*order, etc). ▪ Construct variables required by distributional form for the outcome variable (e.g., total number of trials for proportions as modeled by the binomial distribution).
3	Construct Level-2 file (subject characteristics) in SPSS (or other database program) appropriately, including: ▪ Linking subject identifiers to Level-1 file. ▪ Centering subject characteristics (e.g., at grand mean) to prepare for more meaningful interpretation.
4	Import data into HLM (or other multilevel modeling software). Select appropriate model type/distribution. Indicate all possible Level-1 and Level-2 variables for availability for model building.
5	Run full multilevel model with all Level-1 parameters included (i.e., all possible main effects and two- and three-way interaction effects). For ABAB designs, this may include: intercept, time, phase, order, time*phase, phase*order, time*phase*order. Pull out non-significant fixed effects one-by-one, until only significantly contributing Level-1 terms remain. [Start with the term with the p-value furthest from significance (i.e., highest), being careful to keep in main effects when interactions including those terms remain.]
6	As indicated by variance components, enter Level-2 variables into the model to try to explain any significant variance that remains on any Level-1 effect, as appropriate.
7	Once final model is identified in HLM, re-run model in WinBUGS (or other software) to estimate effects via fully Bayesian estimation.
8	Translate/interpret terms from all models as needed.
9	Graph final model(s) over original data and examine fit visually.
10	Compare parameters, standard errors, etc. and decide on best suited model.

While this study is strong, it is not without limitations. In order to make stable comparisons across models, the methods that were demonstrated here were all reframed as special cases of the multilevel model. However, using multilevel models in this way is not exactly the same as duplicating the alternative methods, since multilevel modeling software accommodates technicalities of the data not assumed by all alternative methods (e.g., computing the statistical significance of between-subjects variance on estimates). Also, in doing so, several of the methods reviewed in the literature could not be carried out in the demonstrations because their underlying models did not align with the regression model underlying the multilevel modeling method.

In addition, only two datasets were employed for this demonstration. The simulated data were intentionally designed to include many of the unique aspects of single case design data. The published study dataset was selected for its realistic limitations in number of data points, number of subjects, and variability within. However, the statistical impacts of one of the most considered technicalities of single case design data, autocorrelation, could not be demonstrated for either dataset because, even though a moderate level of autocorrelation was included in the simulated data, little to no autocorrelation existed in either actual dataset.

Lastly, only two estimation approaches were discussed and demonstrated here. Beyond ML-EB and fully Bayesian estimation, other estimation approaches exist and each approach can lead to somewhat different estimates of effects. Especially when sample sizes are small, as was the case in both demonstration datasets, it is important to consider which estimation method is best since the choice of estimation method can have implications for interpretation of the results. Demonstrating other estimation methods may have revealed slightly different estimation of effects.

The multilevel modeling method is not without limitations itself. As with any analysis method, the power of this approach in re-analyzing data is limited by the breadth of information reported by a study's original author. Reports of studies may not provide enough information about subjects or study characteristics to serve as potential explainers of variation in intercepts or slopes. Without such information, an analyst may not be able to explain enough of this variation, as was illustrated in the demonstration usng published data. Failure to explain substantial proportion of between-subject or between-study variance can hamper interpretation and practical utility of results. Also, the flexibility of the multilevel modeling approach leads to several choices in the model-building process. Although there is some protocol, it is essentially up to the analyst to make these decisions at each turning point. Thus, different models and conclusions (some appropriate, some not) may be drawn from one dataset. Finally, though this method is better able to handle data with fewer observations and even fewer subjects, the HLM software will not run for data on one single subject (N=1). Therefore, this method is not an option for the analysis of data from a true single subject study.

Future research should consider and build upon the work presented here. Additional replications should be conducted to compare the performance of these methods for analyzing data from single case design studies, including data with more severe autocorrelation. Datasets from studies using more complex designs (e.g., multiple treatments, multiple baselines) could be utilized to continue testing the flexibility of these methods. Additional statistical models may also be demonstrated, perhaps under a different framework, one other than the multilevel model. Other estimation approaches could be demonstrated and compared to illustrate implications on model estimates.

In addition, model comparison methodologies could be used to help make comparisons between models even more straightforward. Akaike's information criterion, known as AIC, is such a tool for comparing models of the same data. It ranks competing models according to the information lost when each is used to describe data, also taking into account parsimony of the model when assigning an AIC value. The AICc method includes a correction term for small sample size, suitable for comparing models of single case design datasets (Burnham & Anderson, 2004).

This paper has illustrated the many ways in which standard methods proposed for analyzing single case design data often fall short and that the additional complexity of the multilevel modeling approach is sometimes necessary. In an article by the APA Task Force on Statistical Inference (1999), guidelines for choosing a minimally sufficient analysis are presented as follows:

> "Although complex designs are sometimes necessary to address research questions effectively, simpler classical approaches can often provide elegant and sufficient answers to important questions. … If the assumptions and strength of a simpler method are reasonable for your data and research problem, use it. Occam's razor applies to methods as well as theories (Wilkinson et al, 1999, p. 598)."

While it's true that we should not always skip straight to the most sophisticated method without reason, in some cases, more complex methods are indeed necessary to appropriately accommodate the realities of the data and/or design. Some of the standard methods proposed for analyzing single case design data are missing essential parameters that limit their ability to express patterns in the data. These shortcomings were known or at least suspected before analyses were demonstrated here. However, the demonstration analyses presented have shown just how insufficient these approaches can be. Single case design data can provide valuable information about the effects of an intervention on individuals, but only if analyzed appropriately

to take full advantage of the richness of the data. Of the methods compared here, the combination of visual inspection and multilevel modeling with fully Bayesian estimation has been shown to provide for the most appropriate, rich, and rigorous analysis and interpretation of data from single case designs.

Appendix

SPSS Syntax for Simulating Data (Demonstration #1)

```
**generate 161 empty cases.
input program.
loop #i = 1 to 161.
end case.
end loop.
end file.
end input program.

**compute IV's.
COMPUTE case = $casenum – 1.
COMPUTE id = 1 + trunc((case - 1)/20).
COMPUTE day = case - 20*(id-1).
COMPUTE order = trunc((day-1)/10).
COMPUTE phase = (trunc((day-1)/5)) - 2*order.
COMPUTE ph.ord = phase * order.
COMPUTE ph.day = phase * day.
COMPUTE day.ord = day * order.
COMPUTE ph.day.ord = phase * day * order.
COMPUTE err = rv.normal(0,5).
Execute.

**compute AR
COMPUTE e.25 = rv.norm(0,5).
COMPUTE f.25 = e.25 + .268 * lag(e.25).
COMPUTE f.25.1 = lag(f.25).
Execute.

**compute DV with AR =.25
if(id=1) y1 = 80 -2*day -20*phase +5*order -10*ph.ord + f.25.
if(id=2) y1 = 75 -2*day -17*phase +6*order - 6*ph.ord + f.25.
if(id=3) y1 = 70 -2*day -22*phase +5*order -11*ph.ord + f.25.
if(id=4) y1 = 82 -2*day -22*phase +6*order - 9*ph.ord + f.25.
if(id=5) y1 = 85 -2*day -27*phase +4*order - 8*ph.ord + f.25.
if(id=6) y1 = 90 -2*day -35*phase +5*order - 7*ph.ord + f.25.
if(id=7) y1 = 85 -2*day -32*phase +4*order -12*ph.ord + f.25.
if(id=8) y1 = 87 -2*day -31*phase +6*order -12*ph.ord + f.25.
Execute.

*center day
COMPUTE day.5 = day-5.
Execute.

**graph data created
IGRAPH /VIEWNAME='Scatterplot' /X1 = VAR(day) TYPE = SCALE /Y = VAR(y1) TYPE = SCALE
/STYLE = VAR(phase) /PANEL =  VAR(id)
 /COORDINATE = VERTICAL  /X1LENGTH=3.0 /YLENGTH=3.0 /X2LENGTH=3.0
/CHARTLOOK='C:\Program Files\SPSS\Looks\Grayscale.clo'
 /CATORDER VAR(id) (ASCENDING VALUES OMITEMPTY) /CATORDER VAR(phase) (ASCENDING
VALUES OMITEMPTY) /SCATTER COINCIDENT = NONE.
EXE.
```

WinBUGS input code for Simulated Data (Demonstration #1)

```
model
{  for (i in 1:103)
   { logit(p[i]) <- base[subj[i]] + trt[subj[i]] * phase[i]
      r[i] ~ dbin(p[i],10) }

   for (j in 1:4)
     {  base[j] ~ dnorm(mu0, prec0)
        trt[j] ~ dnorm(mu1, prec1) }

    mu0 ~ dnorm(0,.001)
    mu1 ~ dnorm(0,.001)
    prec0 ~ dgamma(.01,.01)
    prec1 ~ dgamma(.01,.01)
    sd0 <- sqrt(1/prec0)
    sd1 <- sqrt(1/prec1)
}

#  initial values
list(mu0=5,mu1=0,prec0=1,prec1=1)
```

subj[ ]	r[ ]	phase[ ]
1	4	0
1	5	0
1	5	0
1	4	0
1	5	0
1	10	1
1	8	1
1	10	1
1	7	1
1	10	1
1	10	1
1	10	1
1	9	1
1	10	1
1	9	1
1	10	1
1	9	1
1	5	0
1	5	0
1	6	0
1	6	0
1	10	1
1	10	1
1	9	1

1	9	1
1	9	1
2	2	0
2	3	0
2	3	0
2	5	0
2	3	0
2	10	1
2	9	1
2	10	1
2	10	1
2	10	1
2	8	1
2	7	1
2	8	1
2	9	1
2	8	1
2	9	1
2	9	1
2	6	0
2	4	0
2	3	0
2	9	1
2	8	1
2	10	1
2	10	1
2	9	1
3	6	0
3	6	0
3	7	0
3	6	0
3	6	0
3	10	1
3	9	1
3	10	1
3	9	1
3	9	1
3	10	1
3	10	1
3	9	1
3	9	1
3	9	1
3	9	1
3	9	1
3	7	0
3	7	0

```
3       6       0
3       6       0
3       10      1
3       9       1
3       10      1
3       9       1
3       9       1
4       4       0
4       3       0
4       3       0
4       4       0
4       5       0
4       5       1
4       6       1
4       8       1
4       8       1
4       9       1
4       7       1
4       8       1
4       10      1
4       9       1
4       9       1
4       8       1
4       9       1
4       4       0
4       5       0
4       5       0
4       3       0
4       8       1
4       8       1
4       7       1
4       9       1
4       9       1
END DATA
```

WinBUGS input code for Published Data (Demonstration #2)

```
model
{  for (i in 1:160)
   { mu[i] <- b0[subj[i]] + b1[subj[i]]*L2int[i] + b2[subj[i]]*day[i] + b3[subj[i]]*trt[i] +
b4[subj[i]]*trt[i]*L2trt[i] + b5[subj[i]]*daytrt[i]
     y[i] ~ dnorm(mu[i],precy)
   }

   for (j in 1:8)
    {    b0[j] ~ dnorm(mu0, prec0)
         b1[j] ~ dnorm(mu1, prec1)
                        b2[j] ~ dnorm(mu2, prec2)
                        b3[j] ~ dnorm(mu3, prec3)
                        b4[j] ~ dnorm(mu4, prec4)
                        b5[j] ~ dnorm(mu5, prec5)  }

          sdy <- sqrt(1/precy)
          precy ~ dgamma(.01,.01)
    mu0 ~ dnorm(0,.001)
    mu1 ~ dnorm(0,.001)
    mu2 ~ dnorm(0,.001)
    mu3 ~ dnorm(0,.001)
    mu4 ~ dnorm(0,.001)
    mu5 ~ dnorm(0,.001)
    prec0 ~ dgamma(.01,.01)
    prec1 ~ dgamma(.01,.01)
    prec2 ~ dgamma(.01,.01)
    prec3 ~ dgamma(.01,.01)
    prec4 ~ dgamma(.01,.01)
    prec5 ~ dgamma(.01,.01)
    sd0 <- sqrt(1/prec0)
    sd1 <- sqrt(1/prec1)
    sd2 <- sqrt(1/prec2)
    sd3 <- sqrt(1/prec3)
    sd4 <- sqrt(1/prec4)
    sd5 <- sqrt(1/prec5)
}

#  initial values
list(mu0=50, mu1=0, mu2=-5, mu3=-10, mu4=0, mu5=-5, prec0=0.1, prec1=0.1,
prec2=0.2, prec3=0.1, prec4=0.1, prec5=0.1)
```

subj[ ]	y[ ]	day[ ]	trt[ ]	daytrt[ ]	L2int[ ]	L2trt[ ]
1	71.58	-4	0	0	0	2
1	78.24	-3	0	0	0	2

1	79.78	-2	0	0	0	2
1	65.97	-1	0	0	0	2
1	71.46	0	0	0	0	2
1	48.40	1	1	1	0	2
1	44.12	2	1	2	0	2
1	45.24	3	1	3	0	2
1	49.54	4	1	4	0	2
1	34.88	5	1	5	0	2
1	68.14	6	0	0	0	2
1	68.80	7	0	0	0	2
1	70.82	8	0	0	0	2
1	62.56	9	0	0	0	2
1	53.19	10	0	0	0	2
1	27.17	11	1	11	0	2
1	34.11	12	1	12	0	2
1	12.82	13	1	13	0	2
1	17.61	14	1	14	0	2
1	9.84	15	1	15	0	2
2	67.40	-4	0	0	-2	3
2	69.73	-3	0	0	-2	3
2	69.58	-2	0	0	-2	3
2	65.68	-1	0	0	-2	3
2	57.86	0	0	0	-2	3
2	41.20	1	1	1	-2	3
2	49.38	2	1	2	-2	3
2	43.98	3	1	3	-2	3
2	50.95	4	1	4	-2	3
2	38.71	5	1	5	-2	3
2	52.69	6	0	0	-2	3
2	51.56	7	0	0	-2	3
2	58.85	8	0	0	-2	3
2	62.06	9	0	0	-2	3
2	53.41	10	0	0	-2	3
2	22.43	11	1	11	-2	3
2	26.31	12	1	12	-2	3
2	20.70	13	1	13	-2	3
2	12.85	14	1	14	-2	3
2	18.22	15	1	15	-2	3
3	63.41	-4	0	0	-3	1
3	67.57	-3	0	0	-3	1
3	65.59	-2	0	0	-3	1
3	68.18	-1	0	0	-3	1
3	61.15	0	0	0	-3	1
3	40.92	1	1	1	-3	1
3	37.70	2	1	2	-3	1
3	29.29	3	1	3	-3	1

3	32.65	4	1	4	-3	1
3	33.09	5	1	5	-3	1
3	53.88	6	0	0	-3	1
3	51.14	7	0	0	-3	1
3	47.35	8	0	0	-3	1
3	53.75	9	0	0	-3	1
3	45.16	10	0	0	-3	1
3	11.95	11	1	11	-3	1
3	6.17	12	1	12	-3	1
3	3.94	13	1	13	-3	1
3	10.34	14	1	14	-3	1
3	2.18	15	1	15	-3	1
4	87.41	-4	0	0	0	1
4	79.99	-3	0	0	0	1
4	72.92	-2	0	0	0	1
4	72.91	-1	0	0	0	1
4	82.59	0	0	0	0	1
4	53.55	1	1	1	0	1
4	51.41	2	1	2	0	1
4	51.08	3	1	3	0	1
4	48.15	4	1	4	0	1
4	43.63	5	1	5	0	1
4	67.58	6	0	0	0	1
4	69.85	7	0	0	0	1
4	60.77	8	0	0	0	1
4	59.96	9	0	0	0	1
4	61.11	10	0	0	0	1
4	34.57	11	1	11	0	1
4	30.42	12	1	12	0	1
4	24.32	13	1	13	0	1
4	15.70	14	1	14	0	1
4	19.68	15	1	15	0	1
5	79.69	-4	0	0	1	-1
5	81.50	-3	0	0	1	-1
5	79.20	-2	0	0	1	-1
5	78.16	-1	0	0	1	-1
5	72.37	0	0	0	1	-1
5	51.06	1	1	1	1	-1
5	49.65	2	1	2	1	-1
5	39.83	3	1	3	1	-1
5	40.76	4	1	4	1	-1
5	40.15	5	1	5	1	-1
5	64.60	6	0	0	1	-1
5	64.23	7	0	0	1	-1
5	70.96	8	0	0	1	-1
5	67.72	9	0	0	1	-1

5	58.42	10	0	0	1	-1
5	23.25	11	1	11	1	-1
5	21.50	12	1	12	1	-1
5	27.51	13	1	13	1	-1
5	18.23	14	1	14	1	-1
5	6.05	15	1	15	1	-1
6	78.56	-4	0	0	3	-3
6	86.81	-3	0	0	3	-3
6	85.87	-2	0	0	3	-3
6	87.62	-1	0	0	3	-3
6	79.55	0	0	0	3	-3
6	35.28	1	1	1	3	-3
6	40.82	2	1	2	3	-3
6	43.52	3	1	3	3	-3
6	36.24	4	1	4	3	-3
6	38.03	5	1	5	3	-3
6	71.63	6	0	0	3	-3
6	69.32	7	0	0	3	-3
6	64.52	8	0	0	3	-3
6	74.52	9	0	0	3	-3
6	79.14	10	0	0	3	-3
6	22.69	11	1	11	3	-3
6	17.05	12	1	12	3	-3
6	15.05	13	1	13	3	-3
6	11.77	14	1	14	3	-3
6	20.12	15	1	15	3	-3
7	81.95	-4	0	0	1	-2
7	78.31	-3	0	0	1	-2
7	78.07	-2	0	0	1	-2
7	86.34	-1	0	0	1	-2
7	73.06	0	0	0	1	-2
7	35.19	1	1	1	1	-2
7	48.40	2	1	2	1	-2
7	34.54	3	1	3	1	-2
7	27.46	4	1	4	1	-2
7	25.60	5	1	5	1	-2
7	67.63	6	0	0	1	-2
7	66.95	7	0	0	1	-2
7	61.18	8	0	0	1	-2
7	71.81	9	0	0	1	-2
7	69.34	10	0	0	1	-2
7	17.40	11	1	11	1	-2
7	11.84	12	1	12	1	-2
7	3.32	13	1	13	1	-2
7	.33	14	1	14	1	-2
7	6.87	15	1	15	1	-2

```
8       86.23 -4      0       0       2       -2
8       75.37 -3      0       0       2       -2
8       80.38 -2      0       0       2       -2
8       81.60 -1      0       0       2       -2
8       78.60 0       0       0       2       -2
8       48.07 1       1       1       2       -2
8       43.79 2       1       2       2       -2
8       35.52 3       1       3       2       -2
8       26.80 4       1       4       2       -2
8       29.93 5       1       5       2       -2
8       67.27 6       0       0       2       -2
8       70.68 7       0       0       2       -2
8       73.76 8       0       0       2       -2
8       70.18 9       0       0       2       -2
8       60.87 10      0       0       2       -2
8       14.38 11      1       11      2       -2
8       17.26 12      1       12      2       -2
8       15.58 13      1       13      2       -2
8       15.43 14      1       14      2       -2
8       6.59  15      1       15      2       -2
END DATA
```

References

Afshartous, D. & de Leeuw, J. (2005). Prediction in multilevel models. *Journal of Educational and Behavioral Statistics, 30*(2), 109-139.

Agresti, A. (2002). *Categorical Data Analysis* (2nd ed.). New York: John Wiley & Sons, Inc.

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

Allison, D. B., & Gorman B. S. (1994). "Make things as simple as possible, but no simpler." A rejoinder to Scruggs and Mastropieri. *Behaviour Research & Therapy, 32*, 885-890.

Allison, D. B., & Gorman, B. S. (1993). Calculating effect sizes for meta-analysis: The case of the single case. *Behaviour Research & Therapy, 31*, 621-631.

Barlow, D.H. & Hersen, M. (Eds.) (1984). *Single-case experimental designs: Strategies for studying behavior change (2nd edition).* New York: Pergamon Press.

Baron, A., & Derenne, A. (2000). Quantitative summaries of single case studies: What do group comparisons tell us about individual performances? *The Behavior Analyst, 23*, 101-106.

Blumberg, C.J. (1984). Comments on "A simplified time-series analysis for evaluating treatment interventions". *Journal of Applied Behavior Analysis, 17*(4), 539-542.

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*(5), 531-563.

Browne, W.J. & Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis, 1*(3), 473-514.

Burnham, K.P. & Anderson, D.R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods in Research, 33*(2), 261-304.

Busk, P.L. & Marascuilo, L.A. (1992). Statistical Analysis in Single-Case Research: Issues, Procedures, and Recommendations, with Applications to Multiple Behaviors. In T.R. Kratochwill & J.R. Levin (Eds.), *Single-Case Research Design and Analysis: New Directions for Psychology and Education* (pp. 159-186). Hillsdale: Lawrence Erlbaum Associates.

Busk, P.L. & Serlin, R.C. (1992). Meta-analysis for Single-Case Research. In T.R. Kratochwill & J.R. Levin (Eds.), *Single-Case Research Design and Analysis: New Directions for Psychology and Education* (pp. 197-212). Hillsdale: Lawrence Erlbaum Associates.

Busse, R. T., Kratochwill, T. R., & Elliott, S. N. (1995). Meta-analysis for single-case consultation outcomes: Applications to research and practice. *Journal of School Psychology, 33*, 269-285.

Campbell, J.M. (2003). Efficacy of behavioral intervention for reducing problem behavior in persons with autism: A quantitative synthesis of single case research. *Research in Developmental Disabilities, 24*, 120-138.

Center, B. A., Skiba, R. J., & Casey, A. (1985). A methodology for the quantitative synthesis of intra-subject design research. *Journal of Special Education, 19*, 387-400.

Critchfield, T. S., Newland, M. C., & Kollins, S. H. (2000). The good, the bad, and aggregate. *The Behavior Analyst, 23*, 107-115.

Crosbie, J. (1993). Interrupted time-series analysis with brief single case data. *Journal of Consulting and Clinical Psychology, 61*, 996-974.

DeProspero, A. & Cohen, S. (1979). Inconsistent visual analysis of intrasubject data. *Journal of Applied Behavior Analysis, 12*(4), 573-579.

Edgington, E.S. (1992). Nonparametric 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* (pp. 133-158). Hillsdale: Lawrence Erlbaum Associates.

Edgington, E.S. (1980). Validity of randomization tests for one-subject experiments. *Journal of Educational Statistics, 5*(3), 235-251.

Edgington, E.S. & Onghena, P. (2007). *Randomization tests* (4th ed.). Boca Raton, FL: Chapman & Hall.

Franklin, R.D. Allison, D.B., & Gorman, B.S. (1996a). *Design and Analysis of Single-Case Research.* Mahwah, NJ: Lawrence Erlbaum Associates.

Franklin, R.D. Allison, D.B., & Gorman, B.S. (1996b). Meta-analysis of single-case research. In R.D. Franklin, D.B. Allison, & B.S. Gorman, *Design and Analysis of Single-Case Research* (pp. 244-278). Mahwah, NJ: Lawrence Erlbaum Associates.

Franklin, R.D. Allison, D.B., Beasley, T.M. & Gorman, B.S. (1996). Graphical display and visual analysis. In R.D. Franklin, D.B. Allison, & B.S. Gorman, *Design and Analysis of Single-Case Research* (pp. 119-158). Mahwah, NJ: Lawrence Erlbaum Associates.

Gelman, A. & Hill, J. (2007). *Data Analysis using Regression and Multilevel/Hierarchical Models*. New York: Cambridge University Press.

Gentile, J.R., Roden, A.H., & Klein, R.D. (1972). An analysis-of-variance model for the intrasubject replication design. *Journal of Applied Behavior Analysis, 5*(2), 193-198.

Gingerich, W.J. (1984). Meta-analysis of applied time-series data. *Journal of Applied Behavioral Science, 20,* 71-79.

Glass, G.V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*, 3-8.

Glass, G.V., McGaw, B., & Smith, M.L. (1981). *Meta-Analysis in Social Research.* Beverly Hills, CA: Sage Publications.

Gorman, B.S. & Allison, D.B. (1996). Statistical alternatives for single-case designs. In R.D. Franklin, D.B. Allison, & B.S. Gorman, *Design and Analysis of Single-Case Research* (pp. 159-214). Mahwah, NJ: Lawrence Erlbaum Associates.

Gorsuch, R. L. (1983). Three methods for analyzing limited time-series (*N* of 1) data. *Behavioral Assessment, 5*, 141-154.

Gottman, J. M. (1981). *Time-series analysis: A comprehensive introduction for social scientists.* Cambridge: Cambridge University Press.

Horner, R.H., Carr, E.G., Halle, J., McGee, G., Odam, S., & Wolery, M. (2005). The use of single case research to identify evidence-based practice in special education. *Exceptional Children, 7*(2), 165-179.

Hoyle, R.H. (1999). *Statistical Strategies for Small Sample Research.* Thousand Oaks, CA: Sage Publications.

Huitema, B.E. (1985). Autocorrelation in applied behavior analysis: A myth. *Behavioral Analysis, 7*, 107-110.

Huitema, B.E. & McKean, J.W. (2000). Design specification issues in time-series intervention models. *Educational and Psychological Measurement, 60*(1), 38-58.

Jenson, W.R., Clark, E., Kircher, J.C. & Kristjansson, S.D. (2007). Statistical reform: Evidence-based practice, meta-analyses, and single case designs. *Psychology in the Schools, 44*(5), 483-493.

Kavale, K. A., Mathur, S. R., Forness, S. R., Quinn, M. M., & Rutherford, R. B. (2000). Right reason in the integration of group and single case research in behavioral disorders. *Behavioral Disorders*, 25, 142-157.

Kollins, S.H. Newland, M.C., & Critchfield, T.S. (1999). Quantitative integration of single case studies: Methods and misinterpretations. *Behavior Analyst, 22,* 149-157.

Kratochwill, T.R. (1992). Single-Case Research Design and Analysis: An Overview. In T.R. Kratochwill & J.R. Levin (Eds.), *Single-Case Research Design and Analysis: New Directions for Psychology and Education* (pp. 1-14). Hillsdale: Lawrence Erlbaum Associates.

Kratochwill, T.R. & Levin, J.R. (Eds.). (1992). *Single-Case Research Design and Analysis: New Directions for Psychology and Education*. Hillsdale: Lawrence Erlbaum Associates.

Kromrey, J.D. & Foster-Johnson, L. (1996). Determining the efficacy of intervention: The use of effect sizes for data analysis in single case research. *Journal of Experimental Education, 65*(1), 73-84.

Lambert, M.C., Cartledge, G., Heward, W.L., & Lo, Y. (2006). Effects of response cards on disruptive behavior and academic responding during math lessons by fourth-grade urban students. *Journal of Positive Behavior Interventions, 8*(2), 88-99.

Light, R.J. & Pillemer, D.B. (1984). *Summing up: The science of reviewing research.* Cambridge, MA: Harvard University Press.

Ma, H-H. (2006). An alternative method for quantitative synthesis of single case researches: Percentage of data points exceeding the median. *Behavior Modification, 30*(5), 598-617.

Matyas, K.M. & Gorman, B.S. (1996). Serial dependency in single-case time series. In R.D. Franklin, D.B. Allison, & B.S. Gorman, *Design and Analysis of Single-Case Research* (pp. 215-243). Mahwah, NJ: Lawrence Erlbaum Associates.

McCleary, R. & Wayne, W.N. (1992). Philosophical and statistical foundations of time-series experiments. In T.R. Kratochwill & J.R. Levin (Eds.), *Single-Case Research Design and Analysis: New Directions for Psychology and Education* (pp. 41-92). Hillsdale: Lawrence Erlbaum Associates.

Nagler, E. (2007). *Analyzing data from small-N designs using multilevel models: A demonstration.* Unpublished Manuscript.

Nagler, E., Rindskopf, D., & Shadish, W. (2006). *Analyzing data from small N designs using multilevel models: A procedural handbook.* Unpublished Manuscript.

New York City Department of Education (2007). *Mayor's Management Report: New York City Department of Education*. Retrieved January 5, 2008 from http://www.nyc.gov/html/ops/downloads/pdf/_mmr/doe.pdf

Nugent, W.R. (1996). Integrating single-case and group-comparison designs for evaluation research. *Journal of Applied Behavioral Science, 32*(2), 209-226.

Onghena, P. & Edgington, E.S. (1994). Randomization tests for restricted alternating treatments designs. *Behaviour Research & Therapy, 32*(7), 783-786.

Ottenbacher, K.J. (1990). When is a picture worth a thousand *p* values? A comparison of visual and quantitative methods to analyze single case data. *The Journal of Special Education, 23*(4), 436-449.

Owen, D.B. (1987). Some comments concerning "The quantitative synthesis of single case research." *Remedial and Special Education, 8*(2), 34-39.

Parker, R.I. & Brossart, D.F. (2003). Evaluating single-case research data: A comparison of seven statistical methods. *Behavior Therapy, 34*, 189-211.

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 (pp. 15-40). Hillsdale: Lawrence Erlbaum Associates.

Raudenbush, S., Bryk, A., Cheong, Y.F., Congdon, R., & du Toit, M. (2004). *HLM6: Hierarchical Linear and Nonlinear Modeling.* Lincolnwood, IL: Scientific Software International, Inc.

Raudenbush, S.W. & Bryk, A.S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods, Second Edition.* Newbury Park, CA: Sage.

Raudenbush, S.W. & Bryk, A.S. (1985). Empirical Bayes meta-analysis. *Journal of Educational Statistics, 10*(2), 75-98.

Richards, S.B., Taylor, R.L., Ramasamy, R., & Richards, R.Y. (1999). *Single subject research: applications in educational and clinical settings.* Belmont, CA: Wadsworth Group/Thomson Learning.

Robbins, N.B. (2005). *Creating more effective graphs.* Hoboken, NJ: John Wiley & Sons.

Salzberg, C.L., Strain, P.S., & Baer, D.M. (1987). Meta-analysis for single case research: When does it clarify, when does it obscure? *Remedial and Special Education, 8*(2), 43-48.

Schlosser, R.W. (2005). Meta-analysis of single case research: how should it be done? *International Journal of Language & Communication Disorders, 40*(3), 375-378.

Schreiber, J.B. & Griffin, B.W. (2004). Review of multilevel modeling and multilevel studies in *The Journal of Educational Research* (1992-2002). *The Journal of Educational Research, 98*(1), 24-33.

Scruggs, T.E. & Mastropieri, M.A. (1998). Summarizing single case research: Issues and applications. *Behavior Modification, 22*(3), 221-242.

Scruggs, T.E. & Mastropieri, M.A. (1994). The utility of the PND statistic: A reply to Allison and Gorman. *Behaviour Research and Therapy, 32*(8), 879-883.

Scruggs, T.E., Mastropieri, M.A., & Casto, G. (1987a). The quantitative synthesis of single case research: Methodology and validation. *Remedial and Special Education, 8*(2), 24-33.

Scruggs, T.E., Mastropieri, M.A., & Casto, G. (1987b). Reply to Owen White. *Remedial and Special Education, 8*(2), 40-42.

Scruggs, T.E., Mastropieri, M.A., & Casto, G. (1987c). Response to Salzberg, Strain, and Baer. *Remedial and Special Education, 8*(2), 49-52.

Shadish, W.R., Brasil, I.C., Illingworth, D.A., White, K.D., Galindo, R., Nagler, E.D., & Rindskopf, D.M. (2009). Using UnGraph to extract data from image files: Verification of reliability and validity. *Behavior Research Methods, 41*(1), 177-183.

Shadish, W.R. & Rindskopf, D.M. (2007). Methods for evidence-based practice: Quantitative synthesis of single case designs. In G. Julnes & D.J. Rog (Eds.), *Informing Federal Policies on Evaluation Methodology: Building the Evidence Base for Method Choice in Government Sponsored Evaluation* (pp. 95-109). San Francisco: Wiley Periodicals.

Sharpley, C.F. & Alavosius, M.P. (1988). Autocorrelation in behavioral data: An alternative persepective. *Behavioral Assessment, 10*, 243-251.

Shine, L.C. & Bower, S.M. (1971). A one-way analysis of variance for single case designs. *Educational and Psychological Measurement, 31*, 105-113.

Strain, P.S., Kohler, F.W., & Gresham, F. (1998). Problems in logic and interpretation with quantitative synthesis of single-case research: Mathur and colleagues (1998) as a case in point. *Behavioral Disorders, 24*, 74-85.

Swanson, H.L. & Sachse-Lee, C. (2000). A meta-analysis of single case design intervention research for students with LD. *Journal of Learning Disabilities, 33*, 114-136.

Suen, H.K. (1987). On the epistemology of autocorrelation in applied behavior analysis. *Behavioral Assessment, 9*, 113-124.

Thoresen, C.E. & Elashoff, J.D. (1974). "An analysis-of-variance model for intrasubject replication design": Some additional comments. *Journal of Applied Behavior Analysis, 7*(4), 639-641.

Todman, J.B. and Dugard, P. (2001). *Single-case and small-n experimental designs: A practical guide to randomization tests.* Mahway, N.J.: Lawrence Erlbaum Associates.

Tryon, W.W. (1984). "A simplified time-series analysis for evaluating treatment interventions": A rejoinder to Blumberg. *Journal of Applied Behavior Analysis, 17*(4), 543-544.

White, O.R. (1987). Some comments concerning: "The quantitative synthesis of single case research: Methodology and validation". *Remedial and Special Education, 8*(2), 34-39.

White, D. M., Rusch, F. R., Kazdin, A. E., & Hartmann, D. P. (1989). Applications of meta analysis in individual-subject research. *Behavioral Assessment, 11*, 281-296.

U.S. Department of Education. (2002). *No Child Left Behind Act of 2001*. Retrieved January 5, 2008, from http://www.ed.gov/policy/elsec/leg/esea02/107-110.pdf.

U.S. Department of Education. (2008). *WWC procedures and standards handbook, Version 2.0*. Retrieved June 2, 2010, from http://ies.ed.gov/ncee/wwc/pdf/wwc_procedures_v2_standards_handbook.pdf

U.S. Department of Education. (2010). *Race to the Top Fund*. Retrieved January 24, 2010, from http://www2.ed.gov/programs/racetothetop/index.html.

Van Belle, G. (2008). *Statistical rules of thumb*. Hoboken, NJ: John Wiley & Sons.

Van den Noortgate, W. & Onghena, P. (2007). The aggregation of single-case results using hierarchical linear models. *The Behavior Analyst Today, 8*(2), 196-208.

Van den Noortgate, W. & Onghena, P. (2003a). Hierarchical linear models for the quantitative integration of effect sizes in single-case research. *Behavior Research Methods, Instruments, and Computers, 35*(1), 1-10.

Van den Noortgate, W. & Onghena, P. (2003b). Combining single-case experimental data using hierarchical linear models. *School Psychology Quarterly, 18*(3), 325-346.

Velicer, W.F. (1994). Time series models of individual substance abusers. *NIDA Research Monographs, 142*, 264-301.

Velicer, W.F. & McDonald, R.P. (1994). Cross-sectional time series designs: A general transformation approach. *Multivariate Behavioral Research, 26*(2), 247-254.

Software

HLM (Version 6.03 for Windows) [Computer software]. Lincolnwood, IL: Scientific Software International, Inc.

SPSS (Version 12.0.1 for Windows) [Computer software]. Chicago, IL: SPSS, Inc.

UnGraph (Version 5.0) [Computer software]. Cambridge, UK: Biosoft.

WinBUGS (Verson 1.4.3) [Computer software]. Cambridge, UK: MRC Biostatistcs Unit, University of Cambridge.