# SIMULATION ANALYSIS AND META-ANALYSIS OF SINGLE-CASE EXPERIMENTAL DESIGNS

A Dissertation

by

KEVIN R. TARLOW

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

| | |
|---|---|
| Chair of Committee, | Daniel F. Brossart |
| Committee Members, | Kimberly J. Vannest |
| | Wen Luo |
| | Victor L. Willson |
| Head of Department, | Shanna Hagan-Burke |

August 2017

Major Subject: Counseling Psychology

# ABSTRACT

Single-case experimental designs remain outside of mainstream methodology despite their substantial contributions to our understanding of human behavior. An obstacle to wider adoption is the lack of consensus regarding the analysis and meta-analysis of single-case data. Many single-case statistical methods have been proposed; nearly all are limited by the incompleteness of their models or their lack of formal statistical development, both limitations that inhibit research synthesis and knowledge building.

This dissertation, presented in three manuscripts, introduces a simulation-based method of analysis and meta-analysis for single-case experimental designs. Interrupted Times-Series Simulation (ITSSIM) estimates treatment effect sizes by modeling level, trend, variance, and autocorrelation parameters. Parameter estimates are naturally imprecise in brief time-series. ITSSIM compensates for this imprecision by using an iterative procedure to model many plausible parameter values given the observed data. ITSSIM calculates an effect size by comparing a distribution of plausible "null effects"—the no-treatment predictions based on baseline data—to a distribution of plausible treatment effects. ITSSIM effect size estimates, reported as correlation coefficients, standardized mean differences, or unstandardized effects, are interpretable for both clinical practice and quantitative research synthesis.

Three studies provide evidence for the content validity, construct validity, and criterion validity of ITSSIM effect size estimates using theoretical, comparative, and

deductive strategies, respectively. The first study establishes the theoretical rationale for single-case simulation methods generally, and ITSSIM specifically. ITSSIM produced effect size estimates comparable to five sophisticated multilevel methods when a study of disruptive classroom behavior was reanalyzed. In the second study, ITSSIM produced mean effect size estimates consistent with similar meta-analyses of group-design research. The third study field tests a new software tool for simulation research of single-case statistics. ITSSIM performed reliably under a variety of simulation conditions, controlling for baseline trend and autocorrelation.

The results from these three studies indicate that ITSSIM is a powerful, comprehensive method for analysis and meta-analysis of single-case experimental designs. ITSSIM effect size estimates are consistent with other, previously published single-case statistics, and it yields reasonable results even under extreme simulation conditions. ITSSIM is recommended to single-case investigators who wish to better understand their single-case data and treatment effects.

# ACKNOWLEDGMENTS

Emily Dickinson said, “To live is so startling it leaves little time for anything else,” and I suppose one could say the same thing about writing a dissertation. Luckily, I had a legion of family, friends, mentors, and traveling companions on the journey that brought me—finally—here. The next few pages are addressed to you, Helpers. Acknowledging your support and (often invisible) contributions to this document seems inadequate to say the least. I will do my best to thank you each in person and support you in all the ways you supported me over the past thousand days of dissertating.

First, thank you to Dr. Brossart. Thank you for supporting me while I pursued a project that, even now, seems improbable. You politely listened to me think my simulation models out loud on too many occasions; you asked me the right questions when I was ready to answer them; you got excited about statistics with me. During one of our discussions you reminded me to “be a good scientist”. In that moment, your instruction was what I needed to hear most. I’ve returned to those works again and again since that meeting. Thanks to your guidance I will always, first and foremost, try to be a good scientist.

To my committee, Drs. Vannest, Luo, and Willson, thanks for tolerating my ridiculous 113-page dissertation proposal, and for graciously allowing me the opportunity to submit a (heavily edited) 10-page summary. Honestly, what was I thinking.

I owe a special acknowledgment to Dr. Bruce Thompson, the only statistics

professor I ever had. Thank you for being, to use your words, “emotionally and intellectually withholding.” In your first EPSY 640 lecture (my first stats class) you said that you wanted each student to complete our statistics training having surpassed our own initial expectations of our abilities. I still can’t believe I wrote a dissertation about time-series modeling and simulation. So I guess mission accomplished.

Thank you to Dr. Tim Elliott and Dr. Charles Ridley, who did not directly contribute to my dissertation research, but were nonetheless mentors to me during the past five years of graduate school. Dr. Elliott—thank you for your boundless wit and wisdom. Dr. Ridley, thank you for challenging me (and all your students) to always think deeply.

I am thankful to my work family at the Texas A&M University Student Counseling Service. As a graduate assistant and full-time student, y’all were my second job; as a predoctoral intern you are now my first job. But in five years you never expected the SCS to be my only job. So many individuals at the office have supported me in large ways and small ones while I worked on my dissertation research. Thank you everyone.

I am a grateful member of the community of scholars, healers, and dear friends in the Counseling Psychology Ph.D. program at Texas A&M University. I can’t imagine going through this without you. Special thanks to Angel Glover, Christina Jeffrey, Jeremy Saenz, and Ally Sequeira. Our friendship and late-night wine-filled conversations made the worst parts of grad school bearable, and the best parts better. I’m so excited for us all to be doctors very soon.

To Mom, Dad, and Zanne—there is no version of my life where I have a Ph.D. that doesn't also have you in it. You are my steady supporters, my source of strength, and my safe base. This dissertation is for all four of us.

Now, as the next 150 pages would suggest, I am rarely at a loss for words. However, nothing I write here can possibly capture how thankful I am for Mattie Squire, who, in addition to providing endless support, also agreed to marry me ten days from today. Mattie, thank you for being understanding and loving and encouraging and patient and a hundred thousand other things since I started this project. I simply would not have made it without you, and I can't wait to start our new life together. Our relationship is the best small-sample experiment I have ever participated in ($n = 2$).

—KRT (May 17, 2017)

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

# CHAPTER I

# INTRODUCTION

Single-case experimental designs are "poised for a resurgence" in psychological research (Smith, 2012, p. 510). Single-subject time-series experiments have a rich history in psychology, and they are used with increasing frequency in published studies. Many early discoveries in learning, behavior, and cognition were due to carefully controlled longitudinal studies of individuals (e.g., Ebbinghaus, 1885; Fechner, 1889; Jones, 1924; Piaget, 1952; Watson & Rayner, 1920; Skinner, 1938). Yet these designs have failed to achieve mainstream methodological status because of deep epistemological differences between the single-case design and groups sample designs, which have dominated the field since the mid-20th century (Baer, 1977; Morgan & Morgan, 2001).

Single-case experimental designs have many advantages. Similar to randomized controlled trials (RCTs), single-participant interrupted time-series experiments can demonstrate causal treatment effects (APA Presidential Task Force on Evidence-Based Practice, 2006; Barlow & Hersen, 1984; Campbell & Stanley, 1963). "Gold standard" RCTs have been criticized for producing results that are not generalizable to the idiosyncratic and idiographic nature of individualized treatment (Castelnuovo et al., 2004; Garfield, 1996; Seligman, 1995; Westen et al., 2004). However, practitioners in applied settings can conduct clinically relevant single-case designs (Hilliard, 1993; Jones et al., 1993; Persons & Silberschatz, 1998). Not only is there fidelity between single-case

designs and the settings in which most psychological interventions are delivered to consumers, but the studies can also be conducted with a fraction of the resources required of "large-*n*" RCTs (Barlow & Nock, 2009). Single-case research methods are also valuable to practitioners who wish to demonstrate their treatment efficacy, given the growing emphasis on evidence-based treatments in education and healthcare (Morgan & Morgan, 2001; Shadish et al., 2008).

## Validity of Single-Case Statistical Methods

Most single-case investigators evaluate treatment effects by visual inspection of graphed data (Brossart et al., 2006; Busk & Marascuilo, 1992; Kratochwill & Brody, 1978). Unfortunately, visual analysis is often unreliable, with different judges assigning different ratings to the same graphs (Danov & Symons, 2008; DeProspero & Cohen, 1979; Harbst, Ottenbacher, & Harris, 1991; Park, Marascuilo, & Gaylord-Ross, 1990; Lieberman et al., 2010; Ximenes, Manolov, Solanas, & Quera, 2009). Statistical methods, on the other hand, are not threatened by inter-rater reliability problems. As Franklin et al. (1996) stated, "Analysts using statistical techniques may disagree about the appropriate model for evaluating differences in dependent variables but, once applied, statistical models always provide the same result when calculated properly" (p. 150). Statistical methods are a reliable alternative to visual analysis, but reliability does not guarantee validity.

An obstacle to wider adoption of single-case methods is the lack of consensus regarding the statistical analysis of brief interrupted time-series single-case data. Large-sample clinical researchers use statistics to reliably quantify treatment effects and

describe participant characteristics. Single-case investigators also need to know if an experimental treatment is effective and how effective it is relative to other interventions. However, well-established time-series analysis methods (e.g., Box & Jenkins, 1970; Glass, Willson, & Gottman, 1975) require longer data sets than are typically available to single-case investigators.

In the absence of one agreed upon analytic approach, a host of single-case statistics have proliferated (Brossart et al., 2011; Campbell, 2004; Ma, 2006; Parker et al., 2005, 2006, 2007, 2011; Parker & Hagan-Burke, 2007; Manolov & Solanas, 2009, 2013; Manolov et al., 2011; Parker & Vannest, 2009; Parker, Vannest, & Davis, 2011; Shadish, Hedges, & Pustejovsky, 2014; Tarlow, in press; Vannest et al., 2012; Wolery et al., 2010). Most methods are similar in that they estimate treatment effects by comparing outcomes under a no-treatment condition (i.e., the baseline/A phase) with the outcomes from a treatment condition (i.e., the experimental/B phase). However, that similarity aside, these statistics differ in as many ways as there are methods. How statistics model interrupted times-series data—and the parameters that are included and excluded from analysis—is an important point of contrast between effect size indices. Some effect size indices quantify treatment effects in terms of average outcomes, or "last day" outcomes, or phase nonoverlap between A and B phases, or some other conceptual framework. Some methods account for participants' change over time beyond the effects of treatment, which could include baseline trend patterns or autocorrelated error structures (i.e., serial dependency). Some methods assume data are normally distributed (an assumption with convenient statistical implications), whereas other methods make no

such assumption. Some methods make direct A phase-to-B phase comparisons when calculating an effect size, whereas other methods contrast observed treatment outcomes (i.e., B phase data) with predicted "null effect" outcomes inferred from baseline phase observations. Unfortunately, it is rarely clear to the single-case investigator which parametric assumptions—and consequently, what statistical methods—are appropriate for any given data set. Given the diversity of data analytic approaches to single-case designs, it is not surprising that one data set may yield substantially differing estimates of treatment effect, depending on the effect size statistic used. Of course there is only one true treatment effect for an experiment, but it is rarely known which statistical model best represents reality.

If different statistics give different effect size estimates for the same experiment, then not all statistics can be valid for that experiment. Methodologists have tackled this problem with theoretical, comparative, and deductive strategies. Respectively, these strategies address the content validity, construct validity, and criterion validity of the single-case statistical methods.

**Content Validity**

A purely theoretical approach involves deep scrutiny of the assumptions underlying different statistics. Methods are rejected when their assumptions are deemed untenable for a given experimental design, i.e., they lack the content validity (or face validity) necessary for effect size estimation (e.g., Allison & Gorman, 1993; Baer, 1977; Tarlow, in press; Wolery et al., 2010).

**Construct validity**

In comparison-based strategies, the effect size estimates of multiple statistics are compared to determine how well different methods concur. When different statistics tend to yield the same conclusions (e.g., their effect size estimates are highly correlated), convergent validity (an aspect of construct validity) is conferred to the methods, or withheld in the case of disagreement (e.g., Brossart et al., 2006; Campbell, 2004; Crosbie, 1995; Parker & Brossart, 2003; Parker & Vannest, 2009; Parker, Vannest, & Davis, 2011).

**Criterion Validity**

The deductive approach often involves the use of artificial simulated single-case data sets with known parameters. A statistical method is used to estimate effect sizes for these hypothetical time-series, with an a priori hypothesis about the results based on the knowledge of simulation parameters. For example, a "null effect" interrupted-times series could be created in which the treatment had no effect on the participant's outcome; however, a baseline trend is included so that the outcome of the hypothetical participant is slowly improving throughout both phases of the experiment. In this example, the investigator can hypothesize that a valid statistical method should yield an effect size estimate of zero. Yet many single-case statistics fail to model baseline trend, and would thus yield nonzero positive treatment effect size estimates. These methods lack predictive validity (a type of criterion validity). This deductive approach has been used to "stress-test" single-case statistics to determine how violations of trend, autocorrelation, phase length, and distributional assumptions threaten the methods'

validity (Allison & Gorman, 1994; Crawford & Garthwaite, 2006; Gorsuch, 1983; Manolov & Solanas, 2008, 2009, 2012, 2013; Manolov et al., 2011; Matyas & Greenwood, 1990; Smith et al., 2012; Solanas, Manolov, & Onghena, 2010; Tarlow, in press; Ugille et al., 2012).

## Replication, Meta-Analysis, and Advanced Statistical Methods

Single-case research is intrinsically tied to experimental replication (Morgan & Morgan, 2001). Investigators wish to find effective treatment protocols for one individual, determine if those treatments work in other similar individuals, and when they fail to replicate, explore why not in an ideographic way. Single-case research is therefore epistemologically—but not always statistically—complementary with meta-analysis, i.e., the aggregation and synthesis of many research findings, or the "analysis of analyses" (Glass, 1976). There are many methods of estimating treatment effect sizes in single-case experiments, but only some of those methods permit the quantitative synthesis of results across cases and/or studies. Standard meta-analytic methods are not practical with many single-case effect size indices because those statistics lack the necessary formal statistical development—for example, their distributions may be unknown (e.g., unspecified standard errors), or they may fail to adequately model data parameters such as autocorrelation, which can influence the weights of combined effects (Shadish, 2014b; Shadish et al., 2008). Meta-analysis of single-case research is also complicated by an "apples and oranges" problem (Jenson et al., 2007). Effect sizes from single-case designs are often nominally larger than those for group design studies, so it is unclear how, and if, effects can be aggregated across both types of studies (see also

Parker et al., 2005).

Sophisticated multilevel methods have been applied to single-case data to model complex data structures and pool variance from stratified cases in order to improve the reliability of parameter estimates (Moeyaert et al., 2014; Rindskopf, 2014; Shadish, Hedges, & Pustejovsky, 2014; Shadish, Zuur, & Sullivan, 2014; Swaminathan et al., 2014; Van den Noortgate & Onghena, 2003, 2007, 2008). One limitation of multilevel modeling is the statistical sophistication and interpretive nuance required to conduct these analyses. Multilevel methods may not be a realistic option for some single-case investigators, who are often practitioners first and clinical scientists second (Parker & Vannest, 2012). Simpler statistical methods remain popular in single-case research, whereas complex ones (which may be more valid) are neglected due to their lack of utility (Schlosser et al., 2008).

As discussed above, computer-intensive simulation methods are often used to evaluate the criterion validity of single-case statistics; however, simulation models can themselves be used to estimate treatment effects. Simulation methods are based on bootstrapping. Wilcox (2001) stated, "The basic idea behind all bootstrap methods is to use the data obtained from a study to approximate the sampling distributions used to compute confidence intervals and test hypotheses" (p. 95). Boostrap methods offer a conceptual middle-ground between Bayesian estimation, which is based on applying prior beliefs to predictions, and frequentism, which emphasizes the probabilities of statistics under repeated use (Efron, 2013). In a bootstrap analysis, observed data are resampled many times in order to produce a sampling distribution for a statistical metric.

This strategy is useful because it does not make strict assumptions about the distribution underlying the observed data (as in a frequentist approach)—and in that way observed data are treated similarly to Bayesian priors. At the same time, creating a sampling distribution from repeated statistical measurements is an essentially frequentist strategy—but without the parametric assumptions (normality, etc.) which often vex frequentist methods. As Efron put it, "The bootstrap is a frequentist machine that produces Neyman-like confidence intervals far beyond the point where theory fails us" (p. 140).

Simulation Modeling Analysis (SMA), developed by Borckardt et al. (2008), is the only simulation-based method for single-case effect size estimation that has been widely adopted. SMA was developed to address the problem of autocorrelation estimation in single-case time-series. Autocorrelation can greatly distort effect size and probability estimates when not accounted for in many single-case statistics (Brossart et al., 2006; Crosbie, 1987; Ferron, 2002; Manolov & Solanas, 2008; Matyas & Greenwood, 1991). Unfortunately, autocorrelation estimation is unreliable in brief time-series (Huitema & McKean, 1991, 2000a; Solanas, Manolov, & Sierra, 2010). To address this problem, SMA uses the autocorrelation estimate from observed data to simulate thousands of artificial time-series. Simulated data are then used to create a sampling distribution for SMA's main effect size metric, a Pearson *r* correlation, which will yield probability estimates that are more accurate than standard *p* values. Simulation methods like SMA are useful because they can be easily implemented via user-friendly software and they do not require many data sets like multilevel modeling.

## Summary of the Problem

Single-case experimental designs are increasingly popular across many fields in education and psychology. These designs can aid in the development of clinically relevant evidence-based treatments, and they are accessible to a wide range of practitioners and applied scientists. However, the lack of agreed upon statistical methods is a barrier to wider adoption of single-case designs. While there are many proposed statistical methods for single-case effect size measurement, none have emerged as superior. Few of the available statistics yield effect size estimates appropriate for meta-analysis, further complicating quantitative synthesis and knowledge-building. Multilevel modeling is one area of recent innovation, but many investigators and research consumers may find those methods difficult to implement and interpret. Multilevel modeling also requires many single-case data sets, so it is also not a true "*N*-of-1" method. Computer-intensive simulation methods have been applied to single-case data analysis with promising results (Borckardt et al., 2008), though this area requires further development. New statistical methods, simulation-based or otherwise, should be evaluated with theoretical, comparative, and deductive strategies to determine their content validity, construct validity, and criterion validity, respectively.

## Plan for Studies

This dissertation, presented in three manuscripts, proposes a new simulation-based method of measurement and meta-analysis for single-case experimental designs. Interrupted Time-Series Simulation (ITSSIM) models a comprehensive set of data parameters, including level, slope, variance, and autocorrelation effects, and its effect

size estimates are based on empirical distributions of simulated data, similar to a bootstrap design. ITSSIM also yields a standardized mean difference effect size estimate that is appropriate for meta-analysis.

The first manuscript (Chapter II) presents the theoretical rationale for ITSSIM in order to provide evidence for content validity of its results. ITSSIM is also compared to five multilevel methods outlined in a special issue of the *Journal of School Psychology*, to determine if ITSSIM effect size estimates concur with the results from other sophisticated methods.

In the second manuscript (Chapter III), ITSSIM is used to meta-analyze ten recently published single-case studies of cognitive therapy for depression, with 53 total cases. The ITSSIM meta-analysis results are also compared to meta-analyses performed with five other (non-multilevel) statistical effect size indices. This study provides additional evidence of the construct validity of ITSSIM's effect size estimates.

The third manuscript (Chapter IV) presents a software tool, Interrupted Time Series Lab (ITSLAB), for conducting computer-intensive simulation research on single-case effect size statistics. ITSLAB is used to "stress test" ITSSIM (via "simulations of simulations") and five other single-case statistics under a range of simulation models. This study provides evidence of criterion validity by demonstrating how ITSSIM yields the results expected under various parametric assumptions.

# CHAPTER II

# A COMPREHENSIVE METHOD OF SINGLE-CASE DATA ANALYSIS: INTERRUPTED TIME-SERIES SIMULATION (ITSSIM)


## Abstract

Single-case experimental data are analyzed with a variety of statistical methods, but no one effect size measure has demonstrated clear superiority. The time-series nature of single-case designs requires special consideration for deterministic processes like baseline trend and autocorrelation when estimating treatment effect size. However, standard correction methods are limited because they assume perfectly precise statistical estimation. Two emerging approaches address the poor precision of single-case effect size indices: multilevel modeling and computer-intensive simulation. A new simulation-based method, Interrupted Time-Series Simulation (ITSSIM), is introduced and compared to multilevel methods. ITSSIM performed similarly to multilevel methods in a small field-test. It may be a useful option for single-case investigators because it yields easily interpretable effect size and reliability (standard error) estimates and it models several important data parameters, including baseline trend, level- and slope-change effects, and autocorrelation. ITSSIM is also accessible as a user-friendly standalone software application that requires no knowledge of statistical computing or syntax.


## Introduction

Single-case experimental designs are an increasingly popular method in

educational and clinical research (Smith, 2012). Single-case methods can resolve major challenges faced by investigators in applied behavioral, educational, and other social science fields. Case-based interrupted time-series experiments can demonstrate causal treatment effects, similar to randomized controlled trials (APA Presidential Task Force on Evidence-Based Practice, 2006; Barlow & Hersen, 1984). Because these designs are easy to implement in clinical settings, they can bridge the scientist-practitioner divide (Borckardt, 2008). They are also accessible to a broad range of investigators because they do not require the considerable resources of large sample studies (Barlow & Nock, 2009; Morgan & Morgan, 2001). An unresolved issue in single-case research is the lack of consensus about the analysis of brief interrupted time-series data. Most single-case studies evaluate data with visual analysis—a time-honored but often unreliable approach (Brossart, Parker, Olson, & Mahadevan, 2006; Danov & Symons, 2008; DeProspero & Cohen, 1979; Harbst, Ottenbacher, & Harris, 1991; Park, Marascuilo, & Gaylord-Ross, 1990; Lieberman et al., 2010; Ximenes, Manolov, Solanas, & Quera, 2009). Statistical methods are a useful complement to visual analysis; however, no single method of statistical analysis has demonstrated superiority over the other methods. This is a problem for investigators, as different statistical methods may lead to different conclusions about the same data.

An ideal single-case statistical method should meet several criteria. First, a single-case statistic should at a minimum yield an easy-to-interpret effect size for comparing the magnitudes of experimental treatment outcomes (Maggin & Odom, 2014). Second, a single-case statistic should quantify the precision of the effect size

estimate; for example, small highly reliable effects may be more desirable than large inconsistent effects. Reliability indicators, such as a standard error, are also useful for meta-analysis of multiple single-case experiments. Third, a good single-case statistic should model interrupted time-series data accurately; it should incorporate the data parameters which are known to affect participant responses over time. There has been a great deal of discussion about which parameters should be included in the statistical analysis of single-case data. The most discussed parameters involve deterministic processes which affect behavior over time, such as baseline trend and serial dependency (i.e., autocorrelation). Trend and autocorrelation are among the most vexing challenges to single-case statistical analysis (Tarlow, in press; Wampold, 1988). Fourth, analytic methods should be accessible to as wide a range of investigators as possible (Parker & Vannest, 2012; Shadish, 2014b). This includes practitioners and applied researchers who may or may not have received advanced training in statistical methodology. Accessibility is enhanced when investigators can understand and apply statistical methods to their own data and interpret results appropriately in order to answer their research and/or practice questions.

Interrupted Time-Series Simulation (ITSSIM) is a new method of statistical analysis that aims to address the four criteria outlined above. ITSSIM effect sizes are estimated via Monte Carlo simulation modeling, where data parameters are estimated from an observed single-case data set and treatment effects are calculated for a range of plausible conditions. Essential to the ITSSIM method is the assumption that *one observed time-series may be explained by many plausible effects and conditions*.

ITSSIM determines what treatment effect size is most likely (and how reliable that effect is) based on the many possible conditions which could plausibly yield the observed data.

ITSSIM differs from many other single-case statistics in that it does not attempt to "control" for data patterns such as baseline trend and autocorrelation. Instead, ITSSIM models the uncertainty introduced by these time-dependent processes when calculating an effect size. Before ITSSIM is formally introduced, some limitations of conventional single-case control methods will be illustrated. These limitations will establish the rationale for simulation-based effect size statistics like ITSSIM.

## Limitations of Statistical Control

Historically, single-case experiments were the primary tool of behavioral researchers and applied behavioral analysts. Early operant conditioning researchers demonstrated subtle, and often elegant, manipulations of human and animal behavior via the controlled case study (e.g., Skinner, 1948). Those investigators, and the applied methods they pioneered, emphasized experimental control above nearly all other aspects of research. Early behaviorists pointed out that with a very high degree of control over the extraneous (and potentially confounding) aspects of the experiment, the investigator could uncover the underlying mechanisms of behavior and intervene accordingly to produce change in the subject.

Sidman (1960) stated, "Experimental control is as basic to our understanding of behavior as it is to our manipulation of behavior" (p. 342). Skinner (1956) said even more succinctly, "Control your conditions and you will see order" (p. 223). Behaviorists argued that sufficient experimental control eliminated the need for large samples and

statistical analyses. Their position signaled an enduring philosophical divide in psychology (see Morgan & Morgan, 2001) between experimental and statistical methodologies, described famously by Cronbach (1957):

> The well-known virtue of the experimental method is that it brings situational variables under tight control. It thus permits rigorous tests of hypotheses and confident statements about causation. The correlational method, for its part, can study what man has not learned to control or can never hope to control. (p. 672)

Single-case research has since outgrown the behavioral laboratory. Brief interrupted time-series methods are frequently applied to a host of applied and clinical problems that, for ethical and practical reasons, investigators have little hope of controlling experimentally (Barlow & Hersen, 1984). And just as Cronbach (1957) described, single-case investigators have come to rely on statistical methods to control what their designs cannot.

The following discussion will illustrate the limitations of statistical control with two challenges faced by single-case investigators: the problem of baseline trend and the problem of autocorrelation. Both problems involve analyzing data that has some time-dependent deterministic process which confounds a straightforward pre-treatment/post-treatment (or A phase/B phase) comparison.

**Baseline Trend**

For the trend problem, consider the dilemma of the recovering patient. If prior to treatment the patient is already improving during the baseline phase, and then the patient fully recovers during treatment, one cannot infer that the recovery was due solely (or

maybe even partly) to the effects of the treatment. Perhaps the patient would have recovered just the same without intervention. Or perhaps the patient would have recovered more quickly were it not for some unexpected noxious effect of the treatment.

Single-case investigators have proposed a number of statistical trend control methods which aim to remove the influence of baseline trend from both phases (Tarlow, in press). Most methods control trend by first estimating a baseline trend parameter, and then statistically removing the effect of the estimated trend from both A and B phases. In theory, the "corrected" baseline data no longer contain observable trend, and whatever effects are leftover in the corrected treatment phase data (level-change, slope-change, or otherwise) are assumed to exist beyond the influence of baseline trend, which has been statistically removed. There is a general consensus about this approach among single-case statistics that account for baseline trend, though the methods of trend estimation and correction vary by method. Most, but not all, proposed trend correction methods assume linear trend patterns; trend may be estimated with parametric or nonparametric statistical models; it may be assumed that all data are trended and in need of correction, or it may be assumed that data are stable and untrended unless proven otherwise.

Unfortunately, despite the variety of correction methods, few single-case statistics explicitly account for the reliability or unreliability of baseline trend estimates. This limitation raises several questions. Should any detected trend be corrected? What if observed trends are simply due to the chance variability of the sample data, and no real trend exists? Parker, Vannest, Davis, and Sauber (2011) recommended that investigators control trend when an observed baseline trend coefficient falls above a given cutoff

point. But what if the observed trend, which might be based on only a few data points, is very different than the true (unobserved) trend value? In that case, the effect size calculated from "corrected" data will bear little resemblance to the true treatment effect. Tarlow (in press) recommended trend correction only when observed baseline trends were statistically significant; essentially, a null hypothesis of no trend must be rejected before the investigator alters observed data. While this approach emphasizes the problem of low power in trend detection methods, null hypothesis significance testing is not guaranteed to yield accurate parameter estimates (Cohen, 1994). An investigator with a highly statistically significant baseline trend coefficient may know very little about the true trend that needs correction—the investigator knows only that the true trend is probably not zero. Other single-case effect size statistics disregard hypothesis testing completely, and incorporate a baseline trend correction automatically and regardless of the trend's magnitude or statistical significance (Allison & Gorman, 1993; Manolov & Solanas, 2009, 2013; White & Haring, 1980).

Trend correction is risky when the accuracy of the trend parameter estimate is unknown. To illustrate, consider the hypothetical time-series in Figure 1. This time-series of 100 data points was generated from a random white noise signal with normally distributed error residuals and unit variance ($s^2 = 1.00$). The series follows a linear trend with slope $\beta = 0.20$. Suppose an investigator observes a baseline of ten points from the Figure 1 data (the other 90 points are unobserved). What might the investigator conclude about trend from the observed data, and how accurately would those observations reflect the true trend parameter of $\beta = 0.20$?

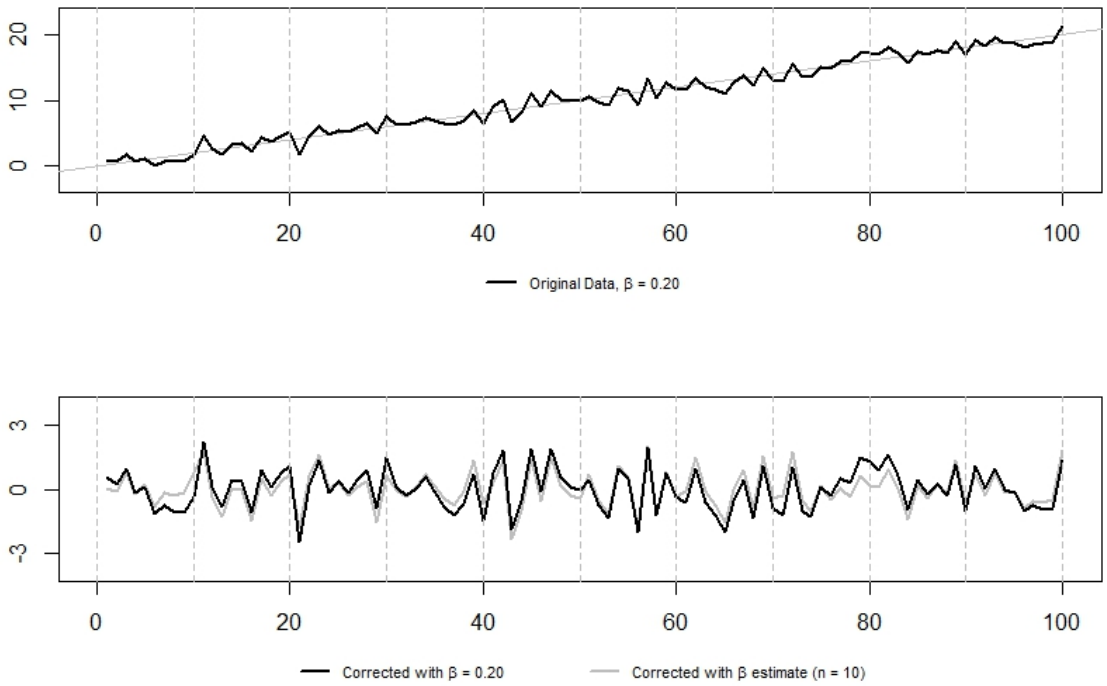



Figure 1. Hypothetical time-series with corrections for trend (n = 100). Original data is displayed in the top graph. In the bottom graph, trend was removed using the known slope parameter, $\beta = 0.20$ (dark line) and removed using an estimated slope from each interval of 10 data points (light line), resulting in an unwanted reduction of variance.

The Figure 1 time-series is partitioned into brief time-series of ten points each. Suppose each sample of ten points is corrected for baseline trend via ordinary least squares (OLS) regression—but the correction is based only on the estimated trend in the ten-point sample, not the "unknown" true parameter of $\beta = 0.20$. In some cases, the estimated trend will be less than the true parameter, and in other cases it will be greater than the true parameter. Presented the bottom graph of Figure 1 are two corrected time-series. In one corrected series, baseline trend was controlled with the trend estimates of

each ten-point sample; in the other, trend was controlled with the true population parameter (it is noted that investigators never know the true parameters of their data—only the precision of their statistical estimates).

Although the two corrected time-series in Figure 1 appear similar, there is one important difference. The variance of the time-series which was corrected with the true trend parameter is unchanged from the original data, $s^2 = 1.00$. However, the variance of the time-series corrected with the sampled trend parameter estimates was reduced by over 20%, $s^2 = 0.78$. The loss of variance represents an overfitting which occurs when trend estimates are influenced by random noise in the time-series sample. Simply put, this time-series has lost 20% of the information contained in the original data—essentially, one fifth of the original information was mistaken for trend and removed. Any inferences made from the overcorrected data should take into account this reduction in variance; failure to do so could distort conclusions about the influence of trend, treatment future treatment effectiveness, and any other data properties.

Here is the crucial point of this heuristic example: *nearly all baseline trend control methods in single-case research assume that trend estimates are perfectly precise*. A trend coefficient is estimated and then data are corrected and analyzed. There is rarely any account for the unreliability of trend parameter estimates. In addition, as the length of baseline series decreases, so does the accuracy of trend parameter estimates. If the Figure 1 data were instead partitioned into brief time-series of five points, the corresponding reduction in variance from trend correction would increase from 20% to nearly 40% ($s^2 = 0.63$). This result should be of great concern to single-case

investigators, who may be drawing conclusions from "trend controlled" data that have little resemblance to reality.

**Autocorrelation**

Just as single-case methods rarely account for the precision or imprecision of trend parameter estimates, the same is true for autocorrelation parameter estimates. Some have advocated for ARIMA "back-casting" or "cleansing" autocorrelation before single-case data are analyzed (Parker et al., 2006, 2011). The rationale is similar to trend correction: an autocorrelation parameter is estimated from observed data, then data are altered to remove the influence of the estimated autocorrelation, then an effect size is calculated. Also like trend correction, this type of data cleansing can only be performed with an exact estimate of the autocorrelation parameter—i.e., the estimate is assumed to be perfectly precise. Unfortunately, autocorrelation estimation requires a very large number of data points to yield reliable parameter estimates, far more data points than are typically available to single-case investigators (Box & Jenkins, 1970; Glass, Willson, & Gottman, 1975). Back-casting may remove the autocorrelation present in the sample—just as trend control will remove baseline trend in the observed data—but the procedure may lead to misleading conclusions about the effect of treatment.

The hypothetical time-series data in Figure 2 illustrates the hazard of back-casting to remove autocorrelation from brief time-series. Similar to the example in Figure 1, the Figure 2 data was generated using the lag-1 autoregressive parameter $\varphi_1$ = 0.20 and unit variance ($s^2$ = 1.00). Two corrected time-series are presented. In one, the influence of autocorrelation is "cleansed" using the true $\varphi_1$ parameter; the total variance

of the time-series is nearly unchanged, $s^2 = 0.96$. In the other corrected time-series, autocorrelation is cleansed one ten-point sample at a time using the $r_1$ estimator; the result is an overall reduction in variance to $s^2 = 0.79$, or again roughly 20%. The results from Figure 1 and Figure 2 beg the question: If a large percentage of the information contained in the observed data is erroneously discarded during the statistical correction process, how useful are the effect size estimates calculated from corrected data?

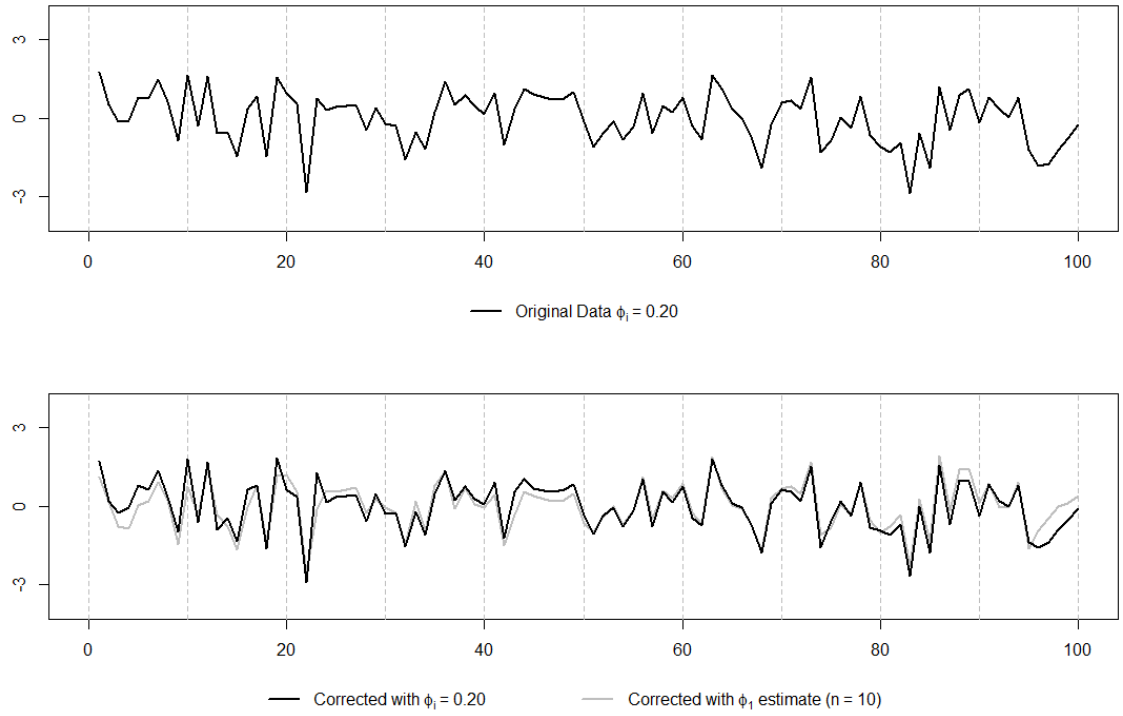



Figure 2. Hypothetical time-series with corrections for autocorrelation (n = 100). Original data is displayed in the top graph. In the bottom graph, trend was removed using the known autocorrelation parameter, $\varphi_1 = 0.20$ (dark line) and removed using an estimated slope from each interval of 10 data points (light line), resulting in an unwanted reduction of variance.

The problem of imprecise autocorrelation estimation is further illustrated in Figure 3, which illustrates the confidence intervals for $r_1$ using the error variance estimator proposed by Huitema and McKean (1991, 2000a).[1] The $r_1$ estimator is quite imprecise with the sample sizes typical of single-case experimental designs. While data cleansing may appear useful—it does remove the *estimated* $r_1$ in the *observed* time-series—the procedure is unlikely to control for the true $\varphi_1$ parameter because the estimator itself is imprecise with small samples. The single-case investigator may wish to control autocorrelation in a brief time-series, but it is rarely clear what value of $r_1$ should be used for statistical control. Due to the sampling properties of $r_1$, and the brief lengths of most single-case studies, a range of $r_1$ values is often plausible for an observed single-case data set.

[1] Huitema and McKean's (2000) error variance estimator, $\text{var}(r_1) = [(N-2)^2/N^2(N-1)]$, demonstrated less bias than Bartlett's (1946) popular statistic, $\text{var}(r_1) = 1/N(1-r_1^2)$, when applied to the small-sample, low-autocorrelation time-series that characterize single-case experiments.

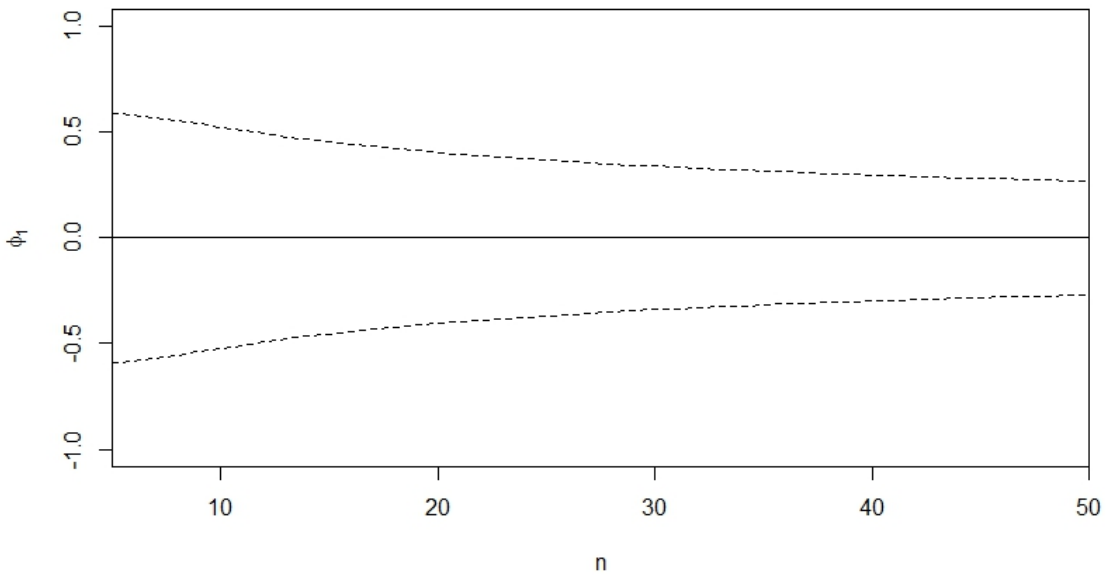



Figure 3. 95% confidence intervals for the lag-1 autocorrelation coefficient when $\varphi_1$ = 0.0. Confidence intervals (illustrated by the dashed lines) are based on Huitema and McKean's (1991, 2000a) $r_1$ variance estimator.

**Simulation Methods Can Address Limitations of Statistical Control**

As the examples above demonstrate, statistical control of brief time-series data is limited by the precision of baseline trend and autocorrelation parameter estimates. The reliability of those and other parameter estimates are of paramount importance to overall analysis, as they determine the degree of confidence one can place in statistical results, i.e., the effect size estimate. However, many popular single-case statistical methods assume parameter estimates are perfectly precise—even though we know this assumption is false (particularly for autocorrelation, but also for trend and other data

parameters). In most statistical control methods, a baseline trend and/or autocorrelation coefficient is calculated and used to adjust observed data for effect size calculation, but those final results are almost never reported within the context of the reliability, or unreliability, of the parameter estimates.

Computer-aided simulation methods have the potential to address this limitation. Rather than calculate one effect size from one set of trend and autocorrelation parameter estimates that are assumed to be perfectly precise, many effect sizes can be calculated from many sets of parameter estimates, with each set of parameters having a different probability based on the observed data. It may in fact be quite meaningful for investigators to consider the range of plausible treatment effects that might have yielded their observed results. Using this approach, investigators would be informed about the reliability of their effect size estimate (e.g., "The calculated effect size suggested an effective treatment, but the range of plausible effects was large and included the possibility that the true effect of treatment is trivially small"). A simulation approach, based on the assumption that many different effects could plausibly lead to the observed data, offers context for interpretation absent in most single-case statistical analyses.

**Multilevel Modeling**

Another approach to single-case measurement and meta-analysis is multilevel modeling, which has recently grown in popularity (Shadish et al., 2008). Multilevel models increase the efficiency of parameter estimation by pooling data within and across cases (Van den Noortgate & Onghena, 2003). Whereas simulation methods like ITSSIM assume many effects could plausibly lead to the observed data, multilevel models

assume many single-case participants share some fixed underlying parameters that can be estimated by pooling data across groups of cases. Multilevel methods are a powerful tool for synthesizing single-case research findings; however, they often require multiple (sometimes many) cases in order to yield accurate effect size estimates (Ugille et al., 2012; Shadish et al., 2014). In a sense, multilevel modeling is not a true "N-of-1" single-case method, as this approach requires many data from many cases. For this reason, multilevel methods have been criticized for being inaccessible to clinical scientists and investigators who lack large single-case data sets or the statistical training to conduct and interpret multilevel studies. Parker and Vannest (2012) described this as a distinction between "bottom-up" and "top-down" research synthesis:

> The bottom-up strategy is distinct from a top-down strategy in which an overall or omnibus analytic model is fit to the entire design. Whereas top-down appears more elegant, it entails a marked risk, which is to ignore the idiosyncrasies or uniqueness of a design and its data patterns. It is true that any template can be modified, but to do so in [multilevel modeling], for example, requires statistical skills beyond those of most interventionists. This raises the broadest concern with the top-down analytic strategy because the behavior analyst is not able to maintain decision-making control, may not even be able to confirm the legitimacy of a model fit, and may not even be able to interpret the results. (p. 263)

Multilevel methods have a unique potential for organizing the emerging field of single-case meta-analysis. However, investigators would be remiss to ignore the essentially pragmatic nature of single-case research (Iwakabe & Gazzola, 2009; Fishman, 2005; Shadish & Rindskopf, 2007). Sophisticated methods which can be applied to true "N-of-1" single-case experiments, and are accessible to applied researchers and practitioners, would be a valuable supplement to multilevel modeling. ITSSIM aims to be such a method.

## Interrupted Time-Series Simulation: ITSSIM

Interrupted Time-Series Simulation (ITSSIM) follows a three-stage process to yield an effect size estimate for an interrupted (AB) time-series design. The three stages of ITSSIM are: (1) parameter estimation, (2) time-series simulation, and (3) effect size calculation. In the first stage, estimates and standard errors are calculated for seven data parameters (A phase level, B phase level, A phase trend, B phase trend, A phase error variance, B phase error variance, and cross-phase autocorrelation). These parameter estimates are used to construct two models. A *null effect model* (based on the A phase data) describes the participant's response pattern prior to treatment. An *experimental effect model* (based on the B phase data) describes the participant's response pattern during/after treatment. In the second stage of ITSSIM, time-series data sets are simulated from the null effect model and experimental effect model. This stage yields thousands of artificial time-series that represent the null effect and treatment effect within a range of plausible parameter values. Essentially, the simulation stage produces two distributions of time-series, one representing the range of possible outcomes without treatment, and one representing the range of possible outcomes with treatment. In the third stage, the null effect distribution and the experimental effect distribution are compared for a standardized mean difference effect size (essentially a Cohen's *d*-type statistic). Due to its desirable statistical properties, this effect size statistic is suitable for meta-analysis.

### How Is ITSSIM Different from Other Single-Case Statistics?

While reviewing the ITSSIM method below, it is useful to keep in mind three differences between ITSSIM and other single-case statistics. First, the ITSSIM effect

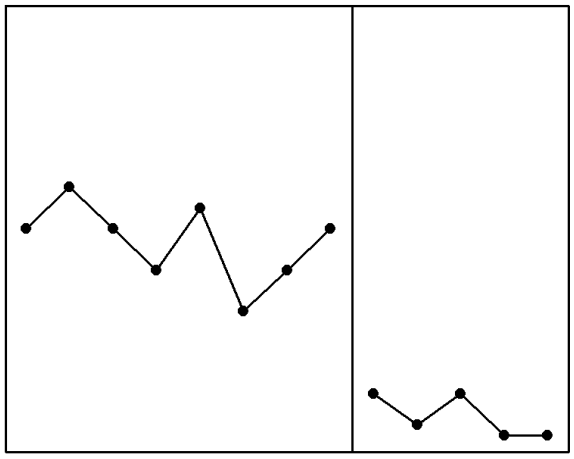
size is not calculated from exact estimates of level, trend, etc., nor is it calculated directly from observed data. Instead, the ITSSIM effect size is calculated from many simulated time-series that are based on a range of parameter values which could plausibly account for the observed data. For example, rather than fit a single estimate of trend to a small sample of data points, ITSSIM models a range of many trend values that could fit the data, with more likely trend values occurring more frequently among the simulated cases. If observed data lead to very precise parameter estimates with small standard errors (e.g., "From the twenty baseline data points, I am confident that the true linear baseline trend coefficient lies somewhere close to $\beta_1 = 0.30$"), then simulated data sets will reflect that confidence. In this way, unlike many other methods, ITSSIM incorporates the reliability of the parameter estimates into effect size estimation.

Second, ITSSIM effect size estimates are calculated from predicted data rather than controlled data. Many single-case statistics that model trend or autocorrelation do so by "correcting" observed data points to make comparisons between A and B phases more tenable (e.g., calculating an effect size from linear regression/ARIMA residuals). This approach is used because A phase and B phase observations are recorded at different time intervals, and time confounds direct A-to-B phase data comparisons when data points are changing deterministically over time (as with baseline trend or autocorrelation). ITSSIM instead simulates B phase data from both the null effect model and the experimental effect model. These artificial B phase time-series, which are simulated predictions, are directly comparable without "correction" because they are predictions for the same time interval. This strategy is similar to single-case effect size

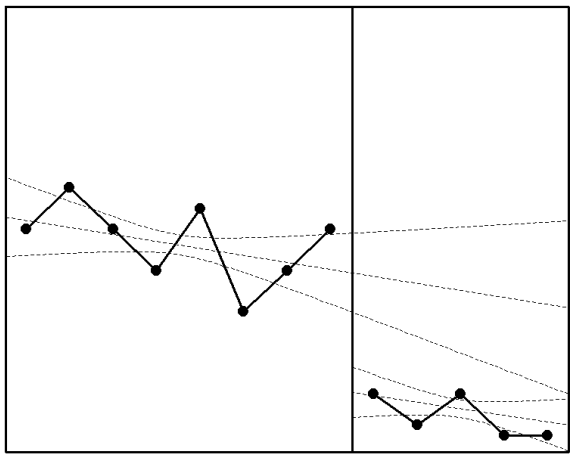
methods proposed by Gorsuch (1983), Allison and Gorman (1993), and Manolov and Solanas (2013), where B phase observations are compared to data predicted by baseline observations.

Third, the ITSSIM effect size is not calculated by a comparison of A phase and B phase data. Instead, a ITSSIM effect size is calculated from *distributions* of simulated time-series. The null effect distribution (based on A phase data) and the experimental effect distribution (based on B phase data) are both composed of simulated time-series—predictions—for the B phase time interval. These distributions do not directly represent baseline and experimental phase data, but they do represent a range of plausible data parameters based on the A and B phases, respectively.

The three stages of the ITSSIM procedure are discussed further below, and the Appendix demonstrates the exact ITSSIM calculations using a hypothetical data set. However, Figure 4 presents a simplified illustration of the three ITSSIM stages.

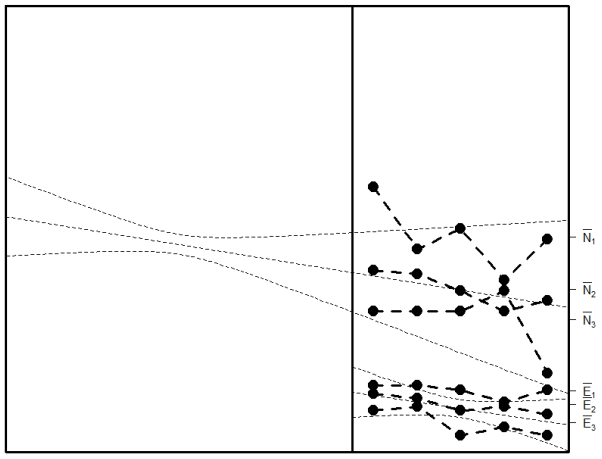


**Original Data**

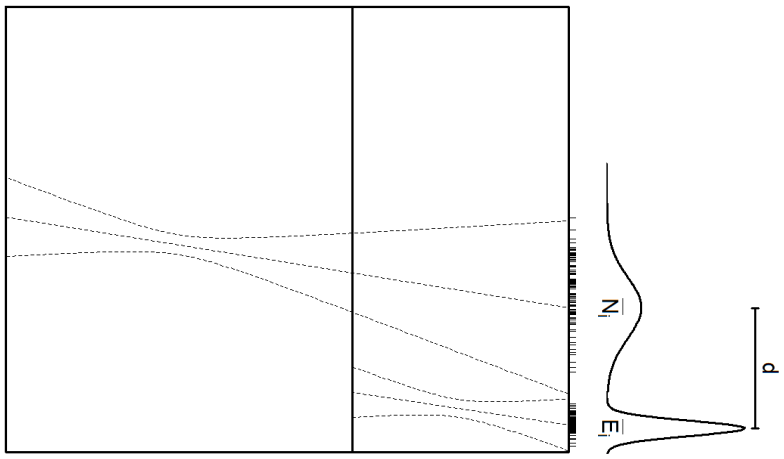


**I. Parameter Estimation**

Figure 4. Simplified illustration of ITSSIM stages.




**II. Time-Series Simulation**




**III. Effect Size Calculation**

Figure 4 Continued

**Stage I: Parameter Estimation**

A statistical model represents an implicit model of reality (Thompson, 2006), or put another way, models "bring nature and theory into closer and closer agreement" (Kuhn, 1996, p. 27). Models based on unrealistic parameters—or models based on an insufficient number of realistic parameters—will do a poor job of describing reality. There has been considerable discussion about which parameters should be included in the analysis of single-case data. Many agree the most fundamental parameters in an interrupted (AB) time-series design are level and trend. Clinical treatments are often expected to affect either the level (i.e., magnitude) of the outcome variable, the trend (i.e., slope) of the outcome variable over time, or both.

Level and trend alone may be insufficient to comprehensively model single-case data. Table 1 presents the parameters for analysis recommended by several well-cited single-case methodology publications. Variability, also described as stability, describes the amount of variance or "bounce" in time-series data. Autocorrelation, or serial dependency, is the degree of correlation between past and future error residuals, i.e., the degree to which past fluctuations in the time-series are carried over into future data points. Overlap, similar to level, is a popular metric for single-case effect size measurement that accounts for the amount of A phase and B phase data points that fall within a common overlapping range. Periodicity describes cyclical or seasonal patterns in data over time (i.e., higher-order autocorrelation). Immediacy is the duration between the introduction of treatment and the treatment response.

Table 1

*Single-Case Data Parameters Recommended for Analysis*

| | **Level** | **Trend** | **Variability** | **Autocorrelation** | **Overlap** | **Periodicity** | **Immediacy** |
|---|---|---|---|---|---|---|---|
| Campbell & Herzinger (2010) | X | X | | X | | | |
| Jones et al. (1977) | X | X | | X | | | |
| Hartmann et al. (1980) | X | X | | X | | | |
| Hayes (1981) | X | X | X | | | | |
| Parsonson & Baer (1992) | X | X | X | X | X | | |
| Kazdin (1982) | X | X | X | X | | | |
| Kratochwill et al. (2010) | X | X | X | | X | | X |
| Smith (2012) | X | X | | X | | | |
| Shadish (2014b) | X | X | | X | | X | |

Four parameters were selected for the ITSSIM model: level, trend, variability, and autocorrelation. Together these parameters offer a comprehensive description of an interrupted (AB) time-series. For an AB single-case design, coefficients for level, trend, and variability are estimated separately for each phase; autocorrelation is estimated for the whole series after controlling for the other three variables.[2] This procedure leads to

[2] Autocorrelation is estimated across both phases because of the low power of autocorrelation estimators in brief time-series (see discussion above, e.g., Figure 3). The ITSSIM model assumes autocorrelation is relatively stable from phase to phase.

an estimation of seven total coefficients: A phase level, A phase trend, A phase variability, B phase level, B phase trend, B phase variability, and across-phase autocorrelation.

The order that parameters are estimated is important. For example, the presence of trend can grossly distort estimates of autocorrelation (Huitema & McKean, 1998). Trend should be estimated and then removed from the time-series in order to accurately measure the degree of autocorrelation. Within-phase variability should also be controlled before estimating cross-phase autocorrelation, because heteroscedasticity within a time-series could distort the estimated autocorrelation coefficient. Given these constraints, a ITSSIM analysis is conducted with the following steps.

**Estimate level and slope parameters: Theil-Sen robust regression.** Regression models have been widely used in single-case measurement and simulation methods (Huitema & McKean, 2000b). Regression is useful for single-case designs because it models both trend and level changes. However, OLS regression is limited by its assumptions and sensitivity to outliers (Brossart et al., 2011). Nonparametric robust regression is a promising alternative to OLS regression in single-case research because it makes fewer distributional assumptions and is less sensitive to outliers. Theil-Sen nonparametric regression (Sen, 1968; Theil, 1950) is used in ITSSIM analysis because it is robust, yields a relatively small standard error, and is efficient in small samples (Wilcox, 1998, 2001). Theil-Sen regression has been applied to single-case behavioral data (Tarlow, in press; Vannest et al., 2012) and has been used in other fields to analyze autocorrelated, non-normal, monotonically-trended time-series data (e.g., Yue et al.,

2002).

Theil-Sen level and trend (i.e., intercept and slope) coefficients are therefore estimated for observed A and B phase data separately. Standard errors for these coefficients are then calculated using a bootstrap procedure (Wilcox, 2001).

**Estimate variability parameters: error variance.** Variability within each phase is calculated as the variance of the Theil-Sen regression residuals. Both the phase error variance ($s^2$) and its standard error are calculated using ordinary least squares methods.

**Estimate lag-1 autocorrelation coefficient: unbiased $r_1$.** An unbiased $r_1$ estimator (Ferron, 2002; Huitema & McKean, 1991, 2000a) is calculated across both A and B phase data after standardizing the Theil-Sen residuals (i.e., dividing residuals by their within-phase standard deviation). There are several ways to estimate the standard error of the $r_1$ statistic, with Bartlett's (1946) method being the most popular. However, Huitema and McKean (1991) demonstrated that Bartlett's method was biased in small samples; their modified var($r_1$) estimator is used in the ITSSIM model.

At the end of the parameter estimation stage, the ITSSIM analysis will yield seven coefficients and their standard errors. These parameter estimates comprise the *null effect* and *experimental effect* models. Table 2 presents a set of hypothetical parameter estimates, organized into the two statistical models. While these models are based on data observed at different time intervals (the A and B phases, respectively), they may be used to make predictions at any point in time—though the predictions become less precise the farther away in time they are from the observed data on which each model was based.

Table 2

*Hypothetical Parameter Estimates in an ITSSIM Analysis*

| | | **Coefficient** | **Std. Err.** |
|---|---|---|---|
| **Null Model** | intercept | 5.375 | (2.752) |
| | slope | -0.083 | (0.540) |
| | *s* | 2.092 | (0.501) |
| **Exp. Model** | intercept | 0.375 | (1.604) |
| | slope | -0.125 | (0.489) |
| | *s* | 0.845 | (0.295) |
| **Autocorrelation** | $r_1$ | 0.133 | (0.244) |

**Stage II: Time-Series Simulation**

It is useful to note that, at the end of the parameter estimation stage, the investigator is left with a time-series that has unit variance and an estimated autocorrelation value. It would be possible at this point to work backwards through the parameter estimation stage, restoring the within-phase error variance and then blending the residuals with the Theil-Sen regression coefficients, to return these standardized residuals to the original data set. In fact, using only the parameter estimates and standard errors from Stage One (e.g., Table 2), it would be possible to work backwards from an artificial white noise signal and arrive at a time-series very similar to the original data set, with differences due only to random variability in the initial standardized residuals. This process of "working backwards" is essentially how the time-series simulation stage of ITSSIM works.

**Randomly sample simulation parameters.** A white noise signal is created with a preset degree of autocorrelation, and then within-phase variability, trend, and level are added. But instead of using the parameter estimates of the original data set (e.g., Table 2), new coefficients are randomly drawn from the estimated parameters' sampling distributions. These new level, trend, variability, and autocorrelation coefficients—though not the exact values calculated from the observed data—plausibly represent the true nature of the participant's response pattern given the known precision of the original parameter estimates. This procedure is arithmetically simplified because the sampling distributions of Theil-Sen, variance, and lag-1 autocorrelation coefficients are normal (Anderson, 1942; Cox, 1966; Mann, 1945; Sen, 1968).

This step is iterated 100,000 times each with the null model and experimental model estimates, yielding two sets of intercept, slope, error variance, and autocorrelation coefficients. The first set of 100,000 coefficients represents the plausible range of participant responding in the absence of treatment, i.e., the null effect. The second set of 100,000 randomly sampled coefficients represents the plausible range of participant responding under the effect of treatment, i.e., the experimental treatment effect. Using each set of sampled coefficients, a random white noise signal could then be manipulated into a brief interrupted time-series with the specified level, trend, variability, and autocorrelation. However, instead of creating AB data sets (like the original observed data), the ITSSIM method uses these simulated parameters to create simultaneous B phase time-series.

**Simulate B phase time-series.** Recall from earlier that ITSSIM does not control

for time-dependent data patterns like trend or autocorrelation. Instead, ITSSIM makes predictions using estimates of trend, autocorrelation, and other parameters. Predictions based on the A phase baseline data are compared to predictions made from the B phase experimental phase data. In order to make the two sets of predictions directly comparable, they are simulated for the same interval of time. After coefficients are randomly drawn from the parameters' sampling distributions, a corresponding B phase time-series is generated from each of the 100,000 simulated null effect models and simulated experimental effect models. The result is 100,000 null effect time-series and 100,000 experimental effect time-series, all corresponding to the same interval of time—the time interval of the B phase.

**Stage III: Effect Size Calculation**

**Calculate standardized mean difference from simulated time-series means.** To find a treatment effect size, the mean of each simulated B phase time series is first calculated. This yields two distributions of means—one for the set of null effect time-series and one for the set of experimental effect time-series. The distributions of means are approximately normal due to the Central Limit Theorem. Thus, as with any large sample study of control group and experimental group data, the ITSSIM null effect distribution and the experimental effect distribution may be compared to find a standardized mean difference, *d* (see Figure 4), which is interpreted as the size of treatment effect accounting for level-, trend-, and variability-changes as well as lag-1 autocorrelation. The *d*-statistic may also be converted to an equivalent $R^2$ effect size for

investigators who may wish to communicate their effect as a percentage of variance accounted for by the treatment effect.

An unstandardized mean difference, *D*, may also be reported as an effect size. This statistic describes the change in participant's response due to treatment after modeling level, trend, variability, and autocorrelation effects. The unstandardized effect size, reported in the original metric of the observed data (e.g., "*D* fewer intervals of disruptive behavior"), may be particularly relevant to practitioners wishing to communicate their treatment results in a clinically relevant way, though it is less useful for synthesizing results across studies with different designs and outcome measures.

**Meta-Analysis with ITSSIM**

An area of growing interest is the meta-analysis of single-case research. Synthesizing results across single-cases experiments is one area where statistical methods are expected to be superior to visual analysis (Beretvas & Chung, 2011; Busk & Serlin, 1992; Maggin & Odom, 2014; Scruggs & Mastropieri, 2001). That said, many single-case effect size statistics do not possess the statistical properties necessary for meta-analysis. However, the standardized mean difference *d*-statistic produced by ITSSIM is ideal for aggregating results across studies.

There are two levels of consideration when combining single-case experimental effects. The most general level is the aggregation of effect sizes across cases (or individuals). However, there may also be multiple effects measured for each individual case in a single-case experiment. Take for instance a reversal ABAB design. In an ITSSIM analysis there will be one effect size for the first AB phase contrast, and a

second effect size for the second AB phase contrast. These effect sizes can be combined to find a meta-analytic mean effect size. Combining effects within individuals and across cases may require multilevel estimation methods.

**ITSSIM Software**

ITSSIM is accessible to single-case investigators via free software download at http://www.ktarlow.com/stats/itssim (Tarlow, 2017a). This is a standalone program that does not require additional statistical computing software. When opened, the program prompts the user to enter A phase and B phase data. After the observed data are inputted, the program estimates the model parameters and performs the simulation. All coefficients, their standard errors, and the standardized mean difference effect size are reported as output. ITSSIM software requires no computing code or syntax to use, and should be easy to implement for applied researchers and practitioners without statistical computing experience. An example of ITSSIM software input and output is presented in Figure 5.

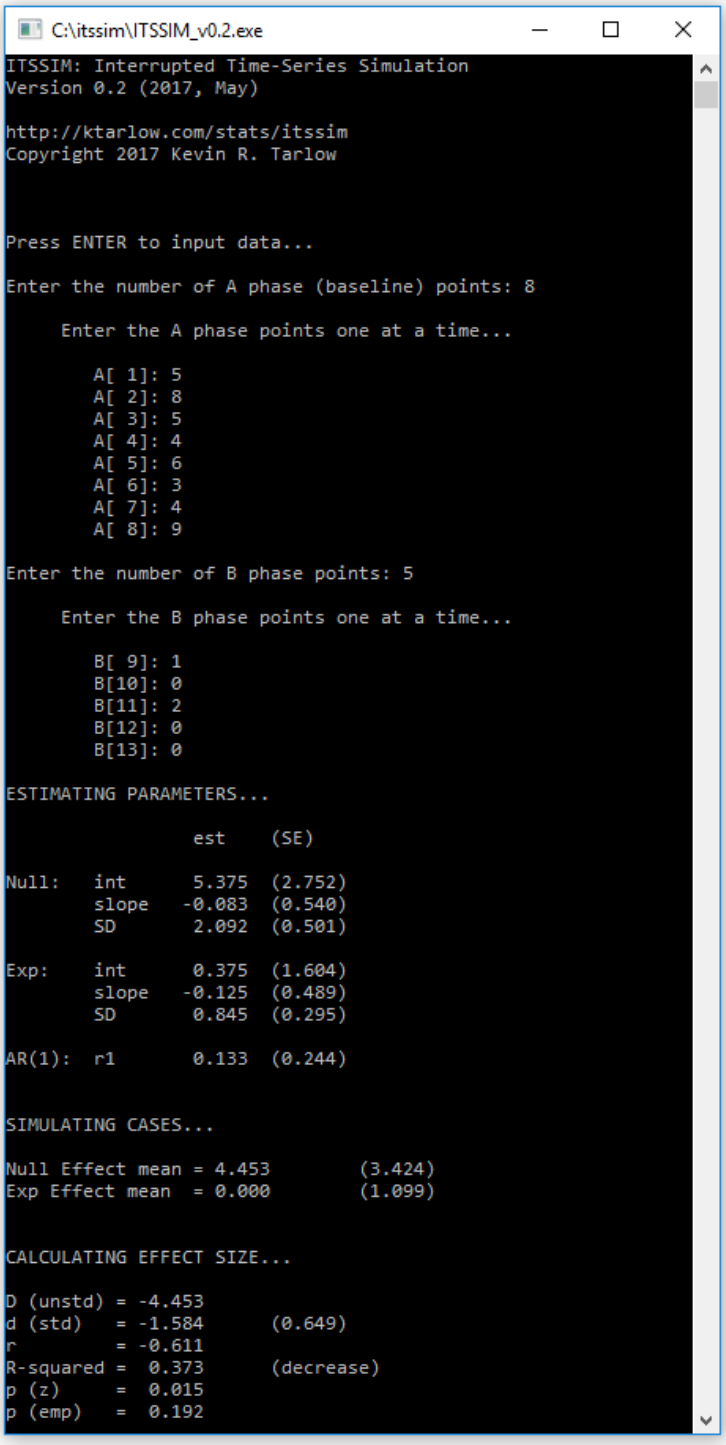



Figure 5. Example of ITSSIM console with input and output. Software is available for download at http://ktarlow.com/stats/itssim (Tarlow, 2017a).

**Assumptions and Limitations**

ITSSIM has several statistical assumptions and limitations. The violations of some assumptions, such as the expectation of stable autocorrelation across phases, may not greatly affect results (the degree of autocorrelation observed in brief time series is relatively small, e.g. Shadish & Sullivan, 2011, so cross-phase differences are also expected to be minimal). The violation of other assumptions, like the assumption of linear trends, may lead to gross distortions of results. A list of ITSSIM assumptions is presented below, and should be considered before the method is applied to experimental data. More investigation is needed to determine how violations may affect results, and to that end, another paper is currently being prepared that uses simulation methods to determine how ITSSIM functions under a range of different parametric conditions.

Some ITSSIM assumptions are:

- A linear trend model is assumed, and ITSSIM analysis may give misleading results when applied to data with nonlinear trends and/or pronounced ceiling and floor effects. However, the Theil-Sen estimator finds a best-fit line for any monotonic (i.e., linear or nonlinear) trend pattern, so the error introduced by nonlinear trends are expected to be reduced.
- Lag-1 autoregressive error is assumed to be unaffected by treatment. This assumption is consistent with the general view that autocorrelated behavior is a function of the organism under study rather than an effect of treatment (e.g., Baer, 1988). An example of a single-case effect size that makes a similar assumption is the Bayesian model proposed by Swaminathan, Rogers, and

Horner (2014), described in more detail below.

- Only lag-1 autoregressive error structures ($\varphi_1$) are modeled. Moving average, integrated, and higher-order error structures are not considered. Other single-case statistics which model autocorrelation often assume a lag-1 autoregressive error structure (e.g., Moeyaert et al., 2014; Shadish, Hedges, & Pustejovsky, 2014). The issue of error model identification has been underexplored in single-case research, and single-case investigators would be wise to consult the time-series analysis and multivariate literature which demonstrate that covariance structures are difficult to identify correctly (Sivo & Willson, 2000; Velicer & Harrop, 1983).
- The outcome variable is assumed to be continuous. Investigators analyzing time-series with count data (or other non-continuous data) should exercise caution in interpreting results. Future updates of the ITSSIM software should include options for non-continuous data modeling.
- Normally distributed data are assumed in the current version of ITSSIM software. Future software releases should include options for modeling non-normal data distributions (e.g., Poisson, binomial, etc.).
- During the data simulation procedure of a ITSSIM analysis, the level, trend, error variance, and autocorrelation parameters are assumed to be independent from each other. There is little published research to suggest, one way or another, if this assumption is tenable. It is expected that relationships between these parameters vary from study to study, though more exploration is needed

to determine if and how violations to this assumption would impact statistical results.

- The current version of the ITSSIM software does not permit missing data—i.e., all observations are assumed to be recorded at equal intervals in time.

## Revisiting the *Journal of School Psychology* Special Issue

This is the first in a series of papers that aim to evaluate ITSSIM and its utility for single-case data analysis. The primary goal of this paper is expository, to introduce the rationale for simulation methods in single-case analysis and explain ITSSIM's use and interpretation. However, an initial field test of ITSSIM was conducted to determine if it performed similarly to other sophisticated single-case analytic methods.

The *Journal of School Psychology* published a special issue, *Analysis and Meta-Analysis of Single-Case Designs* (Shadish, 2014), in which five research teams independently analyzed the same single-case data set with different statistical methods. The methods used in the special issue represent the cutting edge of single-case data analysis. They are more sophisticated than the simpler methods which tend to be more studied but also more statistically limited (e.g., Parker, Vannest, & Davis, 2011).

The data set used in the special issue was reanalyzed with ITSSIM, and results were compared with the five previously published methods. It was hypothesized that ITSSIM would produce comparable results, given that it models data parameters similar to the other methods (level, trend, variability, autocorrelation), albeit with a different statistical framework.

## Data for Analysis

Five single-case research teams analyzed the same data set in a special issue of the *Journal of School Psychology* (Shadish, 2014a). That data set was revisited for this study and reanalyzed to determine if ITSSIM would yield results similar to other sophisticated analytic methods. All six methods (ITSSIM and the five previously published ones) were used to analyze data from Lambert, Cartledge, Heward, and Lo's (2006) single-case study of the effects of response cards on disruptive behavior by fourth-grade math students. Nine participants were treated with an ABAB reversal design. Data were extracted from the published graphs using GetData Graph Digitizer (2013). All extracted data sets were visually compared with the original published graphs to confirm an accurate data extraction.

## Overview of the Meta-Analytic Methods

In the 2014 special issue of *Journal of School Pyschology* (Shadish, 2014a), five single-case research teams applied their novel statistical analyses to the Lambert et al. (2006) data. Analytic approaches varied, with some methods yielding standardized effect sizes and others yielding unstandardized effects. Two groups took a Bayesian approach to data analysis. All methods used multilevel models to detect effects within and across individuals. A brief summary of each method follows, with attention given to the unique strengths and limitations of each analysis.

**Shadish, Hedges, and Pustejovsky (2014).** This method yields a *d*-type standardized mean difference effect size aggregated across multiple $AB^k$ (i.e, AB, ABAB, ABABAB, etc.) cases. Kratochwill and Levin (2014) pointed out that, while the

Shadish et al. *d*-statistic may be useful for synthesis of multiple single-case studies, it is less useful in applied clinical settings. In order to calculate the *d*-statistic, at least three single-case studies are required. Investigators intending to study only one participant (a true "single-case" experiment) would require other methods. The *d*-statistic models level, autocorrelation, and within- and between-case variability effects. However, the *d*-statistic does not model trend. While Shadish et al. did not find evidence of trend in the Lambert et al. (2006) data, their method may be inappropriate for data sets with trend.

**Shadish, Zuur, and Sullivan (2014).** This method applies generalized additive models (GAM), which are similar to linear regression-based models but use smoothing functions to model nonlinear trend over time. In this case, GAMs can also account for non-normal distributions of outcome data—an important issue in single-case research given the prevalence of count data and other non-continuous outcome measures. In the first step, Shadish et al. test the statistical fit of four GAMs that predict outcome by different arrangements of time, treatment condition, and participant variables. Second, after selecting the best-fitting model, seven additional GAMs are tested, all variations on the best-fit model from step one that add predictor variable interactions and autocorrelation effects. Finally, a best-fitting model is selected from step two and its results are interpreted.

The methods offered by Shadish, Zuur, and Sullivan (2014) are quite sophisticated, and while they have the potential to reveal useful insights about within- and across-case treatment effects, they require a high level of statistical expertise to implement and interpret. For example, the authors emphasize one strength of GAM is

that they do not require the investigator to know the functional form of trend and other data patterns—they can test many models and select the one that best fits the observed data. The drawback of this approach is that many models must be conceived, built, and tested—a process that, as the authors demonstrated, requires time, skill, and nuance.

The clear strength of this approach is that GAMs do a good job of modeling nonlinear trends. However, the authors point out the approach is not well-suited for meta-analysis across studies with different outcome measures (it does not easily yield standardized effect sizes). Mixed models that accounted for autocorrelation were also discarded from the study because the Lambert et al. (2006) data lacked a sufficient number of observations to model complex error structures. In order to simplify analyses, data from longer time series were also discarded to make all cases have an equal number of observations—a restriction that regrettably removed over 10% of the data from the study.

**Rindskopf (2014).** This analysis applies Bayesian methods to single-case data. Instead of estimating a fixed parameter (e.g., effect size), Bayesian statistics emphasize the estimation of probabilities—for example, how likely is a participant in the Lambert et al. (2006) study to respond with disruptive behavior (the outcome measure)? And how much more likely is disruptive behavior during a treatment phase? As Rindskopf points out, Bayesian methods are well-suited for analyzing small sample sizes. This is because they take into account data from all analyzed cases when modeling a single individual; however, all cases are not constrained to have the same characteristics.

On limitation of Rindskopf's approach may be that many researchers are

unfamiliar with Bayesian methods. Rindskopf also chose not to model trend in the Lambert et al. (2006) data for the sake of simplicity. It is possible to account for trend with a Bayesian model, although this adds additional work for the investigator. Rindskopf's approach also does not model autocorrelation.

**Moeyaert, Ferron, Beretvas, and Van Den Noortgate (2014).** Just as ITSSIM emphasizes an analysis of the range of plausible parameter values, the method proposed by Moeyaert et al. instructs single-case investigators to consider results from a range of plausible statistical models and assumptions. In addition to considering different models, Moeyaert et al. use a two-level design to combine effects within and across cases.

Four mixed/multilevel models are specified for use with continuous single-case data. The models vary from simple to complex. The simplest analysis includes only level-change effects, within and across individuals. Autocorrelation, variance, and trend effects are then added in subsequent models. Finally, the authors add a class effect term to specify differences between different groups of cases. Two logistic models are also provided for non-continuous/count data—the first is a simple two-level model offering only level-change terms; the second allows for level-change effects to vary within individuals (A1B1 vs. A2B2 effects).

In most cases a single-case investigator will not know the true parameters underlying observed data—hence the need to investigate multiple models. When several models with differing assumptions converge on a narrow range of effect size results, there is evidence that the true effect lies somewhere near that range. Moeyaert et al. suggest that, in the absence of this convergence, investigators should report results from

all models. This approach has the advantage of not limiting the investigator to a strict set of statistical assumptions which may or may not be true depending on the nature of the experiment at hand. The authors point out this approach may lack sufficient statistical power when the number of observations (i.e., phase lengths) or the number of cases is small. This approach also requires familiarity with SAS statistical software.

**Swaminathan, Rogers, and Horner (2014).** Similar to Rindskopf (2014), Swaminathan et al. use a Bayesian approach to measure and combine single-case effect sizes. Two multilevel models are proposed, one that accounts for level- and slope-change effects, and the other including only level-change effects. Both models incorporate estimates of lag-1 autoregressive error structures. As the authors stated, "The Bayesian procedure is intuitively meaningful and appealing, but it is mathematically complex" (p. 220). Bayesian methods are in many ways attractive for single-case investigators—they are statistically powerful in small samples and may provide results that are interpretable and immediately relevant for practitioners. Implementation of these methods, however, requires some knowledge of Bayesian statistics.

**Plan for Analysis**

The purpose of reanalyzing the Lambert et al. (2006) data is to determine if ITSSIM effect size results are comparable to the five other methods introduced in the special issue of the *Journal of School Psychology.* Shadish (2014a) found the results of those five methods "reasonably consistent with each other" (p. 112), and similar ITSSIM effect size estimates would suggest the simulation method is worthy of future study. All five previously published methods yield unstandardized effect sizes, so this will be one

point of comparison. However, not all methods gave an equivalent standardized effect. Some approaches, like ITSSIM, yield standardized mean differences, which are straightforward to compare across studies. Others methods gave log-odds ratios as the standardized outcome metric, which are less intuitive when describing treatment effects. The ITSSIM unstandardized effect size, *D*, will be compared to the unstandardized results of the other five studies, and the standardized mean difference, *d*, will also be compared to the two other studies which reported results in a similar standardized metric.

## Results

Data from the nine students in Lambert et al.'s (2006) study were analyzed with ITSSIM simulation software (Tarlow, 2017a). The study used a reversal ABAB design for each student, so a total of 18 AB phase contrasts were analyzed. Results of the ITSSIM analyses are presented in Table 3. The A1B1 standardized effects ranged from $d = 1.60$ to 4.66; all A1B1 effects were statistically significant at the $p < .05$ level. The A2B2 standardized effects ranged from $d = -0.39$ to 6.89; seven of the nine A2B2 effects were statistically significant. The fixed effects mean of the A1B1 treatment effects was $d = 2.47$, $p < .001$, 95% CI [1.98, 2.96] ($Q = 12.70$, $df = 8$, $p = 0.122$); the mean of the A2B2 treatment effects was $d = 1.75$, $p < .001$, 95% CI [1.30, 2.21] ($Q = 56.23$, $df = 8$, $p < 0.001$).

Table 3

*ITSSIM Standardized Effect Sizes for Lambert et al. (2006) Data*

| | A1B1 Effect | | A2B2 Effect | |
|---|---|---|---|---|
| **Participant** | ***d*** | **SE** | ***d*** | **SE** |
| Student A1 | 1.63 | 0.65 | 3.71 | 0.80 |
| Student A2 | 3.06 | 0.82 | 5.68 | 1.81 |
| Student A3 | 1.90 | 0.73 | 3.72 | 0.85 |
| Student A4 | 2.28 | 0.75 | 1.83 | 0.64 |
| Student B1 | 2.76 | 0.71 | 3.20 | 0.72 |
| Student B2 | 4.48 | 1.10 | 1.16 | 0.57 |
| Student B3 | 4.66 | 1.07 | 6.89 | 1.32 |
| Student B4 | 1.60 | 0.62 | -0.39$^{ns}$ | 0.52 |
| Student B5 | 2.77 | 0.71 | 0.83$^{ns}$ | 0.54 |
| **Combined Effect** | 2.47 | 0.25 | 1.75 | 0.23 |

$^{ns}$ not statistically significant at $p < .05$

An unstandardized treatment effect, *D*, was also calculated for each of the 18 AB phase contrasts. The mean unstandardized effects for ITSSIM and the other five methods is presented in Table 4, in addition to the mean standardized effects (for the methods that yielded an interpretable *d*-type effect size). Several of the multilevel models did not account for trend, autocorrelation, or both.

Table 4

*Six Analyses of Lambert et al. (2006) Data*

| Study | Unstandardized Effect Size | Standardized Effect Size |
|---|---|---|
| Shadish, Hedges, & Pustejovsky (2014) [a] | 5.46 | 2.51 |
| Shadish, Zuur, & Sullivan (2014) [b] | 6.70 | - |
| Rindskopf (2014) [a, b] | 5.70 | - |
| Moeyaert et al. (2014): *continuous outcome* | 5.10–5.76 (A1B1)<br>4.92–5.77 (A2B2) | - |
| Moeyaert et al. (2014): *logistic outcome* [a, b] | 5.61 | - |
| Swaminathan et al. (2014) | 5.38 (A1B1)<br>5.03 (A2B2) | 2.47 (A1B1)<br>2.34 (A2B2) |
| ITSSIM | 7.59 (A1B1)<br>8.47 (A2B2) | 2.47 (A1B1)<br>1.75 (A2B2) |

[a] Trend not modeled
[b] Autocorrelation not modeled

## Discussion

The purpose of this study is to introduce a computer simulation method, ITSSIM, for the measurement and meta-analysis of single-case experimental data. ITSSIM is a comprehensive effect size metric which incorporates baseline trend, level- and slope-change, within-phase error variance, and autocorrelation parameters. The simulation method assumes that one observed time-series may be explained by many plausible treatment effects and conditions. ITSSIM simulates many plausible conditions which could yield the observed data, and outputs the most likely treatment effect size based on the precision of the various parameter estimates. While ITSSIM can be used to analyze

data from a single case, its standardized mean difference effect size, *d*, can also be used to synthesize effects across cases and studies using standard meta-analytic methods.

Single-case data from Lambert et al.'s (2006) study were analyzed with five multilevel modeling methods in a special issue of the *Journal of School Psychology* (Shadish, 2014a). Effects size estimates from those methods were reasonably consistent, indicating an effective treatment which led to a drop in disruptive behavior. The Lambert et al. data were re-analyzed using ITSSIM software (Tarlow, 2017a) to determine how consistent the results from this simulation approach were to the other five methods.

**Comparison of Effect Size Indices**

The results presented in Table 4 suggest two tentative conclusions about ITSSIM compared to the other five effect size methods presented in the special issue. The standardized effects were quite similar—all methods that yielded a *d*-type effect size produced average effects between 1.75 and 2.51 standard deviations of improvement. However, when the unstandardized effects are compared, ITSSIM produced results that were somewhat larger than the other methods, by about one to three intervals of disruptive behavior.

These results indicate that larger treatment effects are more plausible under the ITSSIM simulation model than with the other estimation methods. However, the effect size estimates are less precise with ITSSIM—and less precision means larger standard errors, and therefore relatively smaller standardized effects (or in this case, standardized effects of comparable size). Put another way, using the ITSSIM framework, *the likeliest*

*estimate of treatment effect is relatively large, but so is the degree of uncertainty about that estimate*.

This outcome makes sense in light of ITSSIM's unique approach to parameter estimation. Whereas the other five methods take a fully-multilevel approach to effect size measurement and meta-analysis—and therefore pool information from all participants (using classical or Bayesian analysis) to yield one omnibus effect size estimate—ITSSIM instead estimates each effect independently, resulting in decreased precision. In addition, the nonparametric estimators used in the ITSSIM model make fewer assumptions than least squares methods, and may be more appropriate for some single-case time-series data; however, less strict assumptions come at a cost of statistical power.

ITSSIM also models baseline trend, level- and slope-change, and autocorrelation effects when estimating effect size—several of the comparison methods excluded one or more of those parameters. ITSSIM's comprehensive modeling approach may account for the larger unstandardized effect size estimate. The presence of un-modeled autocorrelation may inflate some parametric statistics (Manolov & Solanas, 2008), although serial dependency has been shown to attenuate nonparametric effect size indices (Tarlow, in press). The effect of un-modeled autocorrelation on multilevel methods has not been thoroughly investigated (Ugille et al., 2012).

The discrepancy in unstandardized effect size could also be in part due to some methods' failure to model trend and slope. This is true in particular for the A2B2 phase contrast, where ITSSIM yielded the largest effect and the multilevel methods gave

relatively small ones. The Lambert et al. (2006) data sets include an ABAB reversal design. A trend is visually apparent in many of the A2 phases as the participants' disruptive behavior returns to the baseline level of functioning. One argument for excluding trend from this analysis would be the expectation that disruptive behaviors will level off after they return to their baseline level. However, one could also argue that effect size indices should account for orthogonal slope changes in the A2B2 phase contrast—and that to ignore the influence of slope would lead to an underestimation of treatment effect. Ma (2006) and others have explored this very issue as a limitation of nonoverlap measures of effect size, in particular when applied to ABAB designs (Allison & Gorman, 1993; Schlosser et al., 2008).

## Conclusion

Additional study of ITSSIM is needed to determine if the findings of this field test generalize to other single-case effect size indices and data sets. However, it is encouraging that ITSSIM gave effect size estimates similar to the five multilevel methods reviewed in the *Journal of School Psychology* special issue. The similarities and differences among these results are interpreted with caution, because of the different analytic frameworks implemented by each statistical method. ITSSIM and the five multilevel methods are new, and not as well researched as older (often simpler) single-case statistics. Additional investigation is planned to compare ITSSIM some of those more established methods (see Chapters III and IV).

Given the tentative results of this field test, investigators may consider using ITSSIM to analyze their single-case data. The simulation-based method yields effect size

estimates that are easy to interpret, either as unstandardized or standardized measures of treatment effect. All ITSSIM effect sizes are presented in context of their baseline trend, level- and slope-change, error variance, and autocorrelation parameter estimates, which may further aid with the interpretation of experimental results. And unlike with multilevel methods, ITSSIM does not require the investigator to pool many single-case data sets into one omnibus treatment effect estimate. Perhaps most importantly, ITSSIM is available to clinicians and applied scientists as a user-friendly standalone application that does not require prior familiarity with statistical computing or syntax. Computer simulation procedures have the potential to address several of the challenges encountered in single-case data analysis, and ITSSIM is one application that may be useful to both researchers and practitioners.

# CHAPTER III

# COGNITIVE THERAPIES FOR DEPRESSION: A META-ANALYSIS OF SINGLE-CASE EXPERIMENTAL DESIGNS


## Abstract

Single-case experimental designs can offer unique contributions to the field of psychotherapy research. These flexible easily implemented designs can test treatment outcomes for individuals with complex symptom presentations (e.g., comorbidities) and low-incidence clinical populations—both of which are difficult to study with randomized controlled trials (RCTs). Single-case designs also offer a cost-efficient way of pilot-testing new treatments. Multiple methods for statistical analysis and meta-analysis of single-case designs have been proposed. This paper advocates for an approach that blends the sophistication of multilevel modeling with the utility of single-case-specific effect size statistics. Meta-analyses were performed on 53 cases from 10 single-case studies of cognitive therapy for depression. To compare the performance of different single-case statistics, all meta-analyses were replicated with six effect size indices. Overall, single-case studies of cognitive therapies were associated with substantial decreases in depressive symptoms, and effect sizes were similar to meta-analyses of RCT research ($0.58 \leq ES \leq 1.48$). As expected, evidence of treatment efficacy was even clearer when the moderating effect of bipolar diagnosis was included in analyses. These methods and results should encourage psychotherapy researchers to

consider single-case experiments when designing their studies. Two effect size statistics, Baseline Corrected Tau and Interrupted Time Series Simulation (ITSSIM), are also recommended for further study and development due to their superior performance.

## Introduction

Single-case experimental designs are an increasingly popular research method in many areas of psychology (Smith, 2012). Single-case designs—which use a combination of systematic repeated measurements and experimental controls to test treatment outcomes—are powerful because they can demonstrate the causal effects of interventions (APA Presidential Task Force on Evidence-Based Practice, 2006; Barlow & Hersen, 1984; Campbell & Stanley, 1963), often with a fraction of the resources demanded by high-powered randomized clinical trials (Barlow & Nock, 2009). Single-case experiments are also relatively easy to implement for practitioners who wish to demonstrate the efficacy of their treatments (Morgan & Morgan, 2001; Shadish et al., 2008).

As these designs grow in popularity, so too does the quantitative synthesis (i.e., meta-analysis) of single-case studies (Maggin et al., 2011). Indeed, single-case research is philosophically rooted in the replication and synthesis of experimental findings across many cases (Morgan & Morgan, 2001). Campbell and Stanley (1963) stated, "It should be remembered that … a single [time-series] experiment is never conclusive … [the design] is repeated in many different places by various researchers before a principle is established" (p. 42). However, the brief interrupted time-series data which characterize much of the single-case research literature are difficult to quantify and aggregate using

the statistical methods familiar to most investigators trained in large-*n* between-groups methods (Busk & Serlin, 1992; McCleary & Welsh, 1992). Many effect size statistics have been proposed for measuring single-case treatment effect sizes, but no method has been identified as clearly superior (Brossart et al., 2011; Campbell, 2004; Ma, 2006; Parker et al., 2005, 2006, 2011; Parker & Hagan-Burke, 2007; Manolov & Solanas, 2009, 2013; Manolov et al., 2011; Parker & Vannest, 2009; Parker, Vannest, & Davis, 2011; Shadish et al., 2014; Tarlow, in press; Vannest et al., 2012; Wolery et al., 2010). In addition, many of the statistics proposed for analyzing single-case studies are useful in clinical practice, but their lack of formal statistical development complicates quantitative synthesis across cases and studies (Shadish, 2014b; Shadish et al., 2008).

Systematic reviews of single-case meta-analyses found that investigators use a number of analytic methods to quantify and synthesize treatment effects (Beretvas & Chung, 2011; Maggin et al., 2011). Investigators often report multiple statistical indices for the same meta-analysis, given the lack of consensus about an ideal effect size metric. One challenge identified in single-case meta-analyses is the synthesis of results from complex experimental designs. For example, in an ABAB reversal design, should both AB phase contrasts be included? And if so, how should analysts model the statistical dependence of intra-subject effects? Some researchers have proposed multilevel modeling and Bayesian methods to address the within-subject and within-study dependencies which complicate single-case quantitative synthesis (Moeyaert et al., 2014; Rindskopf, 2014; Shadish, Zuur, & Sullivan, 2014; Swaminathan et al., 2014; van den Noortgate & Onghena, 2003, 2007, 2008).

Though multilevel methods have potential for organizing the emerging field of single-case meta-analysis, investigators would be remiss to ignore the essentially pragmatic nature of single-case research (Iwakabe & Gazzola, 2009; Fishman, 2005; Shadish & Rindskopf, 2007). One limitation of multilevel modeling is the statistical sophistication and interpretive nuance required to conduct these analyses. Multilevel methods may not be a realistic option for some single-case investigators, who are often practitioners first and clinical scientists second. Parker and Vannest (2012) described a distinction between "bottom-up" and "top-down" methods:

> The bottom-up strategy is distinct from a top-down strategy in which an overall or omnibus analytic model is fit to the entire design. Whereas top-down appears more elegant, it entails a marked risk, which is to ignore the idiosyncrasies or uniqueness of a design and its data patterns. It is true that any template can be modified, but to do so in [multilevel modeling], for example, requires statistical skills beyond those of most interventionists. This raises the broadest concern with the top-down analytic strategy because the behavior analyst is not able to maintain decision-making control, may not even be able to confirm the legitimacy of a model fit, and may not even be able to interpret the results. (p. 263)

To illustrate the practical limitations of sophisticated multilevel methods, consider that the Percentage of Nonoverlapping Data (PND; Scruggs, Mastropieri, & Casto, 1987) is the most popular statistic for synthesizing single-case research (Beretvas & Chung, 2011; Maggin et al., 2011). PND, which may be hand-calculated (or calculated via web application to attain p-values; Tarlow & Penland, 2016b), remains popular despite its well-documented statistical limitations (e.g., Allison & Gorman, 1993, 1994; Ma, 2006; Manolov & Solanas, 2009; Tarlow & Penland, 2016a; Wolery et al., 2010). Schlosser et al. (2008) stated, "When metrics are discussed in terms of their theoretical strengths and weakness alone, divorced from issues of implementation and

application, we jeopardize the capability of a particular metric to realize these strengths or perhaps minimize weaknesses, whatever they may be" (p. 184). One of the great strengths of single-case methodology is its potential to bridge the scientist-practitioner gap (Borckardt et al., 2008). Methodologists should be pragmatic when developing new techniques for measurement and meta-analysis of single-case designs—and always seek to maximize statistical applications for as wide a range of investigators as possible (Shadish, 2014b).

Single-case research and meta-analyses are implemented most in the fields of special education and behavior therapy (Maggin et al., 2011; Smith, 2012); however, these methods have the potential to contribute clinically useful knowledge to psychotherapy research (Hilliard, 1993; Iwakabe & Gazzola, 2009). Case-based methods are valuable to psychotherapists because they permit investigators to explore treatment effectiveness outside of the limiting and artificial circumstances that characterize many randomized clinical trials (Edwards et al., 2004; Persons & Silberschatz, 1998). Single-case experiments also allow investigators to conduct research with special populations or low-incidence disorders which are not numerous enough to be realistically studied with large-n group designs. Finally, single-case methods offer a cost-effective way of pilot-testing new therapies (Borckardt et al., 2008).

The goals of this paper are twofold. First, the paper will demonstrate how a meta-analysis of single-case psychotherapy studies may be performed—specifically, for single-case studies of cognitive treatments for depression. Case-based methods offer promising opportunities for psychotherapists and psychotherapy researchers; however,

meta-analyses of single-case studies of psychotherapy are rare. A literature search revealed no psychotherapy meta-analyses that utilized single-case research; indeed, neglect of the single-case literature was one criticism leveled at Smith et al.'s (1980) seminal meta-analysis of psychotherapy outcomes (Wilson & Rachman, 1983). The second goal of the paper is to compare six methods for meta-analyzing single-case research. Six statistics will be compared, and their practical and statistical strengths and limitations will be identified. The distributions of the six effect size statistics will be examined for possible ceiling and floor effects, and correlations will be explored to determine how well effect size estimates produced by the six metrics concur with each other.

**Cognitive Therapy for Depression**

The cognitive model which forms the foundation of modern cognitive psychotherapy is most associated with the work of Beck (Beck, 1967, 1976, 2005; Beck et al., 1979), who proposed that distortions in information processing lead to maladaptive cognitive structures (schemas) and, ultimately, the symptoms of many psychological disorders. Cognitive therapies typically include a restructuring of maladaptive schemas, with the goal of relieving emotional distress caused by misattributions and distortions.

Cognitive and cognitive-behavioral therapy (CBT) are among the most researched psychological treatments. Cognitive therapies are effective for treating unipolar depression and other psychiatric diagnoses (Butler et al., 2006), though not necessarily more effective than other psychotherapies (Wampold et al., 2002, 2017).

Despite the proliferation of cognitive therapies, there is ongoing demand for modified or expanded treatment modalities to address the unique needs of specific demographic and diagnostic groups (Beck, 2005; Szentagotai & David, 2010). Single-case experimental studies of novel therapy treatments, like the ones included in this meta-analysis, in many ways represent the cutting edge of psychotherapy research, with implications for future large-scale research and health care policy.

## Methods

### Selection of Studies

For this paper, articles of interest included single-case studies of cognitive treatments for depression. Published studies were identified for inclusion in the meta-analysis through a database search. The PsycInfo and MEDLINE databases were searched for peer-reviewed articles published between 2011 and 2016 with abstracts that included the following terms: ("single case" OR "single-case" OR "single subject" OR "single-subject" OR "time series" OR "time-series" OR "multiple baseline" OR "open case" OR "open-case") AND ("depression" OR "depressed" OR "depressive") AND ("cognitive").

The initial database search identified 99 records. Twenty-two articles were removed as duplicates, so 77 records were screened. Records were excluded from the meta-analysis for several reasons, the most frequent of which were: articles were not psychotherapy outcome studies (e.g., neuroimaging studies were excluded), articles lacked time-series data (i.e., were not true interrupted time-series experiments), articles did not include a measure of depression in their results or depression was assessed only

as a pretest/posttest measure, or articles included only one baseline time point. After reviewing and excluding database search records, ten articles with a total of 53 cases were identified for inclusion. Figure 6 illustrates the flow of studies searched and selected for meta-analysis, and Table 5 summarizes the included studies.

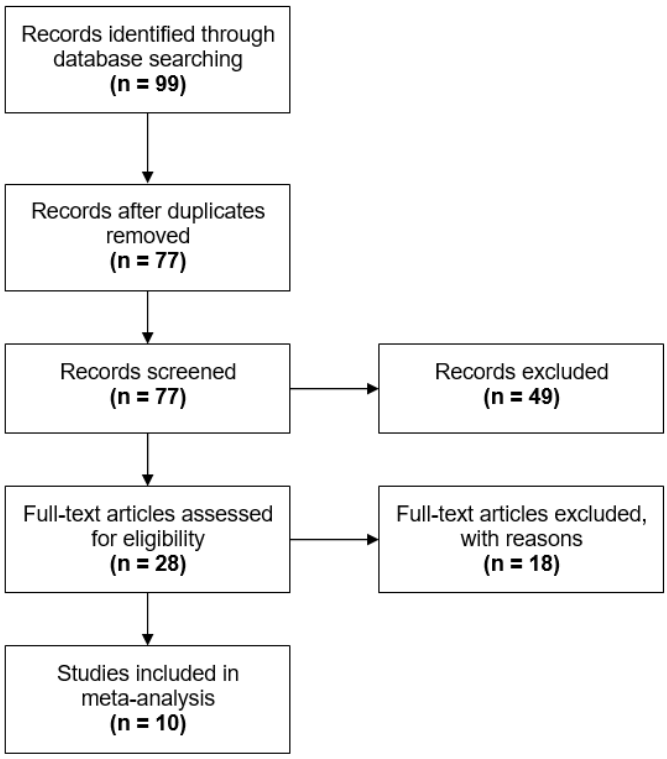



Figure 6. Selection flow diagram for meta-analysis.

Table 5

*Summary of Single-Case Studies of Cognitive Therapy Treatments*

| Study | Authors | *n* | Treatment | Participants | Outcome Measure |
|---|---|---|---|---|---|
| **A** | Holmes et al. (2016) | 14 | Imagery-focused cognitive therapy (Mood Action Psychology Programme; MAPP) | Bipolar I or II diagnosis; recruited from outpatient clinic | QIDS-SR |
| **B** | Searson et al. (2012) | 7 | Think Effectively About Mood Swings (TEAMS) | Bipolar I or II diagnosis; recruited from outpatient clinics | WSAS |
| **C** | Holländare et al. (2015) | 4 | Internet-based CBT (iCBT) | Bipolar II diagnosis; recruited from outpatient clinic | MADRS-S |
| **D** | Jones et al. (2015) | 4 | Telephone-based psychotherapy | Primary brain tumor diagnosis; recruited from hospitals, neurosurgery clinics, and cancer treatment services | DASS-21 |
| **E** | Mehranfar et al. (2012) | 4 | Mindfulness-Based Cognitive Therapy (MBCT) | Mothers of children with cancer with co-occurring depressive and anxiety disorders; recruited from radiology clinic | BDI-II |
| **F** | Akbari et al. (2015) | 3 | Transdiagnostic Treatment of Repetitive Negative Thinking (TTRNT) | Co-occurring anxiety and depressive diagnoses; recruited from university counseling center | BDI-II |
| **G** | Cowles & Nightingale (2015) | 1 | Transdiagnostic CBT | Co-occurring panic, anxiety, and depression; recruited from outpatient clinic | PHQ-9 |

Table 5 Continued

| Study | Authors | *n* | Treatment | Participants | Outcome Measure |
|---|---|---|---|---|---|
| **H** | McManus et al. (2014) | 6 | Transdiagnostic CBT | Co-occurring anxiety disorder diagnoses; recruited from outpatient clinic and website | BDI-II |
| **I** | Maroti et al. (2011) | 2 | CBT for insomnia (CBT-I) | Co-occurring insomnia, anxiety, and depressive disorder diagnoses; recruited from newspaper advertisement | BDI-II |
| **J** | Beck et al. (2016) | 8 | Cognitive Trauma Therapy for Battered Women (CTT-BW) | Women with PTSD diagnosis stemming from intimate partner violence; recruited from research clinic | BDI-II |

*Note:* BDI-II = Beck Depression Inventory, Second Edition (Beck, Steer, & Brown, 1996); DASS-21 = Depression, Anxiety, and Stress Scale (Lovibond & Lovibond, 1995); MADRS-S = Montgomery-Asberg Depression Rating Scale – Self-rated (Montgomery & Asberg, 1979); PHQ-9 = Patient Health Questionnaire-9 (Kroenke et al., 2001); QIDS-SR = Quick Inventory of Depressive Symptomology Self-Report (Rush et al., 2003); WSAS = Work and Social Adjustment Score (Marks, 1986).

Eight of the studies (A-H) were pilot studies of novel treatments, including telepsychology interventions (C and D) and transdiagnostic CBT protocols (G, H, and I). The remaining two studies (I and J) were conducted to extend and replicate earlier findings on new treatments. Half of the studies (D-H) were specifically designed to target clinical populations with complex symptom presentations, such as co-occurring depression and anxiety diagnoses. This is notable given that RCTs are frequently criticized for excluding participants with comorbid disorders—despite the fact that a majority of individuals with psychiatric diagnoses meet criteria for multiple disorders (Castelnuovo et al., 2004; Garfield, 1996; Seligman, 1995; Westen et al., 2004).

Raw data was digitally extracted from graphs in the ten selected studies in order to re-analyze and synthesize results. Data were extracted using GetData Graph Digitizer (2013). Digitization of published time-series graphs is a common procedure in single-case meta-analysis (Maggin et al., 2011) and methodology research with acceptable reliability and validity (Shadish et al., 2009).

**Effect Size Statistics**

Six effect size statistics were selected to analyze and synthesize results from the ten articles. Many single-case meta-analyses report multiple statistical effects, although it is rare for studies to report three or more effect sizes (Maggin et al., 2011). Effect size indices were selected to represent the methodological diversity of the field. Two nonparametric effect size statistics were included, Baseline Corrected Tau (Tarlow, in press) and the Mean Phase Difference (Manolov & Solanas, 2013). Two fully parametric regression-based methods were included, an ordinary least squares (OLS) regression

model (Center, Skiba, & Casey, 1985-1986; Allison & Gorman, 1993; Huitema & McKean, 1998, 2000b) and White et al.'s (1989) standardized mean difference *d* statistic. Two computer-intensive simulation-based methods were also included, Simulation Modeling Analysis (SMA; Borckardt et al., 2008) and Interrupted Time-Series Simulation (ITSSIM; see Chapter II). Table 6 illustrates some similarities and differences between the indices.

Table 6

*Summary of Effect Size Statistics Included in Meta-Analysis*

| **Effect Size** | **Estimation Method** | **Models Trend** | **Models Autocorrelation** |
|---|---|---|---|
| Baseline Corrected Tau | Nonparametric | Yes; Theil-Sen regression | No; but effect size is robust to moderate autocorrelation |
| MPD | Nonparametric | Yes; first-order differencing | No; but effect size is robust to moderate autocorrelation |
| OLS $R^2$ | Parametric | Yes, OLS regression | No |
| White et al.'s *d* | Parametric | Yes; OLS regression | No |
| SMA | Parametric; computer simulation | No | Yes; but only in *p*-value calculation |
| ITSSIM | Both nonparametric and parametric estimators used; computer simulation | Yes; Theil-Sen regression | Yes |

**Baseline Corrected Tau.** Tarlow (in press) proposed Baseline Corrected Tau as an improvement to Tau-U (Parker, Vannest, & Davis, 2011; Parker, Vannest, Davis, & Sauber, 2011). Tau-U has been popularized as an effect size for single-case research because of its conceptual ties to other single-case statistics (Parker et al., 2007; Parker & Vannest, 2009), ease of access via an online calculator (Vannest, Parker, & Gonen, 2011), and suitability for meta-analysis (e.g., Bowman-Perrott et al., 2013). However, Tarlow demonstrated that Tau-U can frequently give "out-of-bounds" results that are difficult to meaningfully interpret, and Tau-U did a poor job of controlling trend under a variety of simulation models.

Baseline Corrected Tau is similar to Tau-U in that it yields a nonparametric rank correlation effect size, Tau. However, the method of modeling baseline trend differs. In Baseline Corrected Tau, the existence of monotonic baseline trend is first tested with a Tau correlation of the baseline phase data with a session/day/time variable. If there is sufficient evidence for monotonic baseline trend, e.g., $p < .05$, then the investigator proceeds to the trend correction step. A nonparametric Theil-Sen regression line is fitted to the baseline data. Detrended data are then calculated as the residuals of both phases from the baseline Theil-Sen regression line. The detrended data—or, if there was no evidence of baseline trend, the original data—are rank correlated with a dummy code phase variable (A phase = 0, B phase = 1) to yield a Tau effect size.

Baseline Corrected Tau is useful because it is suitable for non-normally distributed data. Although Baseline Corrected Tau does not model autocorrelation, simulation research indicated it was robust even to very high levels of it (Tarlow, in

press). A weakness of Baseline Corrected Tau is its low power when testing baseline trend in small samples. When small or moderate trend is present in very brief time series ($N_A < 10$), it is unlikely that a rank correlation will detect it, and the data will be analyzed under the assumption of no trend (leading to results that may offer misleading conclusions about the efficacy of the treatment). This is not only a limitation for Baseline Corrected Tau. Some single-case statistics deal with the low power problem by either assuming no baseline trend (see SMA below) or assuming baseline trend (see MPD below), and in both cases the statistical model may be improperly specified for a given data set. Baseline Corrected Tau is different from the other analyses included in this study because it includes a decision tree whereby a baseline trend correction is performed only when sufficient evidence suggests the trend exists.

Tau coefficients may be combined using a variance estimator proposed by Kendall (1962). The variance of Tau will vary based on the underlying distribution of data. However, Kendall showed that the statistic's variance cannot exceed $(2 / n)(1 - \text{Tau}^2)$. Tarlow (in press) recommended this variance estimator be used as a conservative effect size weight in meta-analyses of single-case data due to the unknown distributions of brief interrupted time series data. Tarlow's (2016) online Baseline Corrected Tau calculator (available at http://ktarlow.com/stats/tau) calculates a Tau effect size from raw data as well as its standard error, $SE_{\text{Tau}}$, which may be used for meta-analysis.

**Mean Phase Difference (MPD).** Manolov and Solanas (2013) proposed the Mean Phase Difference (MPD) effect size, which compares projected baseline trend data with observed B phase data to yield a statistic which accounts for both trend and level

changes. MPD uses a nonparametric first-order differencing procedure to calculate a baseline slope coefficient, where the slope, $b$, is the average of $N_A - 1$ differences, $Y_i - Y_{i-1}$. Projected B phase data is then calculated as $y_{n_A+i} = y_1 + b(n_A + i - 1)$. The mean of the projected B phase is subtracted from the mean of the observed B phase. MPD is useful because it is easy to calculate and interpret (as difference between projected and actual B phase data). MPD is not affected by autocorrelation and may be of particular use to practitioners who want a straightforward effect size index that quantifies treatment effects in the unstandardized metric of the outcome variable (Manolov & Solanas, 2013; Tarlow, in press).

A limitation of MPD is its lack of statistical development for meta-analysis. The MPD effect size, which is a mean difference, may be standardized by dividing the mean difference by the variability around estimated trend lines.[3] However, its distribution is unknown. Rather than weight MPD effects by their inverse variance (which is not defined), they are weighted by the total number of time series observations, $N_A + N_B$, following Faith, Allison, and Gorman's recommendation (1993).

**Ordinary Least Squares (OLS) $R^2$.** A number of similar regression models have been proposed to model level- and slope-change effects in brief interrupted time series experiments (Center, Skiba, & Casey, 1985-1986; Allison & Gorman, 1993; Huitema & McKean, 1998, 2000b). There are slight differences among these

[3] For this study, an additional trend line was fitted to the observed B phase data for the sole purpose of calculating a standardizing term for the standardized MPD. First-order differencing was applied to B phase data just as in the baseline trend estimation step. Difference scores were then calculated within each phase as the observed score minus the predicted score (based on the fitted line). A standardizing term was then calculated as the standard deviation of the difference scores. This follows the standardizing procedure recommended by Maggin et al. (2011) and Swaminathan et al. (2010).

approaches, but the models are identical in most ways. Observed data is regressed onto four predictor terms, $\beta_0$, $\beta_1$, $\beta_2$, $\beta_3$, which correspond to baseline level, baseline trend, level-change, and slope-change, respectively. The regression model is: $Y = \beta_0 + \beta_1 T + \beta_2 D + \beta_3 S$, where $T$ is a session/day/time variable, $D$ is a dummy code phase variable (A phase = 0, B phase = 1), and $S$ is a slope change variable equal to $[T - (n_A + 1)]D$.[4] This method yields an effect size, $R^2$, interpreted as the variance accounted for by level- and slope-change effects. An adjusted $R^2$ was then calculated to account for small sample bias (Faith et al., 1996; Manolov & Solanas, 2008). The adjusted $R^2$ may then be converted to a standardized mean difference, $d$, suitable for meta-analysis.

OLS methods are sensitive to serial dependency. While standard regression methods may be suitable for analysis and meta-analysis when autocorrelation is minimal, $R^2$ is linearly correlated with lag-1 autocorrelation (Manolov & Solanas, 2008, Tarlow, in press). This means that, as autocorrelation increases, so do the estimates of effect size (regardless of the true effect). An analysis of autocorrelation in the depression data sets was conducted, described below, in order to evaluate the influence of serial dependency on effect size estimates. The detection of substantial autocorrelation in the analyzed data would raise concerns about the validity of the OLS $R^2$ statistic.

**White et al.'s *d*.** White, Rusch, Kazdin, and Hartmann (1989) offered a standardized mean difference statistic, $d$, for combining single-case effect size estimates. White et al.'s $d$ is useful because it models trend effects in a straightforward way. First,

[4] Specification of the slope change term, $S$, is a point of difference among the various OLS models. Huitema and McKean (2000) gave a thorough examination of the slope change issue, and their model specification is used here.

regression lines are fitted to the data in each phase. Then, a data point is predicted for the last day of the B phase using each regression line. This is similar in concept to MPD, where a baseline phase trend line is used to make predictions in the B phase. In White et al.'s *d*, the two predicted values are differenced, then divided by a standardizing term which includes the phases' pooled standard deviation controlling for across-phase trend. The equation for *d* is

$$d = \frac{\hat{y}_B - \hat{y}_A}{\sqrt{(1 - r^2)\sqrt{(s_A^2 + s_B^2)/2}}}$$

where $\hat{y}_B$ is the outcome variable prediction of the last day of the B phase made with the B phase regression line, $\hat{y}_A$ is the prediction made with the A phase regression line, $s_A^2$ is the A phase variance, $s_B^2$ is the B phase variance, and *r* is the Pearson correlation between the observed outcome variable and the session/day/time variable.

White et al.'s *d* does not model autocorrelation, and simulation research by Manolov and Solanas (2008) demonstrated its sensitivity to data with an autoregressive error structure. Like the regression-based OLS $R^2$ (Center, Skiba, & Casey, 1985-1986; Allison & Gorman, 1993; Huitema & McKean, 1998, 2000b), White et al.'s *d* was linearly associated with autocorrelation, though not as severely. The autocorrelation analysis described below was designed to determine the level of serial dependency present in the depression treatment data sets, and to what degree autocorrelation predicted the various effect size statistics.

**Simulation Modeling Analysis (SMA).** Borckardt et al. (2008) proposed Simulation Modeling Analysis (SMA) to measure treatment effects in brief interrupted

time series experiments. SMA yields a Pearson *r* correlation between observed data and a dummy code phase variable (A phase = 0, B phase = 1). However, because hypothesis testing with *r* and other OLS statistics is distorted by serial dependency, SMA uses computer simulation methods to account for autocorrelation in the adjusted *p*-value results. To accomplish this, the first-order autoregressive coefficient, $\varphi_1$, is estimated across both A and B phases. Then, thousands of artificial time series are simulated with the programmed level of autocorrelation, $\varphi_1$, with series lengths of $N_A + N_B$. Each simulated time series has no programmed level- or slope-change effects, only the effect of autocorrelation. Each artificial time series in this null effect distribution is then correlated with the dummy code phase variable to yield *r*. The resulting distribution of *r* values is then used to conduct a null hypothesis significance test with the observed effect, i.e., the *r* correlation of the observed data and the dummy code variable. SMA is accessible to investigators via a standalone software application (available at http://www.clinicalresearcher.org/software.htm).

A limitation of SMA is it does not explicitly model baseline trend and slope-change effects. The SMA software allows the investigator to test several different slope vectors to determine if within- or across-phase trend is present. However, due to the brief nature of many single-case experiments, multiple slope models often fit one data set. Effect size values derived from different slope models may also not be directly comparable. For example, one hypothetical data set with phase lengths $N_A = 5$ and $N_B = 5$ may have an effect size $r = 0.80$ when correlated with a dummy code phase variable {0, 0, 0, 0, 0, 1, 1, 1, 1, 1}. A second hypothetical data set of equal phase lengths may

also have an effect size $r = 0.80$, but only when correlated with the slope vector {1, 2, 3, 4, 5, 6, 7, 8, 9, 10} (Slope Vector 4 in the SMA software). These *r* values cannot be interpreted as similar sized effects; the first one suggests an association between outcome and treatment, whereas the second suggests an association between outcome and time (i.e., baseline trend).

Another limitation of SMA is that autocorrelation is estimated without controlling trend—a practice which can lead to uninterpretable results (Huitema & McKean, 1998). SMA also incorporates autocorrelation estimates only into probability testing (i.e., *p*-values). The reported effect size is not adjusted even when large degrees of serial dependency are present in the data, which may lead to interpretation problems.

**Transformation and Comparison of Effect Size Statistics**

The six statistics yield effect size estimates in different metrics. Three of the indices yield correlation statistics as results (Baseline Corrected Tau, SMA, and ITSSIM). The OLS regression model gives a variance-accounted-for $R^2$ as an effect size. White et al.'s method yields a standardized mean difference *d* statistics. And MPD gives an unstandardized effect, where the effect of treatment is expressed in units of the original outcome measure. In order to help with comparison and interpretation, the OLS regression model, White et al.'s *d*, and MPD were transformed to correlation coefficients, which can be interpreted as the strength of association between treatment and outcome (controlling for trend, etc.). Correlation-based measures of effect are in some ways easier to interpret with single-case research because effect sizes are bound between -1 and +1; standardized mean difference *d*-type measures often produce single-

case effects that are quite large and hard to interpret (Jenson et al., 2007; Parker et al., 2005).

Effect size estimates from the six indices were tested with correlations to determine how well the metrics concurred with each other. In general, single-case statistics have demonstrated only moderate agreement (Smith, 2012). The distributions of each effects size measure were also visualized with histograms to detect any floor or ceiling effects, which are known to occur frequently with non-overlap and other methods (Tarlow, in press).

**Meta-Analysis**

Multilevel models are a natural fit for single-case research synthesis, where treatment effects are hierarchically nested within studies and cases. For example, one study may have several cases, and one case may have more than one treatment effect (as in an ABAB reversal design, where there are two phase contrasts). As discussed above, multilevel modeling has been critiqued for being too opaque to the majority of scientists and practitioners conducting single-case research (Parker & Vannest, 2012).

This study aims to find a happy medium between fully-multilevel methods and designs which do not account for random within-group or within-study effects (e.g., Bowman-Perrott et al., 2013; Parker, Vannest, & Davis, 2011; Vannest et al., 2010). In fully-multilevel methods, raw data fitted to an omnibus hierarchical model. However, this approach often requires the investigator to test multiple models, and excess data are frequently discarded. On the other hand, single-level (fixed effects) meta-analyses, which comprise most of the single-case quantitative synthesis literature, combine effect

size estimates without consideration of the statistical interdependence of effects within cases or studies. Treating nested within-group effects as independent when they are not can lead to misleading results and drastically inflated error rates (Cook, 2000). In this paper, individual effects were obtained from the six effect size indices, and the multilevel analysis is used only to aggregate those results across studies.

**Within-study meta-analysis.** A one-level fixed-effects model was used to estimate the average treatment effect for the cases within each study. It should be noted that within-study treatment effects will differ when estimated from a multilevel model, as described below. However, it is useful to consider how one investigator might aggregate their own results—without access to data from other single-case studies. The studies of cognitive treatments for depression included in the meta-analysis were all multiple baseline studies with a single baseline-to-intervention AB phase contrast (i.e., a single treatment effect per case). Therefore, there was no need to add an additional within-case level to account for multiple contrasts.

The fixed-effects model used was:

$$ES_i = \beta_0 + e_i$$

where $ES_i$ is the effect size for case $i$ in the study, $\beta_0$ is the overall treatment effect for the study, and $e_i$ is the difference between the case $i$ effect size and the study's overall effect. The $e_i$ term is assumed to be random, independent, and normally distributed.

The R software (R Core Team, 2016) package 'meta' (Schwarzer, 2007) was used to calculate all fixed-effect means. The package includes the 'metacor' function, which performs fixed- and random-effects meta-analyses with correlation data.

**Multilevel meta-analysis.** The mixed model used to synthesize cases for an overall omnibus effect size assumes hierarchically clustered data with two levels, where cases (and their treatment effect sizes) are nested within studies. This is essentially a random-effects ANOVA, which yields an overall effect size. The hierarchical model used was:

Level 1 (within study) $ES_{ij} = \beta_{0j} + e_{ij}$

Level 2 (between studies) $\beta_{0j} = \theta_{00} + u_{0j}$

where $ES_{ij}$ is the treatment effect size for case $i$ in study $j$, $\beta_{0j}$ is overall effect size for study $j$, $e_{ij}$ is the difference between the case $i$ effect size and that case's study $j$ overall effect, $\theta_{00}$ is the overall effect size for all studies, and $u_{0j}$ is the difference between the study $j$ effect and the overall effect size for all studies. The $e_{ij}$ and $u_{0j}$ terms are assumed to be random, independent, and normally distributed.

Note that Level 1 of the mixed model is identical to the within-study meta-analysis above. However, all within-study treatment effects, $\beta_{0j}$, are distributed around an overall effect size for all studies, $\theta_{00}$.

The SAS PROC MIXED routine was used to perform two-level meta-analyses (SAS Institute, 2012). Singer (1998) offered a thorough tutorial for using SAS in multilevel modeling. Figure 7 presents the syntax used for this study.

```
proc mixed covtest;
  es study;
  model es = /solution cl;
  random intercept/subject=study solution cl;
run;
```

Figure 7. Syntax for multilevel modeling with SAS PROC MIXED.

**Potential moderator effects.** Moderators are variables that affect the strength and/or direction of effect between predictor and dependent variables (Baron & Kenny, 1986). Moderators of interest to researchers are often categorical variables (e.g., gender, race, class) that divide individuals into groups that are meaningful for the treatment under study.

Three potential moderators were initially identified after reviewing the ten articles included in this paper. The moderators were: bipolar vs. unipolar depression, telepsychology versus in-person treatment, and transdiagnostic CBT versus other treatment modality. Only the bipolar diagnosis moderator had sufficient power for further analysis, as the other variables had too few cases to reliably fit a hierarchical model. Three of the studies (A-C) tested treatments of individuals with a bipolar disorder, which included 25 out of the 53 total cases.

There is strong evidence for cognitive/cognitive-behavioral therapy's efficacy when treating individuals with unipolar depression and anxiety (Butler et al., 2006). However, bipolar disorder is more challenging to treat with psychotherapy (Beck, 2005).

Scott (1995) identified three obstacles to treating bipolar disorder: the disorder has strong genetic/biological correlates which suggest it is more biologically based than other affective disorders; there is a false belief among many mental health care providers that individuals with bipolar disorder make a full inter-episode recover, and thus would not benefit from treatment; there is historically a substantial amount of stigma about individual with bipolar disorder, who are seen as poor candidates for psychotherapy. Even researchers who have produced evidence of psychotherapy's efficacy for bipolar disorder acknowledge, "new CBT strategies are needed to increase and enrich the impact of CBT at posttreatment and to maintain its benefits" (Szentagotai & David, 2010, p. 66). The three single-case studies of psychotherapeutic interventions for bipolar (A-C) are good examples of treatment innovation.

**Weights.** Effect size estimates were weighted by their combined phase lengths, $n_A + n_B$, in the within-study (one-level) meta-analyses, following the suggestion of Allison and Gorman (1996). By weighting estimates in this way, effects that were estimated from longer experiments were treated as more reliable indicators of overall treatment effect. Another possible weighting scheme not investigated here would be to weight effects by their baseline phase length only, $n_A$.

A limitation of the two-level omnibus analysis was that it did not permit case-specific weights, as did the one-level model. Weighting schemes are possible in multilevel meta-analysis (Cheung, 2015). However, they require the same number of effects within each cluster, i.e., the same number of cases within each study.

**Autocorrelation**

Autocorrelation is known to influence the effect size estimates of some single-case statistics. The degree to which autocorrelation is present in single-case behavioral data has been debated at length (e.g., see the special issue of *Behavioral Assessment*; Baer, 1988; Busk & Marascuilo, 1988; Huitema, 1985, 1988; Sharpley & Alavosius, 1988; Suen, 1987; Suen & Ary, 1987; Wampold, 1988), with an emerging consensus that serial dependency should not be ignored. How exactly investigators should handle autocorrelation is still unclear. One approach involves identifying statistics which are less affected by autocorrelation (Manolov & Solanas, 2008; Parker, Vannest, Davis, & Sauber, 2011; Tarlow, in press). A limitation to this approach is that serial dependency may nonetheless affect the underlying distribution of these effect size indices, even when the size of effect is robust to autocorrelation (Shadish, 2014b). For example, a statistic could hypothetically report the same effect size for a series with no autocorrelation and another series with a high degree of autocorrelation present—however, the standard errors of those effects (and their relative probabilities) may not be equal. A second approach is to statistically control autocorrelation, much in the same way that baseline trend is statistically controlled in some of the methods discussed above (Parker et al., 2006, 2011). Unfortunately, autocorrelation estimation is very imprecise in brief interrupted time-series, and "corrected" data may do a poor job of reflecting the experimental treatment effect (see Chapter II).

To examine the degree and effect of autocorrelation in this study's depression data sets, $r_1$ (the lag-1 autocorrelation coefficient) was estimated for each of the 53

analyzed cases. Autocorrelation estimates were calculated from regression residuals, not raw data, as level changes and trend effects greatly distort $r_1$ (Huitema & McKean, 1998). The standard autocorrelation estimator is biased in small samples, so a corrected estimator was used (Huitema & McKean, 1991; Ferron, 2002):

$$r_1 = \frac{\sum_{t=2}^{N}(e_t)(e_{t-1})}{\sum_{t=1}^{N} e_t^2} + \frac{P}{N}$$

where $N$ is the number of observations, $e_t$ is the residual at time $t$, and $P$ is the number of estimated parameters in the regression model used to extract residuals (in this case, $P$ = 4).

Effect size estimates were then correlated with $r_1$ to test for any association between treatment effect and serial dependency. Some single-case statistics, including the OLS regression method and White et al.'s (1989) *d* statistic, have demonstrated a linear relationship with $r_1$ in simulation studies (Manolov & Solanas, 2008; Tarlow, in press).

Autocorrelation estimators unfortunately have poor precision with brief time series. Even when a single-case data set yields a large $r_1$ estimate, quite often the $r_1$ value has very large standard errors and is not statistically significant. To increase the power of autocorrelation detection in the present study, $r_1$ estimates were meta-analytically combined using a random-effects model. In a large survey of single-case research (not limited to depression treatments), Shadish and Sullivan (2011) used a similar method to test for autocorrelation. Their analysis of 800 time-series from 113 studies yielded a mean lag-1 autocorrelation of $r_1 = 0.20$ ($p < 0.001$); the mean effect

also had a statistically significant variance component, indicating that the level of autocorrelation was not homogenous across all studies.

## Results

After transforming all effect size estimates to correlation coefficients, the relative associations between statistical methods was examined via correlations. The correlations of the six indices and the lag-1 autocorrelation estimates are presented in Table 7. The six statistics demonstrated moderate to strong agreement, $r$ values ranging from 0.52 to 0.93. All correlations between effect size statistics were statistically significant, $p < 0.05$. On the other hand, there was no evidence of association between any effect size statistic and the lag-1 autocorrelation estimate. This result is considered cautiously, given simulation research which has demonstrated least squares regression-based statistics like the OLS model and White et al.'s method are sensitive to lag-1 autocorrelation (Tarlow, in press; Manolov & Solanas, 2008). This study may have lacked the sufficient power to detect these associations.

Table 7

*Pearson Correlation Matrix of Effect Size Statistics and Lag-1 Autocorrelation for 53 Single-Case Studies of Cognitive Depression Treatments*

| Effect Size Statistic | (1) | (2) | (3) | (4) | (5) | (6) | $r_1$ |
|---|---|---|---|---|---|---|---|
| **(1) Baseline Corrected Tau** | - | 0.57 | 0.75 | 0.66 | 0.86 | 0.72 | 0.05* |
| **(2) MPD** | 0.57 | - | 0.83 | 0.90 | 0.52 | 0.87 | 0.00* |
| **(3) OLS** | 0.75 | 0.83 | - | 0.89 | 0.65 | 0.90 | 0.05* |
| **(4) White et al.** | 0.66 | 0.90 | 0.89 | - | 0.54 | 0.93 | 0.03* |
| **(5) SMA** | 0.86 | 0.52 | 0.65 | 0.54 | - | 0.57 | 0.00* |
| **(6) ITSSIM** | 0.72 | 0.87 | 0.90 | 0.93 | 0.57 | - | -0.03* |
| $r_1$ | 0.05* | 0.00* | 0.05* | 0.03* | 0.00* | -0.03* | - |

* not statistically significant, $p < .05$.

A review of the Table 7 results suggests that the effect size statistics fell into two distinct groups. The first group, which included Baseline Corrected Tau and SMA, were strongly associated, $r = 0.86$. These two methods are similar in that they do not model baseline trend as aggressively as the other four statistics. SMA does not model trend at all. Baseline Corrected Tau uses a nonparametric robust regression method (Theil-Sen) to correct for trend, but only when there is a statistically significant trend in the baseline phase—in contrast with the remaining four statistics which remove the estimated baseline trend from all time-series automatically. Only eight out of the 53 cases (15%) had statistically significant baseline trend using the Baseline Corrected Tau method; the remaining 45 effects were estimated without corrected for baseline trend. This offers at

least a partial explanation for why this statistic was strongly associated with SMA, which used no trend correction on all 53 cases.

The second group of statistics, including MPD, OLS regression, White et al., and ITSSIM, were also strongly associated, with *r* values from 0.83 to 0.93. The high level of agreement could be described by the nature of the trend correction procedure, which is similar with all four methods. In each case, a trend line is fitted to the baseline data and used to make predictions about how the participant would continue to change through the duration of the B phase—and predicted "null effect" values are compared to the observed treatment effect to find an effect size. The correlations across the two sets of statistics ("Conservative/No Baseline Correction" versus "Aggressive Baseline Correction") further suggest these are distinct groups, with relatively moderate *r* values from 0.52 to 0.75.

Effect size estimates for each of the six statistics were then meta-analytically combined within and between studies. The mean effect sizes are summarized in Table 8. The consistency of the statistics when applied to meta-analysis was mixed. The magnitude of mean effects varied by statistic, with Baseline Corrected Tau and ITSSIM giving more moderate mean effects on average. On all but two of the within-study meta-analyses (B and J), the six statistics gave mean effects which concurred on the direction of treatment effect, i.e., mean effect sizes where all negative (indicating improvement) or positive (indicating deterioration of mood). On only three of the studies (A, E and F) did all six statistics give mean effect estimates that were statistically significant. Of those three studies, one (A) had the largest number of cases, and the other two (E and F) had

the largest treatment effects. It is therefore unsurprising that these three studies would have statistically significant mean effects, as statistical power increases with both the number of cases and effect size.

When all effects were combined with the two-level model ($n$ = 53), the six means once again agreed on the direction of average treatment effect, which was negative (treatment associated with a decrease in depression). Mean omnibus effects ranged from -0.279 to -0.743. However, only three of the mean effects were statistically significant: Baseline Corrected Tau, OLS regression, and SMA.

The statistics were most consistent when the moderator effect, bipolar diagnosis, was taken into account. None of the methods yielded a statistically significant mean treatment effect for the bipolar treatment studies (A, B, and C). On the other hand, statistically significant treatment effects were detected in the non-bipolar studies (D-J), with effects ranging from -0.380 to -0.936. This finding would suggest that single-case treatments for depression are effective with individuals not diagnosed with a bipolar disorder; there is less evidence of treatment efficacy for individuals diagnosed with a bipolar disorder.

Table 8

*Mean Effect Sizes by Study with 95% Confidence Intervals*

| Study | *n* | BC-Tau | MPD *r* | OLS *r* | White et al.'s *r* | SMA *r* | ITSSIM *r* |
|---|---|---|---|---|---|---|---|
| **One-Level Model (within-study)** | | | | | | | |
| **A** (Holmes) | 14 | **-0.297** [-0.362, -0.228] | **-0.497** [-0.550, -0.439] | **-0.327** [-0.391, -0.260] | **-0.586** [-0.632, -0.536] | **-0.475** [-0.523, -0.416] | **-0.459** [-0.515, -0.399] |
| **B** (Searson) | 7 | -0.159 [-0.353, 0.049] | 0.129 [-0.079, 0.327] | 0.135 [-0.073, 0.332] | **0.461** [0.282, 0.609] | -0.185 [-0.377, 0.021] | **0.410** [0.223, 0.568] |
| **C** (Holländare) | 4 | 0.002 [-0.372, 0.374] | **0.836** [0.672, 0.922] | 0.297 [-0.086, 0.603] | **0.942** [0.877, 0.973] | -0.083 [-0.442, 0.300] | **0.567** [0.246, 0.776] |
| **D** (Jones) | 4 | -0.164 [-0.425, 0.123] | **-0.548** [-0.712, -0.316] | **-0.332** [-0.561, -0.056] | **-0.980** [-0.988, -0.964] | -0.221 [-0.473, 0.064] | **-0.318** [-0.550, -0.040] |
| **E** (Mehranfar) | 4 | **-0.735** [-0.881, -0.464] | **-0.933** [-0.971, -0.846] | **-0.956** [-0.981, -0.898] | **-0.993** [-0.997, -0.984] | **-0.907** [-0.960, -0.790] | **-0.748** [-0.887, -0.485] |
| **F** (Akbari) | 3 | **-0.596** [-0.767, -0.346] | **-0.922** [-0.958, -0.855] | **-0.977** [-0.988, -0.957] | **-0.997** [-0.998, -0.994] | **-0.709** [-0.837, -0.507] | **-0.912** [-0.953, -0.837] |
| **G** (Cowles) | 1 | -0.344 [-0.821, 0.415] | **-0.928** [-0.985, -0.687] | **-0.917** [-0.983, -0.646] | **-0.979** [-0.996, -0.900] | -0.654 [-0.919, 0.018] | **-0.778** [-0.951, -0.236] |
| **H** (McManus) | 6 | **-0.261** [-0.459, -0.038] | -0.188 [-0.396, 0.040] | **-0.637** [-0.754, -0.480] | **-0.712** [-0.808, -0.580] | **-0.296** [-0.488, -0.075] | **-0.373** [-0.552, -0.161] |
| **I** (Maroti) | 2 | **-0.546** [-0.827, -0.047] | **-0.674** [-0.882, -0.248] | **-0.720** [-0.900, -0.330] | **-0.911** [-0.970, -0.747] | **-0.751** [-0.912, -0.388] | -0.479 [-0.796, 0.044] |
| **J** (Beck) | 8 | -0.016 [-0.200, 0.168] | -0.146 [-0.322, 0.039] | **-0.198** [-0.368, -0.014] | **-0.269** [-0.432, -0.090] | **-0.284** [-0.444, -0.105] | **0.218** [0.035, 0.386] |

Table 8 Continued

| Study | *n* | BC-Tau | MPD *r* | OLS *r* | White et al.'s *r* | SMA *r* | ITSSIM *r* |
|---|---|---|---|---|---|---|---|
| **Two-Level Model (between-studies)** | | | | | | | |
| All Studies | 53 | **-0.281** [-0.450, -0.092] | -0.433 [-0.791, 0.146] | **-0.595** [-0.859, -0.083] | -0.743 [-0.968, 0.148] | **-0.475** [-0.687, -0.187] | -0.279 [-0.631, 0.170] |
| **Moderator Effects** | | | | | | | |
| Bipolar diagnosis | | | | | | | |
| Yes (A, B, C) | 25 | -0.205 [-0.607, 0.281] | 0.192 [-0.960, 0.981] | -0.071 [-0.762, 0.696] | 0.341 [-0.991, 0.998] | -0.305 [-0.734, 0.296] | 0.092 [-0.919, 0.943] |
| No (D, E, F, G, H, I, J) | 28 | **-0.380** [-0.631, -0.055] | **-0.681** [-0.907, -0.152] | **-0.789** [-0.949, -0.300] | **-0.936** [-0.994, -0.460] | **-0.583** [-0.819, -0.181] | **-0.529** [-0.818, -0.028] |

*Note:* Bracketed terms indicate 95% confidence intervals. Negative effect sizes indicate a decrease in depression, i.e., improved mood. **Bold** effect size estimates are statistically significant, $p < .05$

Distributions of the six effect size statistics are illustrated in Figure 8. Only the non-bipolar studies (D-J) are included in the figure. White et al.'s effect size demonstrated severe floor and ceiling effects, suggesting it does a poor job of discriminating between effects of different sizes (nearly all cases fell near +1.000 or -1.000). This replicates a similar finding by Parker et al. (2005). The ceiling effect is due to the extreme *d* values produced by White et al.'s method, which were transformed to *r* statistics during the meta-analysis; 45 out of 53 cases (85%) had a *d* with an absolute value greater than 4; 28 cases (53%) had an absolute *d* value greater than 10. The OLS regression model had a similar, though less severe, problem; 16 cases (30%) had absolute *d* values greater than 4. MPD yields unstandardized effects, to be interpreted as a change in outcome using the original outcome metric (e.g., "a decrease of 10 points on the Beck Depression Inventory"). Manolov and Solanas (2013) suggested it could be standardized for quantitative synthesis of single-case research, though results of this study suggest it may be more interpretable in its original unstandardized form. Out of 53 cases, the absolute standardized *d* value of MPD exceeded 4 on 20 cases (38%). The other three statistics showed less evidence of ceiling/floor effects.

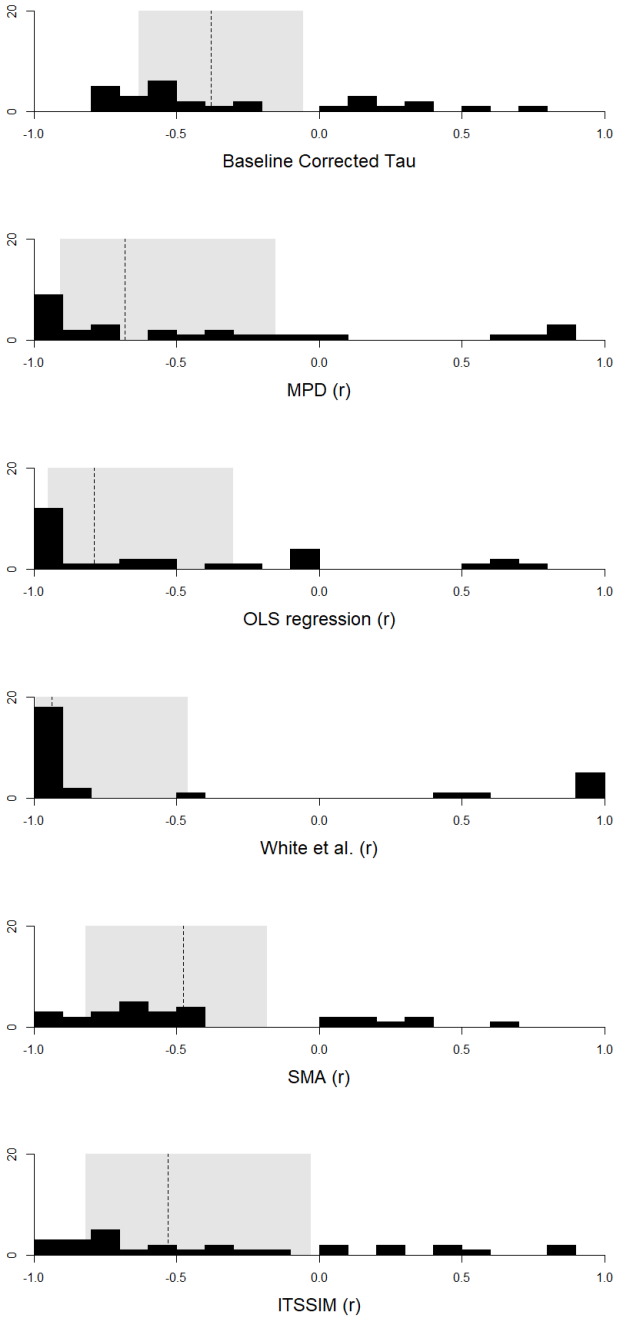



Figure 8. Histograms of effect size distributions, excluding treatments for bipolar disorder ($n$ = 28). Dotted lines indicate multilevel meta-analytic mean effect size; shaded regions indicate 95% confidence intervals.

Mean autocorrelation estimates are summarized in Table 9. Case-specific estimates of lag-1 autocorrelation ranged from $-0.238 \leq r_1 \leq 0.782$. As discussed above, the autocorrelation estimator is quite imprecise in brief time series. When $r_1$ values were combined within studies, they had a more moderate range of $-0.046 \leq r_1 \leq 0.502$. The overall random effects mean $\bar{r}_1 = 0.313$, $p < 0.001$ ($n = 53$). This result is similar to Shadish and Sullivan's (2011) meta-analysis of lag-1 autocorrelation in single-case studies, where they found $\bar{r}_1 = 0.20$, $p < 0.001$ ($n = 800$). However, unlike Shadish and Sullivan's survey, the meta-analysis of the cognitive therapy studies did not yield a statistically significant variance component of the random effects mean ($\tau^2 = 0.001$, Q = 54.68, df = 52, $p = 0.373$).

Table 9

*Mean Autocorrelation Estimates by Study*

| **Study** | ***n*** | $r_1$ | **95% CI** | ***p*** |
|---|---|---|---|---|
| **A** (Holmes) | 14 | 0.330 | [0.263, 0.394] | < 0.001 |
| **B** (Searson) | 7 | 0.269 | [0.067, 0.450] | 0.010 |
| **C** (Holländare) | 4 | -0.049 | [-0.471, 0.391] | 0.834 |
| **D** (Jones) | 4 | 0.120 | [-0.166, 0.388] | 0.412 |
| **E** (Mehranfar) | 4 | 0.006 | [-0.408, 0.417] | 0.980 |
| **F** (Akbari) | 3 | 0.502 | [0.221, 0.705] | 0.001 |
| **G** (Cowles) | 1 | 0.062 | [-0.628, 0.697] | 0.879 |
| **H** (McManus) | 6 | 0.224 | [-0.002, 0.427] | 0.052 |
| **I** (Maroti) | 2 | -0.046 | [-0.545, 0.478] | 0.874 |
| **J** (Beck) | 8 | 0.469 | [0.312, 0.601] | < 0.001 |
| All Studies | 53 | 0.313 | [0.259, 0.366] | < 0.001 |

## Discussion

The purposes of this study were (a) to demonstrate how to perform a meta-analysis of single-case studies of cognitive therapies for depression, and (b) to compare effect size statistics for single-case meta-analysis. Single-case experimental methods have the potential to contribute clinically relevant knowledge to psychotherapy research, but in the past these studies have been difficult to aggregate and synthesize. It is also unclear what statistics would be most appropriate for such a meta-analysis, and how

results might be interpreted alongside group design effect sizes (the "apples and oranges" problem).

**Effect Size Comparison**

The longstanding conventional wisdom of single-case research states "the more promising available statistical analysis methods [yield] moderately different results on the same data series … each available method is equipped to address only a relatively narrow spectrum of data" (Smith, 2012, p. 521). The results of this study suggest otherwise. Statistical methods on average gave consistent results when they used similar approaches to data modeling and were applied to a homogenous set of research studies (e.g., psychotherapy research). Of the six effect size statistics included in this study, two distinct groups were identified, distinguished by their approach to baseline trend correction. In the first group, Baseline Corrected Tau took a conservative approach to baseline trend correction (no correction was performed in 85% of the cases that did not have statistically significant baseline trend), and Simulation Modeling Analysis (SMA) did not correct for trend at all. The statistics in the second group, Mean Phase Difference (MPD), an OLS regression model, White et al.'s *d*, and Interrupted Time-Series Simulation (ITSSIM), took a more aggressive approach, correcting for baseline trend in all cases. Within these two groups of statistics, effect size estimates were strongly correlated, $0.83 \leq r \leq 0.93$.

Converting all statistics to a correlation metric resolved some interpretation issues. All effect size correlations were bounded between -1 and +1, and thus were easier to compare across methods (the $R^2$ effect size has the same advantage; however, it fails

to indicate the direction of treatment effect, positive or negative). Standardized mean difference statistics gave extreme values when applied to cases included in this study, consistent with other findings (Parker et al., 2005). Correlation effect sizes can be interpreted as the association between treatment and outcome (controlling for trend, etc.), though making sense of standardized mean differences for interrupted time-series data, especially extreme values, is more difficult. For example, interpreting an effect size of $r$ = -0.97 is less nettlesome than interpreting $d$ = -10.58, though both are equivalent (and were observed in one of the analyzed cases). Three of the statistics, White et al.'s $d$, the OLS regression model, and MPD, yielded $d$ values greater than 4 on over one-third of the analyzed cases. Ceiling effects were very pronounced on White et al.'s method, where $d > 10$ for over half of the analyzed cases. These findings support Parker et al.'s conclusion that "some effect sizes, notably [White et al.'s $d$], are relatively meaningless unless constrained by information on their reliability" (p. 128). In addition to supplementing effect sizes with reliability indicators (like confidence intervals), it is also suggested that other standardized effect sizes, like correlation coefficients, can aid with interpretation.

Baseline Corrected Tau and ITSSIM gave the most moderate effect size estimates on average (and their estimates were strongly correlated, $r = 0.72$, $p < 0.001$). Baseline Corrected Tau was the only statistic in this study that did not produce any extreme or difficult to interpret effect size estimates. Given concern about the "apples and oranges" problem of comparing single-case and group-design effect sizes, it may be useful to explore and develop these methods further. SMA effect size estimates were

also reasonably distributed and not overly extreme. However, despite this and other strengths, SMA does not offer the option for baseline trend correction; it should therefore be used with caution when analyzing data that may contain trend.

**Meta-Analysis with Single-Case Effect Size Statistics**

Single-case investigators have pursued two approaches to quantitative synthesis of single-case research. Parker and Vannest (2012) contrasted these strategies as "top-down" versus "bottom-up" approaches:

> Although "top-down" models, for example, multi-level or hierarchical linear models, are gaining momentum and have much to offer, interventionists should be cautious about analyses that are not easily understood, are not governed by a "wide lens" visual analysis, do not yield intuitive results, and remove the analysis process from the interventionist, who alone has intimate understanding of the design logic and resulting data patterns. "Bottom-up" analysis possesses benefits which fit will with [single-case research], including applicability to designs with few data points and few phases, customization of analyses based on design and data idiosyncrasies, conformation with visual analysis, and directly meaningful effect sizes. (p. 254)

This study took "middle ground" approach to meta-analysis. Single-case effect size statistics were used to estimate a treatment effect for all 53 cases, similar to a bottom-up strategy. However, like a top-down approach, a hierarchical model was used to combine these effects, because it is assumed that cases within studies are statistically dependent (the same is true for multiple sequential treatment effects within cases). As Glass et al. (1972) pointed out, violating the independence assumption of fixed effect statistical tests is "far more serious" than violating other assumptions. Here multilevel methods can help by accurately modeling within-study (and within-case) covariances. Rather than take a fully-multilevel approach where raw data are fitted to a hierarchical model (e.g., Van den Noortgate & Onghena, 2003, 2007, 2008), the single-case indices of within-case

effect size were inputted into a multilevel analysis. An advantage of this approach (besides correctly modeling within-study dependence) is that omnibus effect size estimates from multilevel meta-analysis were in a metric familiar to the single-case interventionist (e.g., "an average treatment effect of ITSSIM $r$ = -0.279"). A limitations of this approach is the relatively large number of cases needed to have adequate statistical power, especially when testing for moderator effects.

**Autocorrelation**

Serial dependency is difficult to estimate reliably in brief time-series. Investigators must also account for small-sample bias in autocorrelation estimation when working with single-case data (Huitema & McKean, 1991; Ferron, 2002). Lag-1 autocorrelation estimates from all 53 analyzed cases were aggregated via random-effects meta-analysis, similar to strategy used by Shadish and Sullivan (2011). There was a mean autocorrelation of $\bar{r}_1 = 0.313$, $p < 0.001$. This replicates Shadish and Sullivan's finding of small-but-significant positive autocorrelation in single-case data, and confirms the growing consensus that serial dependency should not be ignored in single-case statistical analysis. Unlike Shadish and Sullivan's results, the mean autocorrelation estimate in this study was not significantly heterogeneous. It is hypothesized that this result is due to the relative homogeneity of research studies (all were psychotherapy studies, etc.) compared to Shadish and Sullivan's very large survey of single-case research, which included over 800 cases from over 100 studies, spanning several research domains.

**Evidence of Treatment Efficacy**

There is a rich history of meta-analyses with group studies of psychotherapy outcomes. One need only consult the seminal *Benefits of Psychotherapy* by Smith et al. (1980; see also Smith & Glass, 1977), the first modern meta-analysis, to glimpse the breadth and depth of quantitative synthesis in therapy research. However, a literature review found no meta-analyses of single-case psychotherapy research—a disappointing result given the unique contributions single-case designs can offer the field.

Studies included in the meta-analysis incorporated innovative treatments, complex symptom presentations, and low incidence clinical populations. All studies tested the outcome of some cognitive or cognitive-behavioral intervention. Effect size estimates for individual cases were highly variable, and to a degree that variability limited the precision of within- and between-study effect size means. This supports earlier findings that single-case studies, especially those in clinical settings, tend to produce effects that are more variable than group designs (Jenson et al., 2007; Parker et al., 2005).

Omnibus mean effects for all 10 studies were large and in the expected direction (treatment was associated with a decrease in depressive symptoms/improved mood). However, due to the variability of individual cases, some of the mean effects were not statistically significant. Excluding White et al.'s effect size metric (for the distributional reasons outlined above), the five remaining methods produced a range of omnibus effect size estimates for the 10 analyzed studies ($n = 53$), $0.28 \leq \bar{r} \leq 0.60$. Converted to standardized mean difference ($d$) effect sizes, this is equivalent to $0.58 \leq ES \leq 1.48$.

Despite the "apples and oranges" concern about the typically large values of single-case effect sizes relative to group-design effects, this result is remarkably consistent with recent meta-analyses of cognitive-behavioral therapy RCT research, which yield effect sizes in the $0.71 \leq ES \leq 1.86$ range (Butler et al., 2006). Smith et al.'s (1980) original meta-analysis of psychotherapy outcomes found cognitive and cognitive-behavioral treatment outcomes in the $0.51 \leq ES \leq 1.82$.

Evidence of single-case treatment efficacy was clearer when the bipolar diagnosis moderator was included. Mean treatment effects for the bipolar cases ($n = 25$) were mixed, $-0.39 \leq ES \leq 0.64$ ($-0.19 \leq \bar{r} \leq 0.305$), indicating that, on average, participants may not have reliably improved during treatment, and may have even deteriorated. None of the mean bipolar effect sizes were statistically significant. On the other hand, mean effect sizes for the non-bipolar cases ($n = 28$) were large and statistically significant, $0.82 \leq ES \leq 2.57$ ($0.38 \leq \bar{r} \leq 0.79$).

Overall, these results are consistent with much of the RCT psychotherapy research. Single-case studies of cognitive therapies were, on average, associated with substantial decreases in depressive symptoms. Similar to previous outcomes research, individuals with bipolar disorder did not appear to benefit from psychotherapy interventions as much or as reliably as individuals not diagnosed with bipolar disorder.

## Conclusion

The results of this study indicate single-case experimental designs are a viable and valuable tool for psychotherapy research. Meta-analyses that blend multilevel modeling with single-case effect size estimation produce results that are statistically

reasonable and practically useful. Mean single-case effect sizes were quite similar in magnitude to treatment effects in RCT research, suggesting that, with further development, it may eventually be possible to synthesize research from different designs. While omnibus mean effects were not much larger than those from group-design meta-analyses, the effect sizes of individual cases were highly variable. Fortunately, single-case research allows investigators to examine these case-by-case differences within the context of individuals' demographics, diagnoses, and change over time. Two statistical methods used in this study, Baseline Corrected Tau and ITSSIM, performed well compared to the other four methods. Both gave a reasonable distribution of correlation effect sizes without the pronounced ceiling effects evident in other methods, and both can account for baseline trend when estimating treatment effects. One method from White et al. (1989) is not recommended due to the extreme, difficult-to-interpret results it produced.

# CHAPTER IV

# EVALUATION OF SINGLE-CASE STATISTICAL METHODS WITH COMPUTER-INTENSIVE SIMULATION: SOFTWARE AND APPLICATIONS


## Abstract

Computer-intensive simulation methods are often used to test and compare effect size statistics for singe-case experimental designs. Because there are many established single-case statistics, and because no one statistical method has demonstrated clear superiority, Monte Carlo simulation studies can clarify the conditions under which different methods may be appropriate. Simulations can also test statistical assumptions and define distributional properties for single-case effect size indices that are clinically useful but not well understood from a research perspective. This paper introduces a free software application for conducting simulation research with single-case statistics. The application, Interrupted Time-Series Lab (ITSLAB), allows users to manipulate parameters (including phase length, baseline trend, level- and slope-change effects, autocorrelation, and within-phase error variance) while exploring the distributions and statistical power of several popular single-case effect size measures.


## Introduction

Single-case experimental designs, which have a rich history in behavioral science, are an increasingly popular tool of applied researchers in many areas of education and psychology (Smith, 2012). Single-case methods permit investigators to

establish causal evidence of treatment efficacy (APA Presidential Task Force on Evidence-Based Practice, 2006; Barlow & Hersen, 1984), often with a fraction of the resources required of large-*n* group designs (Barlow & Nock, 2009). Brief single-participant interrupted time-series experiments can also bridge the gap between scientists and practitioners because they can be implemented by practitioners in clinical settings (Borckardt et al., 2008). Single-case experiments may be of particular use to practitioners who wish to demonstrate their treatment efficacy, given the growing emphasis on evidence-based treatments in education and healthcare (Morgan & Morgan, 2001; Shadish et al., 2008).

An obstacle to wider adoption of single-case experiments is the lack of consensus regarding data analytic methods (Smith, 2012). Brief interrupted time-series data violate the independence assumptions of most statistics used in large-*n* between-groups research (Borckardt, 2008; Wampold, 1988). In addition, single-case experiments in applied psychology and education are often too brief to utilize the time-series analysis methods used in other fields (Box & Jenkins, 1970; Glass et al., 1975). As a result, single-case researchers have proposed a range of tools for statistically measuring and combining single-case experimental treatment effects (Parker et al., 2011; Shadish, 2014). No single method has emerged as clearly superior. As a result, single-case investigators must assess which statistics are best suited for their experiment or use non-statistical evaluation methods like visual analysis, which are often unreliable (Brossart, Parker, Olson, & Mahadevan, 2006; Danov & Symons, 2008; DeProspero & Cohen, 1979; Harbst, Ottenbacher, & Harris, 1991; Park, Marascuilo, & Gaylord-Ross, 1990;

Lieberman et al., 2010; Ximenes, Manolov, Solanas, & Quera, 2009) and poorly suited for research synthesis (Beretvas & Chung, 2011; Busk & Serlin, 1992; Maggin & Odom, 2014; Scruggs & Mastropieri, 2001).

Researchers have relied on several strategies to determine how single-case statistics should be used and interpreted given the limitations and assumptions of the various effect size indices. One approach involves the evaluation of one or more effect size statistics by applying them to a sample of published data sets (Brossart et al., 2011; Campbell, 2004; Ma, 2006; Parker et al., 2005, 2006, 2011; Parker & Hagan-Burke, 2007; Parker & Vannest, 2009; Parker, Vannest, & Davis, 2011; Shadish et al., 2014; Tarlow, in press; Vannest et al., 2012; Wolery et al., 2010). This strategy allows investigators to understand how methods perform in "real life" applications. Statistical results from a range of published studies also help identify clinically useful cutoff points and potential limitations of the effect size indices (e.g., Scruggs & Mastropieri, 1998).

A second strategy involves large surveys of published single-case experiments, similar to the approach described above. However, rather than assessing these surveyed data sets with single-case effect size metrics, various statistical properties of the data are examined (Huitema, 1985; Matyas & Greenwood, 1991; Shadish & Sullivan, 2011; Smith, 2012; Solomon, 2013). Systematic reviews of design type, phase length, trend, autocorrelation, data distribution, and other parameters help investigators contextualize their own research and identify promising statistical metrics for further development.

A third approach to the study of single-case statistical methods involves the use of computer-intensive simulation methods (Allison & Gorman, 1994; Crawford et al.,

2006; Gorsuch, 1983; Manolov & Solanas, 2008, 2009, 2012, 2013; Manolov et al., 2011; Matyas & Greenwood, 1990; Smith et al., 2012; Solanas, Manolov, & Onghena, 2010; Tarlow, in press; Ugille et al., 2012). In these studies, artificial time-series are simulated with fixed parameters and analyzed with different effect size statistics. Parameters are then systematically manipulated to determine how changes in trend, autocorrelation, phase length, distribution type, and other characteristics influence effect size. Because the true parametric properties of "real-life" data are unknown, valuable insights can be gained by exploring how outcome metrics perform across a range of conditions. Computer-intensive simulation studies can "stress-test" (Solomon, 2013) effect size indices by examining how violations of statistical assumptions affect results (for example, how effect size measures that do not model slopes are affected by the presence of baseline trend). Simulations also help establish the distributional properties and statistical power of effect size measures which lack formal theoretical development.

Manolov and Solanas (2012), who developed many of the simulation methods used in single-case simulation research, stated, "It is not likely that applied researchers will be easily able to perform the simulations described" (p. 497). Indeed, a limitation of simulation methods is their inaccessibility to many practitioners and researchers. Investigators who wish to perform computer-intensive simulation research would typically need to be fluent in a programming language suitable for high-powered scientific computing, such as FORTRAN or C. The lack of software applications for conducting single-case simulation research excludes many investigators—and perhaps just as importantly, peer-reviewers—from performing and verifying these analyses. As

valuable as computer-intensive simulation tools are for single-case research, it remains a "black box" method for all but a handful of researchers.

This paper introduces a new software application with the goal of making single-case computer-intensive simulation methods accessible to all investigators. *Interrupted Time-Series Lab* (ITSLAB; Tarlow, 2017b) is a user-friendly standalone application which allows the user to experiment with several popular single-case statistics by manipulating a range of data parameters, including phase length, level- and slope-change effects, baseline trend, autocorrelation, and phase variability. No previous experience with statistical computing or syntax is required. To demonstrate the utility of ITSLAB, this paper will present several simulation studies that are easily performed with the application.

## ITSLAB Software

### How to Use ITSLAB

The ITSLAB software is available for download at the author's website, http://ktarlow.com/stats/itslab (Tarlow, 2017b). Once downloaded, the user runs the executable file, which will open a text-based console. The software will prompt the user to select an effect size statistic for simulation study. Then, it will prompt the user to enter the simulation parameters: A phase length, B phase length, baseline trend ($\beta_1$), level-change ($\beta_2$), slope-change ($\beta_3$), A phase standard deviation, B phase standard deviation, autoregressive error term ($\varphi_1$), and moving averages error term ($\theta_1$). The software will then generate 10,000 artificial time-series using the simulation parameters, then analyze the simulated time-series with the selected effect size statistic. ITSLAB then outputs the

mean effect size value, standard error, statistical power (percent of simulated time-series with statistically significant effects), and effect size quartile values.

ITSLAB (Version 1.0) includes six single-case effect size statistics with the plan to add additional indices to future versions. The six statistics included in the initial release of ITSLAB are: Tau-U (Parker et al., 2011; Parker, Vannest, & Davis, 2011; Vannest et al., 2016), Baseline Corrected Tau (Tarlow, in press), Mean Phase Difference (MPD; Manolov & Solanas, 2013), Extended Celeration Line (ECL; White & Haring, 1980), Percentage of Nonoverlapping Data (PND; Scruggs, Matropieri, & Casto, 1987; Tarlow & Penland, 2016), and ITSSIM (see Chapter II). These effect size statistics were selected because they are popular in the single-case research literature (e.g., PND; Maggin et al., 2011) or have otherwise demonstrated promise for the measurement and meta-analysis of single-case data.

**Effect Size Indices**

**Tau-U.** Parker et al. (2011) and Parker, Vannest, and Davis (2011) proposed Tau-U for single-case research, a variant of Kendall's (1962) rank correlation coefficient. There are several variants of Tau-U (Parker, Vannest, Davis, & Sauber, 2011); however, the coefficient used here is the one that purportedly combines a measure of phase nonoverlap with baseline trend control (i.e., Tau-$U_{A\text{ vs. }B - \text{trend }A}$), which can be calculated with an online calculator (Vannest, Parker, & Gonen, 2011). Tau-U is calculated as a Tau rank correlation between time-series observations and a specially coded phase variable. In a standard rank test of homogeneity, an interrupted time-series with $n_A = n_B = 5$ would be correlated with the dummy code variable [0, 0, 0 ,0 ,0 / 1, 1,

1, 1 1]. The Tau-U phase variable [5, 4, 3, 2, 1 / 6, 6, 6, 6, 6] would be substituted, reversing the order of the time variable for the A phase and repeating the initial value of the B phase. Tau-U's authors suggested an arithmetic adjustment to Kendall's original Tau equation "because KRC [Kendall's rank correlation] is not designed for dummy-coded variables" (Parker, Vannest, & Davis, 2011, p. 313). However, Tarlow (in press) pointed out that Kendall in fact described several statistical tests that were essentially Tau analyses with dummy code variables. Tarlow also found the arithmetic adjustment, where the Kendall score (S) is divided by the product $n_A \times n_B$ instead of the standard denominator, can yield inflated and "out-of-bounds" results. Despite these limitations, Tau-U has been widely adopted in single-case research and meta-analysis.

**Baseline Corrected Tau.** Tarlow (in press) proposed Baseline Corrected Tau as an improved rank correlation coefficient effect size statistic, similar to Tau-U. A two-step process is followed to calculate Baseline Corrected Tau. First, the baseline phase is tested for statistically significant monotonic trend. Baseline trend is tested with a Tau correlation of the A phase observations and a time variable. If statistically significant trend is detected, both phases are adjusted with a Theil-Sen (Sen, 1968; Theil, 1950) nonparametric regression. A Theil-Sen regression line is fit to the A phase data by first finding the slopes of all possible pairs of A phase data points; the median slope value is the Theil-Sen slope estimate, *b*. The *y*-axis intercept of the Theil-Sen regression is the median value of all ($y - bx$) values for each A phase data point. The regression line from this procedure is projected into the B phase, and regression residuals are extracted from both phases as corrected data. The Theil-Sen regression residuals will tend to minimize

monotonic trend in the A phase, thus correcting for baseline trend. Residuals are then rank correlated with a phase dummy code variable to produce a Tau effect size. If no statistically significant baseline trend is detected, the correction process is skipped, and the original raw data are correlated with a phase dummy code variable for an uncorrected Tau effect size. Tarlow found that Baseline Corrected Tau outperformed Tau-U on real and simulated data sets. One limitation of Baseline Corrected Tau is the relatively low statistical power when testing for baseline trend; time series with fewer than seven baseline data points rarely yielded statistically significant baseline trend—and thus, the null hypothesis of no baseline trend was not rejected in the final Tau effect size analysis.

**Mean Phase Difference (MPD).** Manolov and Solanas's (2013) Mean Phase Difference (MPD) estimates an effect size by first removing baseline trend and then comparing the A and B phases. MPD estimates baseline trend by taking the average of the first-order differences of the A phase scores. This procedure results in $n_{A-1}$ difference scores, calculated as $n_{t+1} - n_t$. Similar first-order differencing methods have demonstrated usefulness in other single-case data applications (Manolov & Solanas, 2009; Solanas, Manolov, & Onghena, 2010). After estimating the baseline trend, *b*, via differencing, a new B phase data series, $\hat{y}$, is projected from the baseline phase using the equation $\hat{y}_{n_A+i} = y_1 + b(n_A + i - 1)$. The observed B phase data points ($y_i$) are then averaged and differenced from the average of the projected B phase data points ($\hat{y}_i$). The result, a difference of means, may be interpreted as the change between phases after controlling for baseline trend. It is important to note the MPD result is an unstandardized

effect in the original metric of the outcome variable. MPD may be standardized, similar to a *d*-type standardized mean difference. For this study, the standardized MPD *d* was then converted to an *r* correlation, which can be loosely interpreted as the percentage of nonoverlap between the projected and observed B phase data (see Cohen, 1969, for discussion of *r* as an overlap statistic).

**Extended Celeration Line (ECL).** White and Haring (1980) proposed a "split middle" technique for analyzing interrupted single-case data. ECL was also studied as "PEM-T", the Percentage of Data Exceeding a Median Trend (Wolery et al., 2010). First, the baseline phase data points are dividing into two halves by time, which gives the points ($X_i$, $Y_i$) in the first half of the baseline and ($X_j$, $Y_j$) in the second half of the baseline. A trend line is then fit to the medians of each half, i.e., the points $(X_i, Y_i)$ and $(X_j, Y_j)$. Using this method, approximately 50% of the baseline points will fall below the line and 50% will fall above the line. This baseline trend line is then extended into the treatment phase, and the percentage or proportion of treatment phase points exceeding the line is reported as an effect size. Under the null hypothesis of no intervention effect, 50% of the treatment phase points are expected to be above the line and 50% below, as in the baseline phase. The effect size, originally conceived as a percentage, can be scaled into a -1 to +1 metric for easy interpretation and comparison to correlation indices (Parker, Vannest, & Davis, 2011). ECL's strengths as an effect size include its simple calculation and straightforward interpretation as the percentage of change accounted for by the treatment effect, controlling for baseline trend.

**Percentage of Nonovlerapping Data (PND).** PND (Scruggs, Mastropieri, & Casto, 1987) is perhaps the most widely used single-case effect size statistic (Beretvas & Chung, 2008; Maggin, O'Keefe, & Johnson, 2011; Parker et al., 2011; Schlosser, Lee, & Wendt, 2008). PND is popular because it is easy to calculate and interpret. PND is calculated by dividing the number of "nonoverlapping" B phase scores by the total number of scores in the B phase—thus, PND is a "percentage of nonoverlap." Nonoverlapping B phase scores are the data points that exceed the most extreme score in the A phase (extreme in the direction of treatment effect). PND has been criticized for not controlling baseline trend (Allison & Gorman, 1993; Ma, 2006; Salzberg et al., 1987; Wolery et al., 2010) and for yielding effect size estimates that are negatively correlated with baseline length (Allison & Gorman, 1994). Despite its limitations, it remains a popular metric for summarizing single-case research. "In fact," Schlosser et al. stated, "it is unlikely that the field has similar implementation experiences for any other metric at this point in time" (p.184).

**Data Generation**

ITSLAB generates artificial time-series with a modified regression model that is commonly used in single-case simulation research (Manolov & Solanas, 2008, 2009, 2012, 2013; Smith et al., 2012; Tarlow, in press). First, error terms are generated from random normally-distributed deviates and the user's inputted lag-1 autocorrelation parameter, either autoregressive ($\varphi_1$) or moving average ($\theta_1$):

$$\text{AR}(1): \varepsilon_t = (\varphi_1)(\varepsilon_{t-1}) + \mu_t$$

$$\text{MA}(1): \varepsilon_t = (\theta_1)(\mu_{t-1})$$

where $u_t$ is a random normally distributed term with variance of $\sigma^2$. Error terms are then multiplied by the inputted phase-specific standard deviations. The time-series is then generated as

$$Y_t = \beta_1 T_t + \beta_2 D_T + \beta_3 [T_t - (n_A + 1)] D_t + \varepsilon_t$$

where $\beta_1$ is the baseline trend, $\beta_2$ is level-change, $\beta_3$ is slope change, $T_t$ is the time variable, $D_t$ is a dummy code variable for phase, and $[T_t - (n_A + 1)]D_t$ is a slope-change term.

**Software Validation**

Each ITSLAB effect size module was first validated by entering individual single-case data sets and comparing effect size output to hand calculations or peer-reviewed effect size calculators (Tarlow, 2016; Tarlow & Penland, 2016; Vannest et al., 2016). After confirming that ITSLAB calculated all effect sizes accurately, the simulation procedure was validated by replicating analyses in published single-case research articles. Simulation models from Manolov and Solanas (2008) and Tarlow (in press) were inputted into ITSLAB and the output was compared to published tables and figures for agreement. ITSLAB results were also verified with simulation modeling performed in R statistical software (those results not published here). Uncompiled source code for ITSLAB (written in C) is also available from the author for independent review.

**Questions for Study**

ITSLAB allows researchers to study a variety of parameter effects and interactions on the six included effect size indices. For example, researchers have shown a particular interest in the influence of baseline trend and autocorrelation on single-case

statistics (Gorsuch, 1983; Manolov & Solanas, 2008, 2009, 2013; Manolov et al., 2011; Matyas & Greenwood, 1990; Tarlow, in press). For this paper, the following questions are explored:

**Question 1.** How are effect sizes influenced by change in level between phases (i.e., a treatment effect)?

**Question 2.** How are effect sizes influenced by baseline trend?

**Question 3.** How are effect sizes influenced by baseline phase length?

**Question 4.** How are effect sizes influenced by the presence of lag-1 autocorrelation, both autoregressive ($\varphi_1$) and moving average ($\theta_1$) models?

**Question 5.** How are effect sizes influenced by unequal variance between phases?

## Results

**Question 1: Level-change effect.** The simplest type of treatment effect is a change in level, or magnitude, of the outcome variable. In the absence of trend, autocorrelation, or other time-dependent processes, a stable baseline should demonstrate a vertical shift in the direction of treatment effect after an effective intervention is introduced.

Question 1 was explored by setting the following ITSLAB parameters: $n_A = 8$, $n_B = 8$, $\beta_1$ (baseline trend) = 0, $\beta_3$ (slope-change) = 0, $s_A = 1$, $s_B = 1$, $\varphi_1 = 0$, and $\theta_1 = 0$. These parameters gave the simplified regression model, $Y_t = \beta_2 D_t + \varepsilon_t$. A simulation was then performed with varying levels of $\beta_2$, the level-change term. $\beta_2$ was increased from 0

(no effect) to 4 (four standard deviations) by increments of 0.2. Figure 9 illustrates the influence of level-change on effect size.

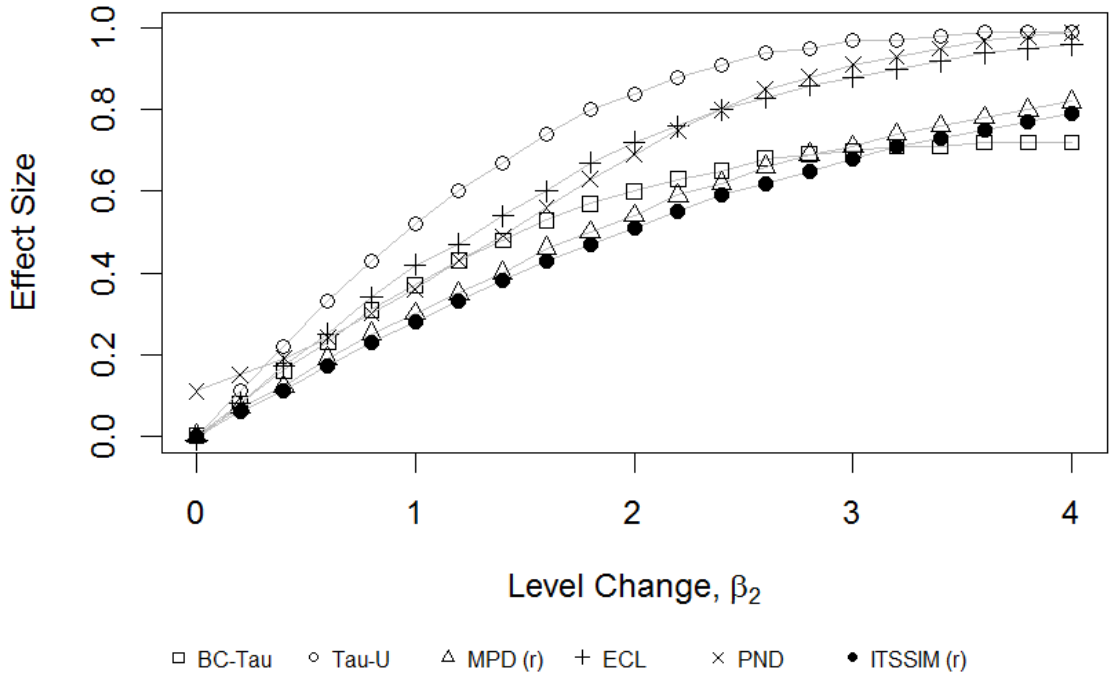



Figure 9. Level-change and effect size. No trend or autocorrelation; $n_A = n_B = 8$.

**Question 2: Baseline trend.** Treatment effects may be difficult to detect without a stable baseline. Consider the methodological challenge posed by the recovering patient: if the patient was already improving prior to treatment, can a recovery be

attributed to the treaetment or to the baseline trend already underway? Five of the six effect size statistics included in ITSLAB model baseline trend, though their trend models differ. How effectively they control for trend may be explored via simulation where the degree of trend is systematically manipulated.

Question 2 was explored with the following ITSLAB parameters: $n_A = 8$, $n_B = 8$, $\beta_2 = 0$, $\beta_3 = 0$, $s_A = 1$, $s_B = 1$, $\varphi_1 = 0$, and $\theta_1 = 0$. These parameters gave the simplified regression model, $Y_t = \beta_1 T_t + \varepsilon_t$. The baseline trend coefficient, $\beta_1$, was increased from 0 to 2 by increments of 0.2. Figure 10 illustrates the influence of baseline trend on effect size.

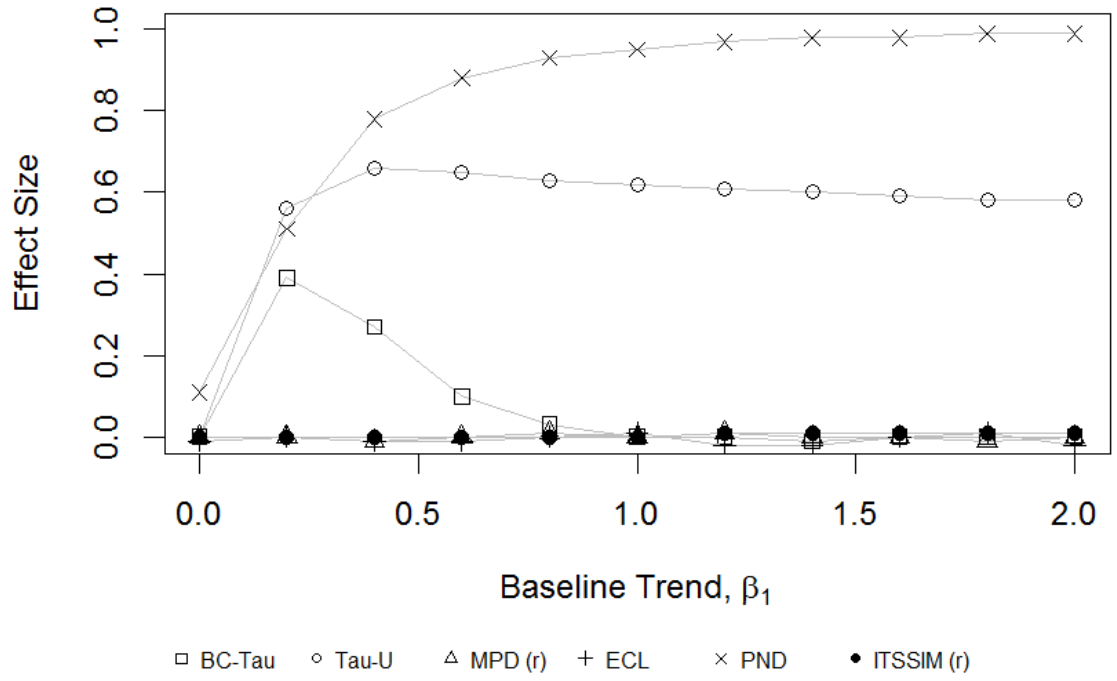



Figure 10. Baseline trend and effect size. No treatment effect or autocorrelation; $n_A = n_B = 8$.

**Question 3: Baseline phase length.** In interrupted time-series designs, baseline phase observations are used to infer treatment effects implemented in the treatment phase. Kazdin (1982) stated

> The [baseline] data serve as the basis for predicting the level of performance for the immediate future if the intervention is not provided … Presumably, if treatment is effective, performance will differ from the projected level of baseline. (pp. 105-106)

Following this logic, many single-case statistics use A phase data to model a "null effect" which is compared to the B phase observations. One's model of the null effect

tends to be more precise as the number of A phase observations increases, and picture of baseline functioning becomes clearer. Investigators are urged to maximize the number of baseline observations to increase the statistical power of their effect size tests (e.g., Tarlow & Penland). It is helpful to understand how baseline phase length effects both the magnitude and precision of single-case effect size indices.

Question 3 was investigated with the following ITSLAB parameters: $\beta_1 = 0$, $\beta_2 = 1$, $\beta_3 = 0$, $s_A = 1$, $s_B = 1$, $\varphi_1 = 0$, and $\theta_1 = 0$. These parameters gave the simplified regression model, $Y_t = D_t + \varepsilon_t$. This model indicates an interrupted time-series with no baseline trend, slope-change, or autocorrelation effects, and a level-change of one standard deviation from A phase to B phase. The B phase length was fixed at $n_B = 8$, and the A phase length varied from $n_A = 3$ to 20. Figure 11 illustrates the effect of phase length on effect size, and Figure 12 illustrates the effect of phase length on the effect size standard error.

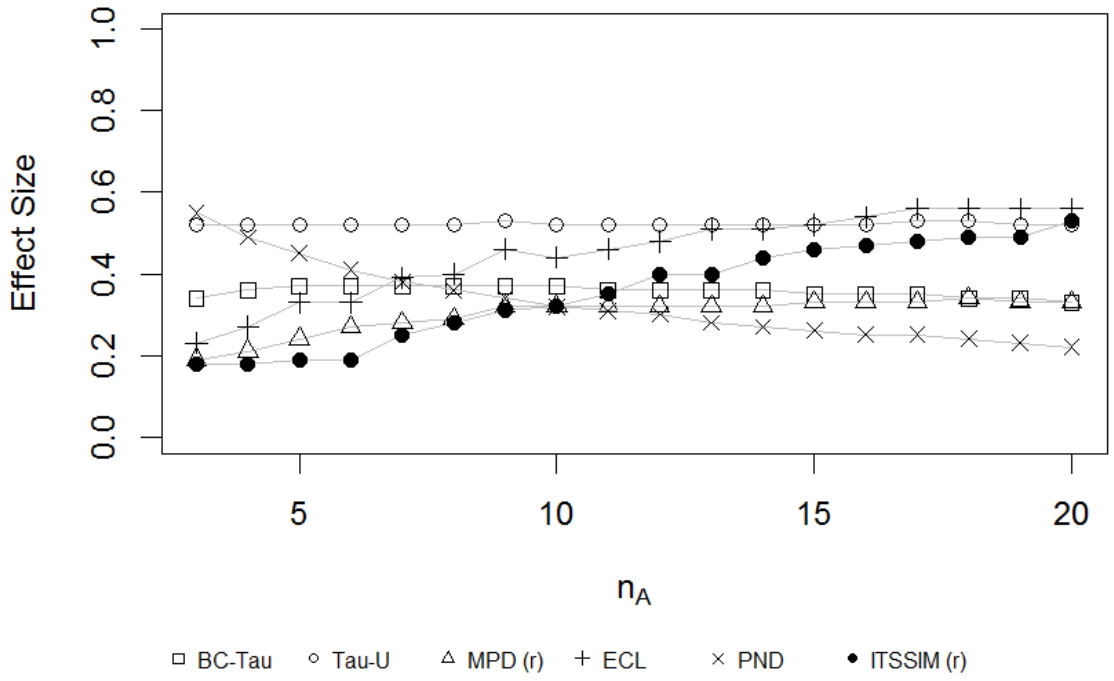



Figure 11. Baseline phase length and effect size. Level-change treatment effect ($\beta_2 = 1$); $n_B = 8$.

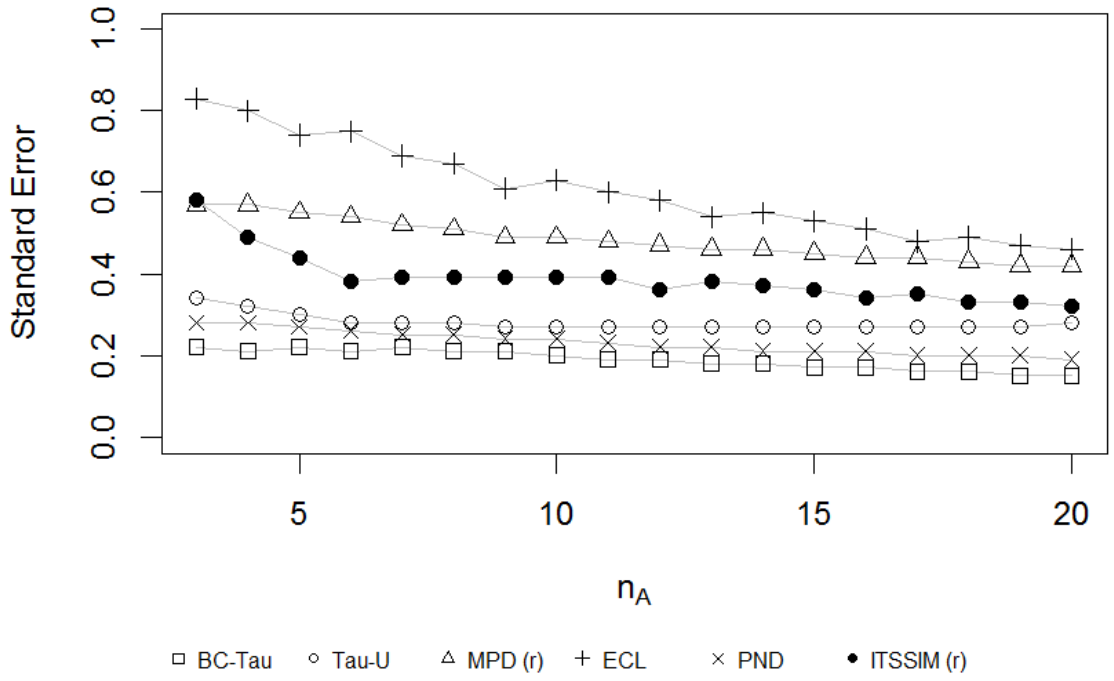



Figure 12. Baseline phase length and standard error of effect size. Level-change treatment effect ($\beta_2 = 1$); $n_B = 8$.

**Question 4: Autocorrelation.** Most single-case effect size statistics do not model serial dependency. Simulation research has been a useful tool for understanding how those indices are influence by different levels and models of autocorrelation. Statistics that are less influenced by autocorrelation are preferable when data are expected to be serially dependent.

To investigate Questions 5 and 6, the following ITSLAB parameters were used: $n_A = 8$, $n_B = 8$, $\beta_1 = 0$, $\beta_2 = 1$, $\beta_3 = 0$, $s_A = 1$, $s_B = 1$. These parameters gave the simplified regression model, $Y_t = D_t + \varepsilon_t$. This model indicates a level-change effect of one standard

deviation. The autocorrelation parameters, $\varphi_1$ and $\theta_1$, varied from -0.9 to 0.9 by increments of 0.3. Figure 13 illustrates effect size values for the various levels of autoregressive error, and Figure 14 illustrates the effect size values for the various levels of moving average error.

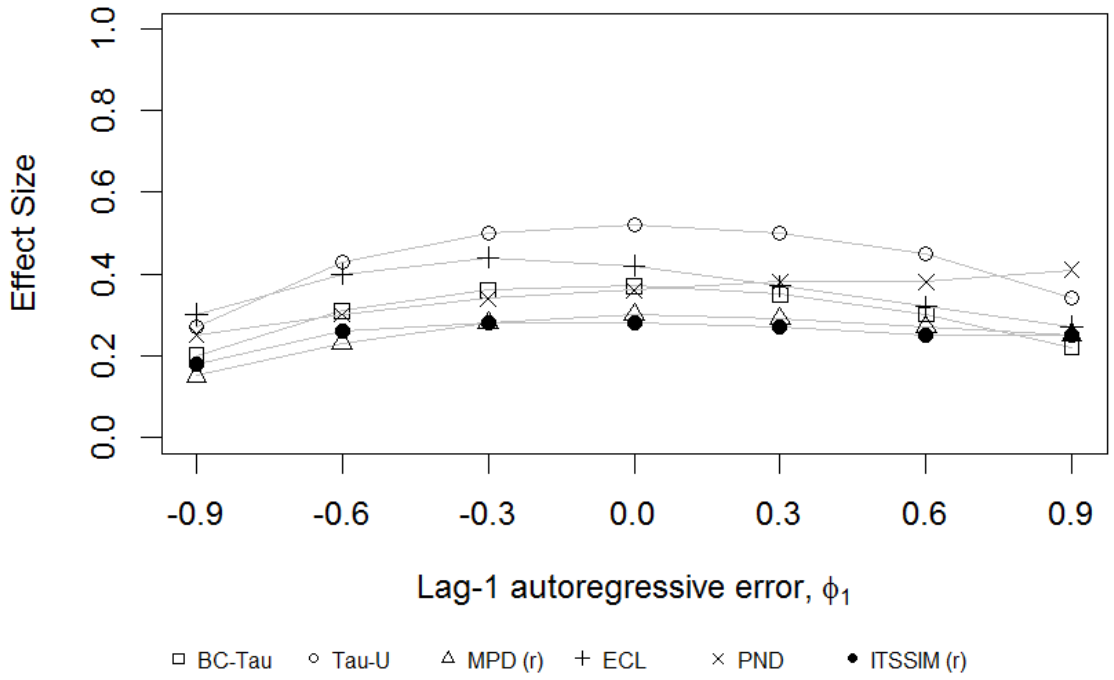



Figure 13. Lag-1 autoregressive error and effect size. Level-change treatment effect ($\beta_2$ = 1); $n_A = n_B = 8$.

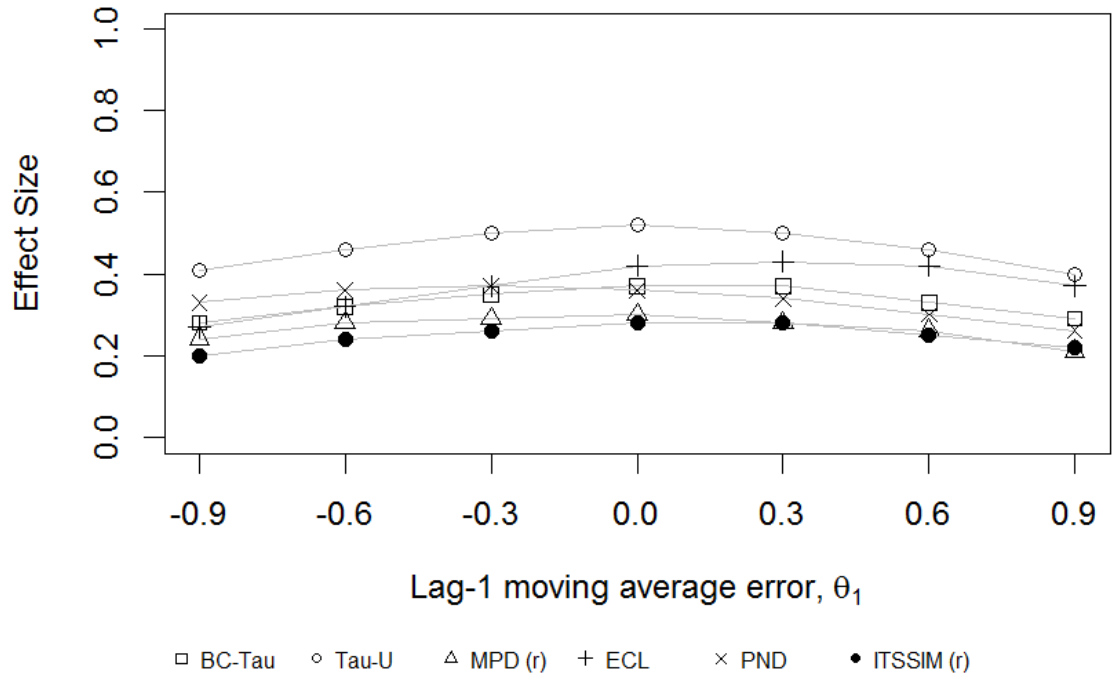



Figure 14. Lag-1 moving average error and effect size. Level-change treatment effect ($\beta_2$ = 1); $n_A = n_B = 8$.

Shadish et al. (2008) pointed out the common misperception that the independence assumption does not apply to nonparametric and nonoverlap effect size estimators (like the Tau-based statistics). In fact, these statistics do assume independence of data points—an assumption violated by autoregressive or moving average error structures. Most simulation studies of autocorrelation focus on the magnitude of the effect under varying levels of autocorrelation, but pay less attention to the distribution of the effect size statistic (e.g., Tarlow, in press). Shadish et al. noted that the *standard errors* of the statistics—and not the actual effect sizes—are more sensitive to serial

dependency. However, the standard errors of statistics are rarely explored with simulation research. (Indeed, the distributional properties are unknown for some nonparametric measures.) To explore this issue further, the standard errors of the six effect size indices were graphed for the varying levels of autocorrelation, illustrated in Figure 15 and Figure 16.

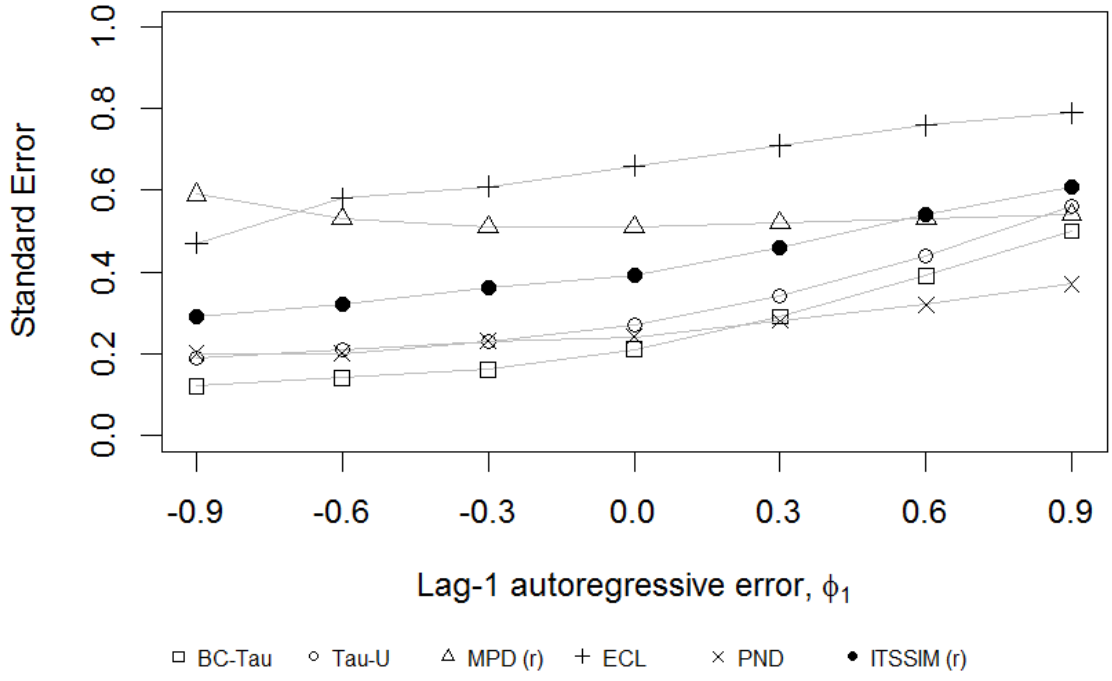



Figure 15. Lag-1 autoregressive error and standard error of effect size. Level-change treatment effect ($\beta_2 = 1$); $n_A = n_B = 8$.

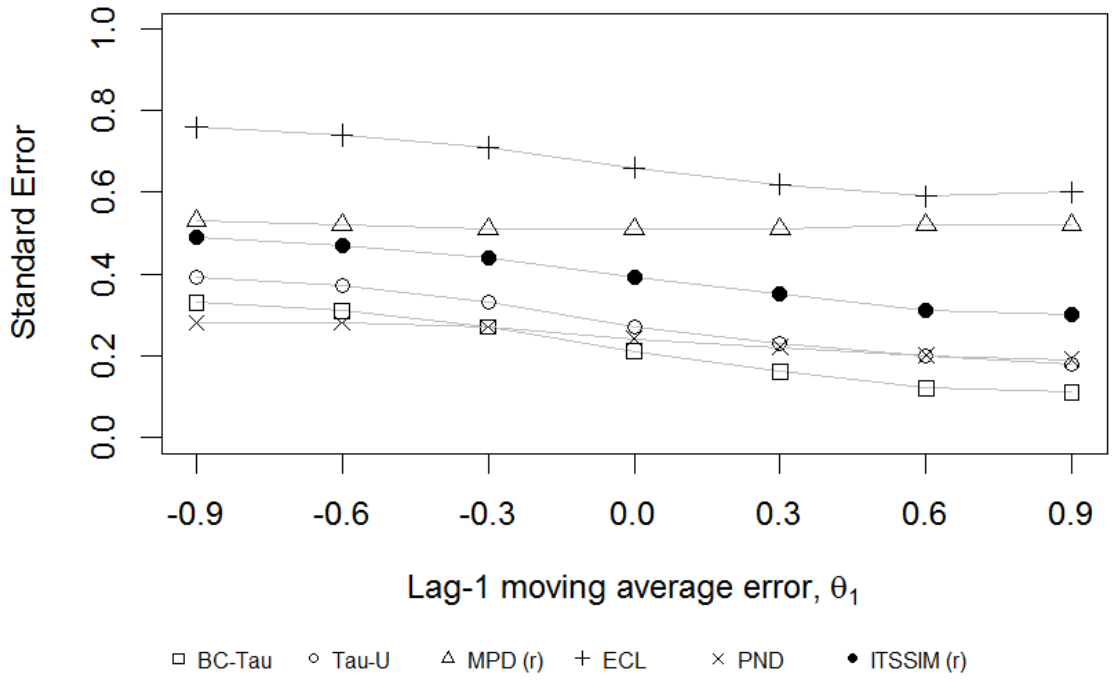



Figure 16. Lag-1 moving average error and standard error of effect size. Level-change treatment effect ($\beta_2 = 1$); $n_A = n_B = 8$.

**Question 5: Heteroscedasticity.** Relatively little attention has been paid to the effect of phase variance on effect size, despite recommendations about including considerations of data spread in single-case analyses (e.g., Kratochwill et al., 2010). The nonoverlap-based effect size indices in particular are sensitive to data variability; however, the influence of variance and unequal variance (heteroscedasticity) between phases is not well understood for most single-case statistics.

Question 5 was explored with these ITSLAB parameters: $n_A = 8$, $n_B = 8$, $\beta_1 = 0$, $\beta_2 = 1$, $\beta_3$ (slope-change) = 0, $\varphi_1 = 0$, and $\theta_1 = 0$, or the regression model $Y_t = D_t + \varepsilon_t$,

once again indicating a time-series model with a level-increase of one (baseline phase) standard deviation. First, the A phase standard deviation was varied from $s_A = 0$ to 2, while keeping the B phase standard deviation set at $s_B = 1$. Then the opposite design was implemented, where A phase standard deviation was kept constant and B phase standard deviation varied. The results of these simulations are presented in Figure 17 and Figure 18.

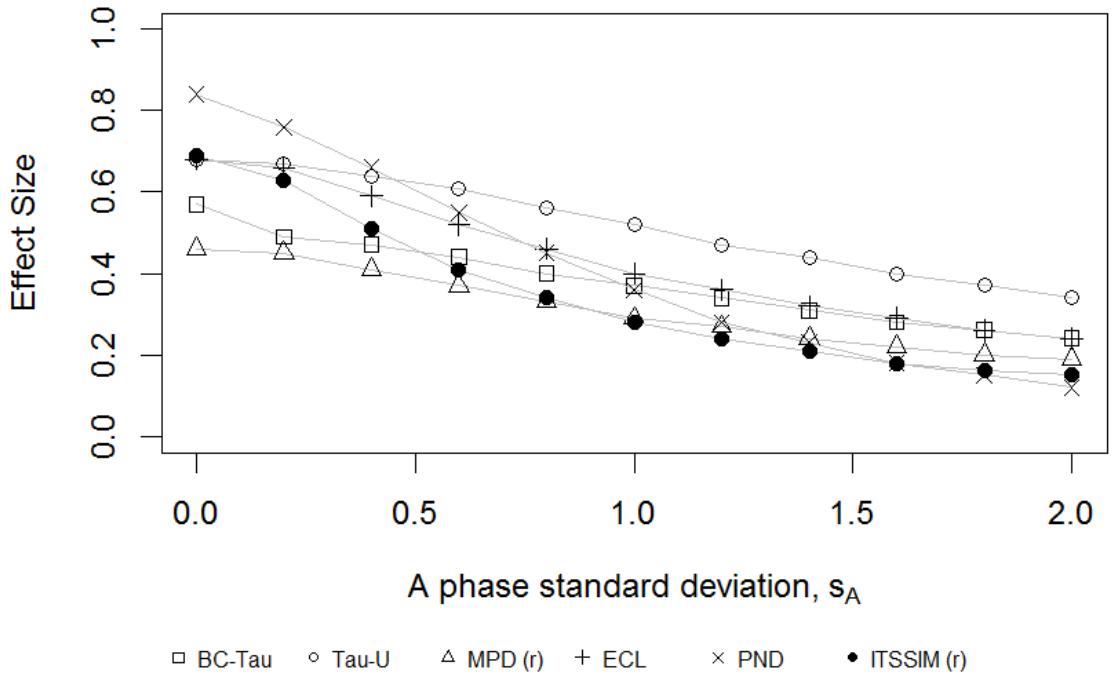



Figure 17. Unequal A phase variance and effect size. Level-change treatment effect ($\beta_2 = 1$); $n_A = n_B = 8$; $s_B = 1$.

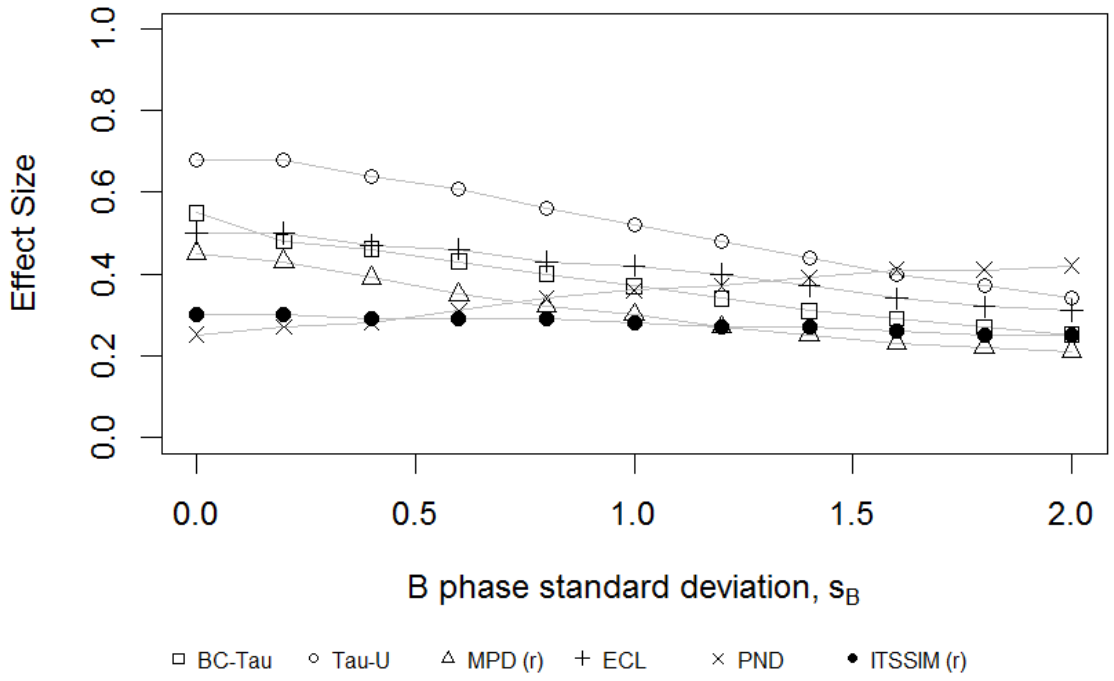



Figure 18. Unequal B phase variance and effect size. Level-change treatment effect ($\beta_2$ = 1); $n_A = n_B = 8$; $s_A = 1$.

## Discussion

Computer-intensive simulation methods are often used to test and compare single-case effect size statistics. Simulation studies allow investigators to understand how effect size indices—some of which are not based in formal statistical theory—perform under different parametric assumptions. The so-called "stress testing" of single-case statistics under different data models can alert investigators to the limitations of their measures. Practitioners may use these insights when designing their studies and

statistical analyses. Researchers can use the results of simulation studies to develop better statistical methods for the measurement and synthesis of single-case data.

The aim of this study was to introduce a new, easy-to-use software application, Interrupted Time-Series Lab (ITSLAB), which makes high-powered simulation research accessible to a wider range of single-case investigators. To demonstrate the utility of ITSLAB, it was used to investigate five research questions regarding the influence of level-change, baseline trend, baseline phase length, autocorrelation, and unequal phase variance on several single-case effect size statistics. It was possible to determine how the effect size indices performed under different data assumptions by systematically manipulating ITSLAB simulation parameters.

The results revealed some unsurprising simulation results, such as the influence of level-change treatment effect on single-case statistics. Other results were unexpected, such as the effect of autocorrelation on effect size indices' precision (standard error). The primary goal of this paper is instructive—to demonstrate how simulation research questions could be investigated with ITSLAB—however, the results of these simulation studies will be briefly discussed.

**Question 1: Level-Change Treatment Effect**

As expected, all effect size indices increased with the magnitude of level-change treatment effect. Tau-U tended to give larger effect size estimates, whereas ITSSIM and MPD effect sizes were more conservative. However, differences between effect size estimates were not dramatic when only level-change effects where systematically manipulated.

**Question 2: Baseline Trend**

MPD, ECL, and ITSSIM controlled for baseline trend in all "no treatment effect" simulation models. Baseline Corrected Tau effectively controlled baseline trend when trend was sufficiently large; however, it was less effective controlling for small amounts of baseline trend. This replicates and extends Tarlow's (in press) simulation study of Baseline Corrected Tau, which found the statistic's ability to detect and correct for trend was product of baseline phase length and magnitude of trend. Baseline Corrected Tau may not be an appropriate statistic for single-case data analysis when baseline phases are very brief and there is the possibility of small, difficult-to-detect baseline trend.

PND and Tau-U failed to control for baseline trend in all simulation models. PND lack of trend modeling is a well-documented limitation of the statistic (Allison & Gorman, 1993; Salzberg et al., 1987; Scruggs et al., 1987; Wolery et al., 2010). As the simulation results (and decades of commentary) demonstrate, PND should not be used when baseline trend is present in single-case data. Manolov and Solanas (2009) proposed the Percentage of Nonoverlapping Corrected Data (PNCD) statistic, which combines PND with a stochastic trend correction procedure to account for baseline trend. It is expected that PNCD will be included in future versions of ITSLAB for additional investigation.

Tau-U's failure to control for trend is more concerning because it is described as a method that accounts for any configuration of baseline trend (Parker, Vannest, & Davis, 2011; Parker et al., 2011). Tarlow (in press) described the limitations of Tau-U, which included its poor trend control, lack of conventional bounds (Tau-U can exceed -1

or +1), and lack of visual graphing method. This study's simulation results confirm and extend those findings. Single-case investigators are strongly advised to consider other effect size statistics, in particular when baseline trend is present.

**Question 3: Baseline Phase Length**

In the simple level-change model under consideration, MPD, ECL, and ITSSIM yielded larger effect size estimates as the number of baseline phase observations increased. All three of those methods estimate an effect size by comparing the B phase observations to predictions based on the A phase, i.e., comparing the observed treatment effect to an unobserved, predicted "null effect". It makes sense that, as the number of baseline observations increases, the accuracy of the null effect predictions improves, and the effect size estimate grows more precise.

Baseline Corrected Tau and Tau-U effect size estimates were stable across varying baseline phase lengths, at least under the level-change model (no trend and no autocorrelation). On the other hand, PND decreased as the number of baseline phase observations increased. This replicates findings by Allison and Gorman (1993), who criticized PND because, among other limitations, it asymptotically approaches zero as the baseline phase grows longer. This occurs because, under normal distribution assumptions, the likelihood of observing a large value in the baseline phase increases with the number of observations—and the PND value is limited by the maximum observation in the baseline phase. Tarlow and Penland (2016) recommended a method for null hypothesis significance testing, which in a way can compensate for this

limitation, because the statistical power of PND increases with additional baseline phase observations, even as its estimated effect size deceases.

All effect sizes indices grew more precise with longer baseline phases. This suggests that, even with statistics that are nominally unaffected by increasing or decreasing baseline phase (e.g., Baseline Corrected Tau, Tau-U), the standard errors grow smaller when baselines are longer. This is an important result for both statistical significance testing and quantitative synthesis of single-case studies.

**Question 4: Autocorrelation**

Surveys of single-case research suggest brief interrupted time-series behavioral data are serially dependent; however, autocorrelation is, on average, small, statistically significant, and heterogeneously distributed in different domains of research (Shadish & Sullivan, 2011; Solomon, 2013). Effect size estimates from least squares regression-based methods tend to be linearly related to the magnitude of autocorrelation (Manolov & Solanas, 2008; Tarlow, in press). By comparison, the six indices included in this study did not yield effect size estimates which were substantially affected by serial dependency. In both lag-1 autoregressive and moving average autocorrelation models, effect size estimates demonstrated an inverted-U pattern, where the largest effect sizes were estimated when the autocorrelation coefficient ($\varphi_1$ or $\theta_1$) was zero (as an exception, PND values slightly increased with large positive lag-1 autoregressive error).

However, despite the lack of linear relationship between serial dependency and effect size, the standard errors of the effect sizes did in fact demonstrate a linear relationship to autocorrelation. To further complicate this finding, the direction of this

relationship depended on whether an autoregressive or moving average error model was specified. Standard errors increased with lag-1 autoregressive error ($\varphi_1$), but they decreased with lag-1 moving average error ($\theta_1$). Put another way, increased autoregressive autocorrelation will lead to increased Type I error, and increased moving average autocorrelation will lead to increased Type II error. Autoregressive error models are usually assumed in single-case research, but there is little rationale for doing so. More investigation is needed to determine how plausible moving average models are in behavioral data. Single-case investigators should also heed Shadish's (2014b) warning that nonoverlap and nonparametric statistics are not free of assumptions, and are indeed affected by serial dependency.

**Question 5: Heteroscedasticity**

In general, effect size estimates increased with more stable observations, i.e., smaller within-phase standard deviations. When unequal variances between phases were modeled, baseline phase variability had a larger influence on effect size estimates than B phase variability. PND was an exception; PND values increased as B phase variability increased (for the same reason as discussed above with regard to baseline phase length). ITSSIM demonstrated the smallest effect of B phase variability, though A phase variability had a large influence on its effect size estimates.

## Conclusion

In describing "the shape of things to come" in single-case research, Shadish (2014b) stated

> Past SCD effect sizes lacked formal development from clear assumptions in statistical theory. As a result, their confidence intervals and significance tests are

> of unclear validity or are nonexistent, their power to detect effects is unknown, and we know little about how they perform in the face of variation in the number of observations per phase (phase length), in how observations change systematically over time even in the absence of treatment (trend), in how correlated observations within each case might be (autocorrelation), and in the outcome-measurement metrics (e.g., count, percentage, normally distributed data). (p. 142)

In the same article, Shadish also made an appeal for better specialized single-case statistical software, noting "many clinical scientists will understandably use simple statistical programs even if they are not state of the art" (p. 144). ITSLAB was developed in response to this request and to the lack of clarity regarding the nonparametric and nonoverlap statistics which have flourished in single-case research. This paper demonstrates how ITSLAB can be used by any investigator, regardless of statistical computing experience, to discover useful insights for both research and clinical practice.

For researchers and methodologists, there are many more questions to explore with ITSLAB. Interactions between simulation parameters should be explored—for example, the interaction between trend and autocorrelation, which are difficult to estimate when both are present in time-series data (e.g., Yue et al., 2002). ITSLAB also outputs quartiles and statistical power for simulation models, and investigations into the distributions and power of single-case effect size indices will be useful. The autocorrelation findings in this study suggests that researchers should determine if moving average autocorrelation models are plausible in single-case data and, if so, how to address that in treatment effect size estimation. Simulation methods like ITSLAB may also provide researchers with a tool for converting effect size estimates between the

various single-case statistics, which could be useful for the meta-analysis of single-case studies.

For practitioners, ITSLAB's statistical power results make what-if analyses possible when designing a single-case experiment. The results from this study also suggest several recommendations to practitioners who wish to integrate statistical methods with their single-case designs. First, practitioners should always maximize the length of their baseline phases. Gorman and Allison's (1996) advice is as relevant now as it was two decades ago: "Perhaps the strongest suggestion for change that we can make is that researchers should collect a greater number of observations" (p. 208). No other single design feature is as important in a single-case experiment. Second, baselines should ideally be stable, with as little variability as possible. Attaining a stable baseline is more important than attaining a stable treatment phase when the goal is to detect treatment effects. Third, Tau-U and PND should not be used when there is any possibility of baseline trend.

ITSLAB is a powerful tool for single-case research, and it will be even more useful to researchers and practitioners with further development. Future versions should include additional effect size indices. It would also be helpful if ITSLAB modeled nonlinear trend patterns, and non-normal error distributions such as Poisson and binomial distributions. These and other developments would make sophisticated computer-intensive simulation methods accessible to virtually any single-case investigator.

# CHAPTER V

# CONCLUSIONS

This dissertation introduced a computer-intensive simulation-based method for the analysis and meta-analysis of single-case experimental designs. This method, Interrupted Time-Series Simulation (ITSSIM), was compared to several established effect size indices using real and simulated single-case data. The results of three studies suggest ITSSIM is a flexible metric that yields effect size estimates consistent with both simple and sophisticated methods. Unlike most other metrics, ITSSIM fits level, trend, variance, and autocorrelation parameters. This comprehensive model is useful when the underlying properties of time-series observations are unknown. ITSSIM is also accessible to single-case investigators via standalone software (Tarlow, 2017a), which does not require prior experience with statistical computing or syntax.

The first study (Chapter II) reviewed the theoretical rationale for simulation-based methods in single-case research. Single-case time-series are often brief, and therefore difficult to analyze in a precise way. Reliability can be improved with multilevel modeling methods, which pool information across cases and studies to give more accurate parameter estimates. Computer-intensive simulation offers another approach to the precision problem, but has received relatively little attention. While ITSSIM does not necessarily improve the precision of parameter estimation, it uses an iterative procedure to model many plausible parameter values given the observed data. By comparing a distribution of plausible "null effects"—the no-treatment predictions

based on baseline data—to a distribution of plausible treatment effects, ITSSIM calculates the most likely treatment effect size and, based on the simulated distributions, reports how precise the estimate is. Single-case data featured in a special issue of the *Journal of School Psychology* was reanalyzed with ITSSIM and compared to results from five sophisticated multilevel analyses of the same data. ITSSIM performed similarly to the multilevel methods, yielding similar standardized effect size estimates and slightly larger unstandardized estimates. An advantage of ITSSIM compared to multilevel modeling is its accessibility (via user-friendly software), easily interpretable effect size metrics (unstandardized effects, *d*-statistic, and $R^2$ or *r*-statistic), and comprehensive simulation model, which includes baseline trend, level- and slope-change effects, and autocorrelation.

The second study (Chapter III) applied ITSSIM to the meta-analysis of 10 single-case studies of cognitive therapies for depression, including a total of 53 cases. A multilevel method was used to synthesize results within- and between-studies. In addition to ITSSIM, five other single-case statistics were also included for comparison. Excluding one method identified as problematic, the five remaining statistics produced mean effect size estimates similar to meta-analyses of RCT psychotherapy research. Overall, these results suggest single-case experimental designs are a viable and valuable options for psychotherapy researchers who are interested in conducting cost-efficient pilot studies of innovative treatments, or studying individuals with complex/co-occurring disorders or low incidence clinical populations—both groups that are difficult to study with large-group designs. ITSSIM and another single-case statistic, Baseline Corrected

Tau, were identified as superior to the other comparison methods because they modeled baseline trend, did not produce extreme/uninterpretable values, and did not demonstrate ceiling effects.

The third study (Chapter III), reviewed a software tool for performing computer-intensive simulation research on single-case statistics, Interrupted Time-Series Lab (ITSLAB). ITSLAB was used to test ITSSIM and five other statistics under a range of parametric conditions, including manipulations of phase length, baseline trend, level-change, heteroscedasticity, and autocorrelation. ITSSIM adequately controlled for small-to-large baseline trend; it produced conservative effect size estimates with shorter baseline phases; its effect size estimates were unaffected by autocorrelation, though, notably, standard errors were influenced by autocorrelation; and effect size estimates were attenuated by increasing variance in the baseline phase. Overall, ITSSIM performed reliably well, even under extreme or unlikely simulation models. ITSLAB, the first program designed specifically for single-case simulation research, makes sophisticated computer-intensive methods accessible to virtually any single-case investigator. With tools like ITSLAB, the single-case research community will be better equipped to test and validate ITSSIM and other statistical methods.

In conclusion, ITSSIM is a powerful method for analysis and meta-analysis of single-case experimental designs. Its simulation model includes level, trend, variance, and autcorrelation parameters, and is therefore more comprehensive than most single-case statistics (which include some, but not all, of those critical parameters). ITSSIM effect size estimates are generally consistent with results from other, previously

validated methods. It also performed predictably under a series of simulation conditions. With additional study and application, single-case investigators can continue to evaluate ITSSIM's potential for modeling treatment outcomes and, ultimately, illuminating how people get better.

* * *

> A prediction of what the *average* individual will do is often of little or no value in dealing with a particular individual … A science of behavior which concerns only the behavior of groups is not likely to be of help in our understanding of the particular case.
>
> —B. F. Skinner, *Science and Human Behavior* (1953)

# REFERENCES


Akbari, M., Roshan, R., Shabani, A., Shairi, M. R., & Zarghami, F. (2015). Transdiagnostic treatment of co-occurrence of anxiety and depressive disorders based on repetitive negative thinking: A case series. *Iranian Journal of Psychiatry, 10*(3), 200-211.

Allison, D. B., & Gorman, B. S. (1993). Calculating effect sizes for meta-analysis: The case of the single case. *Behaviour Research and Therapy, 31*(6), 621-631.

Allison, D. B., & Gorman, B. S. (1994). "Make things as simple as possible, but no simpler." A rejoinder to Scruggs and Mastropieri. *Behaviour Research and Therapy, 32*(8), 885-890.

Anderson, R. L. (1942). Distribution of the serial correlation coefficient. *The Annals of Mathemetical Statistics, 13*(1), 1-13.

APA Presidential Task Force on Evidence-Based Practice (2006). Evidence-based practice in psychology. *American Psychologist, 61*(4), 271-285.

Baer, D. M. (1977). "Perhaps it would be better not to know everything". *Journal of Applied Behavior Analysis, 10*(1), 167-172.

Baer, D. M. (1988). An autocorrelated commentary on the need for a different debate. *Behavioral Assessment, 10,* 295-297.

Barlow, D. H., & Hersen, M. (1984). *Single-case experimental designs: Strategies for studying behavior change* (2nd ed.). Elmsford, NY: Pergamon Press.

Barlow, D. H., & Nock, M. K. (2009). Why can't we be more idiographic in our research? *Perspectives on Psychological Science, 4*(1), 19-21.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*(6), 1173-1182.

Bartlett, M. S. (1946). On the theoretical specification and sampling properties of autocorrelated time-series. *Journal of the Royal Statistical Society, 8*(1), 27-41.

Beck, A. T. (1967). *Depression: Clinical, experimental, and theoretical aspects.* New York, NY: Harper & Row.

Beck, A. T. (1976). *Cognitive therapy and the emotional disorders.* New York, NY: Meridian.

Beck, A. T. (2005). The current state of cognitive therapy: A 40-year retrospective. *Archives of General Psychiatry, 62*(9), 953-959.

Beck, A. T., Rush, A. J., Shaw, B. F., & Emery, G. (1979). *Cognitive therapy of depression.* New York, NY: Guilford Press.

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Manual for the Beck Depression Inventory-II.* San Antonio, TX: Psychological Corporation.

Beck, J. G., Tran, H. N., Dodson, T. S., Henschel, A. V., Woodward, M. J., & Eddinger, J. (2016). Cognitive trauma therapy for battered women: Replication and extension. *Psychology of Violence, 6*(3), 368-377.

Beretvas, S. N., & Chung, H. (2011). A review of meta-analyses of single-subject experimental designs: Methodological issues and practice. *Evidence-Based Communication Assessment and Intervention, 2*(3), 129-141.

Borckardt, J. J., Nash, M. R., Murphy, M. D., Moore, M., Shaw, D., & O'Neil, P. (2008). Clinical practice as natural laboratory for psychotherapy research: A guide to case-based time-series analysis. *American Psychologist, 63*(2), 77-95.

Bowman-Perrott, L., Davis, H., Vannest, K., & Williams, L. (2013). Academic benefits of peer tutoring: A meta-analytic review of single-case research. *School Psychology Review, 42*(1), 39-55.

Box, G. E. P., & Jenkins, G. M. (1970). *Time series analysis: Forecasting and control.* San Francisco, CA: Holden-Day.

Brossart, D. F., Parker, R. I., & Castillo, L. C. (2011). Robust regression for single-case data analysis: How can it help? *Behavior Research Methods, 43*(3), 710-719.

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.

Busk, P. L., & Marascuilo, L. A. (1988). Autocorrelation in single-subject research: A counterargument to the myth of no autocorrelation. *Behavioral Assessment, 10,* 229-242.

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. 187-212). Hillsdale, NJ: Lawrence Erlbaum Associates.

Butler, A. C., Chapman, J. E., Forman, E. M., & Beck, A. T. (2006). The empirical status of cognitive-behavioral therapy: A review of meta-analyses. *Clinical Psychology Review, 26*(1), 17-31.

Campbell, J. M. (2004). Statistical comparison of four effect sizes for single-subject designs. *Behavior Modification, 28*(2), 234-246.

Campbell, J. M., & Herzinger, C. V. (2010). Statistics and single-subject research methodology. In D. Gast (Ed.), *Single-subject research methodology in behavioral science* (pp. 417-453). New York, NY: Routledge.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Chicago, IL: Rand McNally.

Castelnuovo, G., Faccio, E., Molinari, E., Nardone, G., & Salvini, A. (2004). A critical review of Empirically Supported Treatments (ESTs) and Common Factors perspectives in psychotherapy. *Brief Strategic and Systemic Therapy European Review, 1,* 208-224.

Center, B. A., Skiba, R. J., & Casey, A. (1985-1986). A methodology for the quantitative synthesis of intra-subject design research. *Journal of Special Education, 19*(4), 387-400.

Cheung, M. W.-L. (2015). metaSEM: An R package for meta-analysis using structural equation modeling. *Frontiers in Psychology, 5*(2015), e7.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*(12), 997-1003.

Cook, C. (2000, January). *A review of intraclass correlation.* Paper presented at the annual meeting of the Southwest Educational Research Association, Dallas, TX.

Cowles, M., & Nightingale, J. (2015). Diagnosis-specific CBT as a stepping stone to transdiagnostic CBT in a complex case. *The Cognitive Behaviour Therapist, 8* (2015), e18.

Cox, D. R. (1966). The null distribution of the first serial correlation coefficient. *Biometrika, 53*(3/4), 623-626.

Crawford, J. R., Garthwaite, P. H. (2006). Methods of testing for a deficit in single-case studies: Evaluation of statistical power by Monte Carlo simulation. *Cognitive Neuropsychology, 23*(6), 877-904.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist, 12*(11), 671-684.

Crosbie, J. (1987). The inability of the binomial test to control Type I error with single-subject data. *Behavioral Assessment, 9*(2), 141-150.

Crosbie, J. (1995). Interrupted time-series analysis with short series: Why it is problematic; how it can be improved. In J. M. Gottman (Ed.), *The analysis of change* (pp. 361-395). Mahwah, NJ: Lawrence Erlbaum.

Danov, S. E., & Symons, F. J. (2008). A survey of evaluation of the reliability of visual inspection and functional analysis graphs. *Behavior Modification, 32*(6), 828-839.

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

Ebbinghaus, H. (1885/1913). *Memory: A contribution to experimental psychology* (H. A. Rueger & C. E. Bussenius, Trans.). New York, NY: Teachers College, Columbia University.

Edwards, D. J. A., Dattilio, F. M., & Bromley, D. B. (2004). Developing evidence-based practice: The role of case-based research. *Professional Psychology: Research and Practice, 35*(6), 589-597.

Efron, B. (2013). A 250-year argument: Belief, behavior, and the bootstrap. *Bulletin of the American Mathematical Society, 50*(1), 129-146.

Faith, M. S., Allison, D. B., & Gorman, B. S. (1996). Meta-analysis of single-case research. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 245-277). Mahwah, NJ: Lawrence Erlbaum.

Fechner, G. T. (1889). *Elemente der Psychophysik* [Elements of Psychophysics]. Leipzig: Druck und Verlag.

Ferron, J. (2002). Reconsidering the use of the general linear model with single-case data. *Behavior Research Methods, Instruments, & Computers, 34*(3), 324-331.

Fishman, D. B. (2005). Editors introduction to PCSP—From single case to database: A new method for enhancing psychotherapy practice. *Pragmatic Case Studies in Psychotherapy, 1*(1), 1-50.

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

Garfield, S. L. (1996). Some problems associated with "validated" forms of psychotherapy. *Clinical Psychology: Science and Practice, 3*(3), 218-229.

GetData Graph Digitizer. (2013). *GetData Graph Digitizer (2.26).* Available from http://www.getdata-graph-digitizer.com

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*(10), 3-8.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research, 42*(3), 237-288.

Glass, G. V., Willson, V. L., & Gottman, J. M. (1975). *Design and analysis of time-series experiments.* Boulder, CO: Colorado Associated University Press.

Gorsuch, R. L. (1983). Three methods for analyzing limited time-series (N of 1) data. *Behavioral Assessment, 7*(2), 141-154.

Harbst, K. B., Ottenbacher, K. J.,& Harris, S. R. (1991). Interrater reliability of therapists' judgments of graphed data. *Physical Therapy*, *71*(2), 107-115.

Hartmann, D. P., Gottman, J. M., Jones, R. R., Gardner, W., Kazdin, A. E., & Vaught, R. S. (1980). Interrupted time-series analysis and its application to behavioral data. *Journal of Applied Behavior Analysis, 13*(4), 1980.

Hayes, S. C. (1981). Single case experimental design and empirical clinical practice. *Journal of Consulting and Clinical Psychology, 49*(2), 193-211.

Hilliard, R. B. (1993). Single-case methodology in psychotherapy: Process and outcome research. *Journal of Consulting and Clinical Psychology, 61*(3), 373-380.

Holländare, F., Eriksson, A., Lövgren, L., Humble, M. B., & Boersma, K. (2015). Internet-based cognitive behavioral therapy for residual symptoms in bipolar disorder type II: A single-subject design pilot study. *JMIR Research Protocols, 4*(2), e44.

Holmes, E. A., Bonsall, M. B., Hales, S. A., Mitchell, H., Renner, F., Blackwell, S. E., Watson, P., Goodwin, G. M., & Di Simplicio, M. (2016). Applications of time-series analysis to mood fluctuations in bipolar disorder to promote treatment innovation: A case series. *Translational Psychiatry, 6*(1), e720.

Huitema, B. E. (1985). Autocorrelation in applied behavior analysis: A myth. *Behavioral Assessment, 7,* 107-118.

Huitema, B. E. (1988). Autocorrelation: 10 years of confusion. *Behavioral Assessment, 10,* 253-294.

Huitema, B. E., & McKean, J. W. (1991). Autocorrelation estimation and inference with small samples. *Psychological Bulletin, 110*(2), 291-304.

Huitema, B. E., & McKean, J. W. (1998). Irrelevant autocorrelation in least-squares intervention models. *Psychological Methods, 3*(1), 104-116.

Huitema, B. E., & McKean, J. W. (2000a). A simple and powerful test for autocorrelated errors in OLS intervention models. *Psychological Reports, 87*(1), 3-20.

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

Iwakabe, S., & Gazzola, N. (2009). From single-case studies to practice-based knowledge: Aggregating and synthesizing case studies. *Psychotherapy Research, 19*(4-5), 601-611.

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

Jones, M. C. (1924). The elimination of children's fears. *Journal of Experimental Psychology, 7*(5), 383-390.

Jones, E. E., Ghannam, J., Nigg, J. T., & Dyer, J. F. P. (1993). A paradigm for single-case research: The time series study of a long-term psychotherapy for depression. *Journal of Consulting and Clinical Psychology, 61*(3), 381-394.

Jones, R. R., Russell, S. V., & Weinrott, M. (1977). Time-series analysis in operant research. *Journal of Applied Behavior Analysis, 10*(1), 151-166.

Jones, S., Ownsworth, T., & Shum, D. H. K. (2015). Feasibility and utility of telephone-based psychological support for people with brain tumor: A single-case experimental study. *Frontiers in Oncology, 5*(2015), e16.

Kazdin, A. E. (1982). *Single-case research designs: Methods for clinical and applied settings.* New York, NY: Oxford University Press.

Kendall, M. G. (1962). *Rank correlation methods* (3rd ed.). New York, NY: Hafner.

Kratochwill, T. R., & Brody, G. H. (1978). Single subject designs: A perspective on the controversy over employing statistical inference and implications for research and training in behavior modification. *Behavior Modification, 2*(3), 291-307.

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). *Single-case design technical documentation.* Washington, D.C.: What Works Clearinghouse.

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine, 16*(9), 606-613.

Kuhn, T. S. (1996). *The Structure of Scientific Revolutions* (3rd ed.). Chicago, IL: University of Chicago Press.

Lambert, M. C., Cartledge, G., Heward, W. L., & Lo, Y. 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.

Lieberman, R. G., Yoder, P. J., Reichow, B., & Wolery, M. (2010). Visual analysis of multiple baseline across participants graphs when change is delayed. *School Psychology Quarterly, 25*(1), 28-44.

Lovibond, S., & Lovibond, P. (1995). *Manual for the Depression Anxiety Stress Scales* (2nd Ed.). Sydney: Psychology Foundation.

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

Maggin, D. M., O'Keeffe, B. V., & Johnson, A. H. (2011). A quantitative synthesis of methodology in the meta-analysis of single-subject research for students with disabilities: 1985-2009. *Exceptionality, 19*(2), 109-135.

Maggin, D. M., & Odom, S. L. (2014). Evaluating single-case research data for systematic review: A commentary for the special issue. *Journal of School Psychology, 52*(2), 237-241.

Mann, H. B. (1945). Nonparametric tests against trend. *Econometrica, 13*(3), 245-259.

Manolov, R., & Solanas, A. (2008). Comparing N = 1 effect size indices in presence of autocorrelation. *Behavior Modification, 32*(6), 860-875.

Manolov, R., & Solanas, A. (2009). Percentage of nonoverlapping corrected data. *Behavior Research Methods, 41*(4), 1262-1271.

Manolov, R., & Solanas, A. (2012). Assigning and combining probabilities in single-case studies. *Psychological Methods, 17*(4), 495-509.

Manolov, R. & Solanas, A. (2013). A comparison of mean phase difference and generalized least squares for analyzing single-case data. *Journal of School Psychology, 51,* 201-215.

Manolov, R., Solanas, A., Sierra, V., & Evans, J. J. (2011). Choosing among techniques for quantifying single-case intervention effectiveness. *Behavior Therapy, 42*(3), 533-545.

Marks, L. (1986). *Behavioural psychotherapy.* Bristol, UK: John Wright.

Maroti, D., Folkeson, P., Jansson-Fröjmark, M., & Linton, S. J. (2011). Does treating insomnia with cognitive-behavioural therapy influence comorbid anxiety and depression? An exploratory multiple baseline design with four patients. *Behaviour Change, 28*(4), 195-205.

Matyas, T. A., & Greenwood, K. M. (1990). Visual analysis of single-case time series: Effects of variability, serial dependence, and magnitude of intervention effects. *Journal of Applied Behavior Analysis, 23*(3), 341-351.

Matyas, T. A., & Greenwood, K. M. (1991). Problems in the estimation of autocorrelation in brief time series and some implications for behavioral data. *Behavioral Assessment, 13*(2), 137-157.

McCleary, R., & Welsh, 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, NJ: Lawrence Erlbaum Associates.

McManus, F., Clark, G., Muse, K., & Shafran, R. (2015). Case-series evaluating a transdiagnostic cognitive-behavioural treatment for co-occurring anxiety disorders. *Behavioural and Cognitive Psychotherapy, 43*(6), 744-758.

Mehranfar, M., Younesi, J., & Banihashem, A. (2012). Effectiveness of mindfulness-based cognitive-therapy on reduction of depression and anxiety symptoms in mothers of children with cancer. *Iranian Journal of Cancer Prevention, 5*(1), 1-10.

Moeyaert, M., Ferron, J. M., Beretvas, S. N., & Van den Noortgate, W. (2014). From a single-level analysis to a multilevel analysis of single-case experimental designs. *Journal of School Psychology, 52*(2), 191-211.

Montgomery, S. A., & Asberg, M. (1979). A new depression scale designed to be sensitive to change. *British Journal of Psychiatry, 134*(4), 382-389.

Morgan, D. L., & Morgan, R. K. (2001). Single-participant research design: Bringing science to managed care. *American Psychologist, 56*(2), 119-127.

Park, H., Marascuilo, L., & Gaylord-Ross, R. (1990). Visual inspection and statistical analysis of single-case designs. *Journal of Experimental Education, 58*(4), 311-320.

Parker, R. I., & Brossart, D. F. (2003). Evaluating single-case research data: A comparison of seven statistical methods. *Behavior Therapy, 34*(2), 189-211.

Parker, R. I., Brossart, D. F., Vannest, K. J., Long, J. R., Garcia De-Alba, R., Baugh, F. G., & Sullivan, J. R. (2005). Effect sizes in single case research: How large is large? *School Psychology Review, 34*(1), 116-132.

Parker, R. I., Cryer, J., & Byrns, G. (2006). Controlling baseline trend in single-case research. *School Psychology Quarterly, 21*(4), 418-443.

Parker, R. I. & Hagan-Burke, S. (2007). Useful effect size interpretations for single case research. *Behavior Therapy, 38*(1), 95-105.

Parker, R. I., Hagan-Burke, S., & Vannest, K. (2007). Percentage of all non-overlapping data (PAND): An alternative to PND. *Journal of Special Education, 40*(4), 194-204.

Parker, R. I., & Vannest, K. J. (2009). An improved effect size for single-case research: Nonoverlap of All Pairs. *Behavior Therapy, 40*(4), 357-367.

Parker, R. I., & Vannest, K. J. (2012). Bottom-up analysis of single-case research designs. *Journal of Behavioral Education, 21*(3), 254-265.

Parker, R. I., Vannest, K. J., & Davis, J. L. (2011). Effect size in single-case research: A review of nine nonoverlap techniques. *Behavior Modification, 35*(4), 303-322.

Parker, R. I., Vannest, K. J., Davis, J. L., & Sauber, S. B. (2011). Combining nonoverlap and trend for single-case research: Tau-U. *Behavior Therapy, 42*(2), 284-299.

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* (pp. 15-40). Hillsdale, NJ: Lawrence Erlbaum.

Persons, J. B., & Silberschatz, G. (1998). Are results of randomized controlled trials useful to psychotherapists? *Journal of Consulting and Clinical Psychology, 66*(1), 126-135.

Piaget, J. (1952). *The origins of intelligence in children* (M. Cook, Trans.). New York, NY: International University Press.

Rindskopf, D. (2014). Nonlinear Bayesian analysis for single case designs. *Journal of School Psychology, 52*(2), 179-189.

Rush, A. J., Trivedi, M. H., Ibrahim, H. M., Carmody, T. J., Arnow, B, Klein, D. N., … & Thase, M. E. (2003). The 16-item Quick Inventory of Depressive Symptomology (QIDS), Clinician Rating (QIDS-C), and Self-Report (QIDS-SR):

A psychometric evaluation in patients with chronic major depression. *Biological Psychiatry, 54*(5), 573-583.

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

SAS Institute (2012). *SAS Software, Release 9.4.* Cary, NC: Author.

Schlosser, R. W., Lee, D. L., & Wendt, O. (2008). Application of the percentage of non-overlapping data (PND) in systematic reviews and meta-analyses: A systematic review of reporting characteristics. *Evidence-Based Communication Assessment and Intervention, 2*(3), 163-187.

Schwarzer, G. (2007). meta: An R package for meta-analysis. *R News, 7*(3), 40-45.

Scruggs, T. E., & Mastropieri, M. A. (2001). How to summarize single-participant research: Ideas and applications. *Exceptionality: A Special Education Journal, 9*(4), 227-244.

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

Searson, R., Mansell, W., Lowens, I., & Tai, S. (2012). Think Effectively About Mood Swings (TEAMS): A case series of cognitive-behavioural therapy for bipolar disorders. *Journal of Behavior Therapy and Experimental Psychiatry, 43*(2), 770-779.

Seligman, M. E. P. (1995). The effectiveness of psychotherapy: The Consumer Reports study. *American Psychologist, 50*(12), 965-974.

Sen, P. K. (1968). Estimates of the regression coefficient based on Kendall's Tau. *Journal of the American Statistical Association, 63*(324), 1379-1389.

Shadish, W. R. (2014a). Analysis and meta-analysis of single-case designs: An introduction. *Journal of School Psychology, 52*(2), 109-122.

Shadish, W. R. (2014b). Statistical analyses of single-case designs: The shape of things to come. *Current Directions in Psychological Science, 23*(2), 139-146.

Shadish, W. R., Brasil, I. C. 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., Hedges, L. V., & Pustejovsky, J. E. (2014). Analysis and meta-analysis of single-case designs with a standardized mean difference statistic: A primer and applications. *Journal of School Psychology, 52*(2), 123-147.

Shadish, W. R., Hedges, L. V., Pustejovsky, J. E., Boyajian, J. G., Sullivan, K. J., Andrade, A., & Barrientos, J. L. (2014). A *d*-statistic for single-case designs that is equivalent to the usual between-groups *d*-statistic. *Neuropsychological Rehabilitation, 24*(3-4), 528-553.

Shadish, W. R., & Rindskopf, D. M. (2007). Methods for evidence-based practice: Quantitative synthesis of single-subject designs. *New Directions for Evaluation, 113,* 95-109.

Shadish, W. R., Rindskopf, D. M., & Hedges, L. V. (2008). The state of science in the meta-analysis of single-case experimental designs. *Evidence-Based Communication Assessment and Intervention, 2*(3), 188-196.

Shadish, W. R., & Sullivan, K. J. (2011). Characteristics of single-case designs used to assess intervention effects in 2008. *Behavior Research Methods, 43*(4), 971-980.

Shadish, W. R., Zuur, A. F., & Sullivan, K. J. (2014). Using generalized additive (mixed) models to analyze single case designs. *Journal of School Psychology, 52*(2), 149-178.

Sharpley, C. F., & Alavosius, M. P. (1988). Autocorrelation in behavioral data: An alternative perspective. *Behavioral Assessment, 10,* 241-251.

Sidman, M. (1960). *Tactics of scientific research: Evaluating experimental data in psychology.* New York, NY: Basic Books.

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics, 24*(4), 323-355.

Sivo, S. A., & Willson, V. L. (2000). Modeling causal error structures in longitudinal panel data: A Monte Carlo study. *Structural Equation Modeling, 7*(2), 174-205.

Skinner, B. F. (1938). *The behavior of organisms.* New York, NY: Basic Books.

Skinner, B. F. (1948). 'Superstition' in the pigeon. *Journal of Experimental Psychology, 38*(2), 168-172.

Skinner, B. F. (1953). *Science and human behavior.* New York, NY: The Free Press.

Skinner, B. F. (1956). A case history in the scientific method. *American Psychologist,*

Smith, J. D. (2012). Single-case experimental designs: A systematic review of published research and current standards. *Psychological Methods, 17*(4), 510-550.

Smith, J. D., Borckardt, J. J., & Nash, M. R. (2012). Inferential precision in single-case time-series data streams: How well does the EM procedure perform when missing observations occur in autocorrelated data? *Behavior Therapy, 43*(3), 679-685.

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist, 32*(9), 752-760.

Smith, M. L., Glass, G. V., & Miller, T. I. (1980). *The benefits of psychotherapy*. Baltimore, MD: Johns Hopkins University Press.

Solanas, A., Manolov, R., & Onghena, P. (2010). Estimating slope and level change in *N* = 1 designs. *Behavior Modification, 34*(3), 195-218.

Solanas, A., Manolov, R., & Sierra, V. (2010). Lag-one autocorrelation in short series: Estimation and hypothesis testing. *Psicológica, 31*(2), 357-381.

Solomon, B. G. (2013). Violations of assumptions in school-based single-case data: Implications for the selection and interpretation of effect sizes. *Behavior Modification, 38*(4), 477-496.

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

Suen, H. K., & Ary, D. (1987). Autocorrelation in applied behavior analysis: Myth or reality? *Behavioral Assessment, 9,* 125-130.

Swaminathan, H., Rogers, H. J., & Horner, R. H. (2014). An effect size measure and Bayesian analysis of single-case designs. *Journal of School Psychology, 52*(2), 213-230.

Szentagotai, A., & David, D. (2010). The efficacy of cognitive-behavioral therapy in bipolar disorder: A quantitative meta-analysis. *The Journal of Clinical Psychiatry, 71*(1), 66-72.

Tarlow, K. R. (in press). An improved rank correlation effect size statistic for single-case designs: Baseline Corrected Tau. *Behavior Modification.*

Tarlow, K. R. (2016). *Baseline Corrected Tau calculator.* Retrieved from http://www.ktarlow.com/stats/tau/

Tarlow, K. R. (2017a). *ITSSIM: Interrupted Time-Series Simulation, Version 1.0.* College Station, TX: Author. Retrieved from http://ktarlow.com/stats/itssim/

Tarlow, K. R. (2017b). *ITSLAB: Interrupted Time-Series Lab, Version 1.0.* College Station, TX: Author. Retrieved from http://ktarlow.com/stats/itslab/

Tarlow, K. R., & Penland, A. (2016a). Outcome assessment and inference with the Percentage of Nonoverlapping Data (PND) single-case statistic. *Practice Innovations, 1*(4), 221-233.

Tarlow, K. R., & Penland, A. (2016b). *Percentage of Nonoverlapping Data (PND) calculator.* Retrieved from http://www.ktarlow.com/stats/pnd/

Theil, H. (1950). A rank-invariant method of linear and polynomial regression analysis, I, II, and III. *Proceedings of the Royal Netherlands Academy of Sciences, 53,* 386-392, 521-525, and 1397-1412.

Thompson, B. (2006). *Foundations of behavioral statistics: An insight-based approach.* New York, NY: Guilford.

Ugille, M., Moeyaert, M., Beretvas, S. N., Ferron, J., & Van den Noortgate, W. (2012). Multilevel meta-analysis of single-subject experimental designs: A simulation study. *Behavior Research Methods, 44*(4), 1244-1254.

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

Van den Noortgate, W., & Onghena, P. (2007). Combining single-case experimental data using hierarchical linear models. *School Psychology Quarterly, 18*(3), 325-346.

Van den Noortgate, W., & Onghena, P. (2008). A multilevel meta-analysis of single-subject experimental design studies. *Evidence Based Communication Assessment and Intervention, 2*(3), 142–151.

Vannest, K. J., Davis, J. L., Mason, B. A., Burke, M. D. (2010). Effective intervention for behavior with a daily behavior report card: A meta-analysis. *School Psychology Review, 39*(4), 654-672.

Vannest, K. J., & Parker, R. I. (2012). Bottom-up analysis of single-case research designs. *Journal of Behavioral Education, 21*(3), 254-265.

Vannest, K. J., Parker, R. I., Davis, J. L., Soares, D. A., & Smith, S. L., (2012). The Theil-Sen slope for high-stakes decisions from progress monitoring. *Behavioral Disorders, 37*(4), 271-280.

Vannest, K. J., Parker, R. I., & Gonen, O. (2011). Single case research: Web based calculators for SCR analyses (Version 1.0) [Web-based application]. College Station, TX: Texas A&M University. Available from http://www.singlecaseresearch.org/

Velicer, W. F., & Harrop, J. (1983). The reliability and accuracy of time series model identification. *Evaluation Review, 7*(4), 551-560.

Wampold, B. E. (1988). Introduction. *Behavioral Assessment, 10,* 227-228.

Wampold, B. E., Fluckiger, C., Del Re, A. C., Yulish, N. E., Frost, N. D., Pace, B. T., Goldberg, S. B., Miller, S. D., Baardseth, T. P., Laska, K. M., & Hilsenroth, M. J. (2017). In pursuit of truth: A critical examination of meta-analyses of cognitive behavior therapy. *Psychotherapy Research, 27*(1), 14-32.

Wampold, B. E., Minami, T., Baskin, T. W., & Tierney, S. C. (2002). A meta-(re)analysis of the effects of cognitive therapy versus 'other therapies' for depression. *Journal of Affective Disorders, 68*(2), 159-165.

Watson, J. B., & Rayner, R. (1920). Conditioned emotional reactions. *Journal of Experimental Psychology, 3*(1), 1-14.

Westen, D., Novotny, C. M., Thompson-Brenner, H. (2004). The empirical status of empirically supported psychotherapies: Assumptions, findings, and reporting in controlled clinical trials. *Psychological Bulletin, 130*(4), 631-663.

White, O. R. & Haring, N. G. (1980). *Exceptional teaching: A multimedia training package.* Columbus, OH: Merrill.

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.

Wilcox, R. R. (1998). A note on the Theil-Sen regression estimator when the regressor is random and the error term is heteroscedastic. *Biometrical Journal, 40*(3), 261-268.

Wilcox, R. R. (2001). *Fundamentals of modern statistical methods: Substantially improving power and accuracy.* New York, NY: Springer-Verlag.

Wilson, G. T., & Rachman, S. J. (1983). Meta-analysis and the evaluation of psychotherapy outcome: Limitations and liabilities. *Journal of Consulting and Clinical Psychology, 51*(1), 54-64.

Wolery, M., Busick, M., Reichow, B., & Barton, E. E. (2010). Comparison of overlap methods for quantitatively synthesizing single-subject data. *Journal of Special Education, 44*(1), 18-28.

Ximenes, V. M., Manolov, R., Solanas, A. & Quera, V. (2009). Factors affecting visual inference in single-case designs. *Spanish Journal of Psychology, 12*(2), 823-832.

Yue, S., Pilon, P., Phinney, B., & Cavadias, G. (2002). The influence of autocorrelation on the ability to detect trend in hydrological series. *Hydrological Processes, 16,* 1807-1829.

# APPENDIX

# DEMONSTRATION OF ITSSIM CALCULATIONS WITH AN EXAMPLE DATA SET

ITSSIM effect size calculations will be demonstrated with the following hypothetical single-case interrupted time-series:

A phase: 5, 8, 5, 4, 6, 3, 4, 9

B phase: 1, 0, 2, 0, 0

## Stage I: Parameter Estimation

**Step 1: Theil-Sen regression.** Estimate Theil-Sen slope and intercept coefficients for A and B phase data. The Theil-Sen slope, $b$, is the median slope of all possible pairs of ($T_t$, $Y_t$) coordinates (Sen, 1968), where $T_t$ is the day/session/time/etc. variable at time $t$, and $Y_t$ is the observed score at time $t$. There are several methods for estimating the intercept, $a$, but the relatively simple procedure from Wilcox (2001) is used,

$$a = M_Y - bM_T$$

where $M_X$ and $M_Y$ are the medians of $T$ (time) and $Y$ (observed scores), respectively. For the A phase example data above, $a$ = 5.375, and $b$ = -0.083. For the B phase data, the time variable, $T_t$, is re-centered by subtracting $T_{t-nA}$, where $n_A$ is the length of the baseline phase. Re-centering B phase data allows for a more interpretable comparison of A phase and B phase intercept values. Re-centering data does not affect the results of the simulation analysis; however, if B phase data are not re-centered, the calculated intercept

value will be for $T_t = 0$ rather than $T_t = n_A$, which corresponds roughly to the immediate impact of treatment (for more on this topic, see Huitema & McKean, 2000b). The B phase Theil-Sen regression coefficients for the example data are therefore $a = 0.375$, and $b = -0.125$.

**Step 2: Theil-Sen standard errors.** Wilcox's (2001) bootstrap procedure is used to calculate standard errors for the Theil-Sen slope and intercept coefficients. For each phase, a set of $n$ coordinates are resampled from ($Tt$, $Yt$) with replacement; then a Theil-Sen slope and intercept are calculated for this new sample of data points. Wilcox recommended the bootstrap procedure be performed at least 600 times; however, given the brief time series common in single-case experiments, the bootstrap is iterated 10,000 times. The standard error is then calculated as the standard deviation of the distribution of 10,000 coefficients. For the A phase example data, $SE_a = 2.752$, and $SE_b = 0.540$. For the B phase data, $SE_a = 1.604$, and $SE_b = 0.489$.

**Step 3: Error variance.** Error residuals, $\varepsilon_t$, are extracted from the observed data using the regression equation

$$Y_t = a + bT_t + \varepsilon_t$$

When extracting the B phase residuals, the re-centered time variable $T_t - n_A$ is substituted. For the example data, this yields the residuals

$\varepsilon_A$: -0.292, 2.791, -0.126, -1.043, 1.040, -1.877, -0.794, 4.289

$\varepsilon_B$: 0.750, -0.125, 2.000, 0.125, 0.250

The standard deviation of the error residuals is then $SD_A = 2.092$, and $SD_B = 0.845$.

**Step 4: Error variance standard errors.** The standard error of the standard

deviations is calculated using the least squares method. For the example data, $SE_{SD(A)}$ = 0.501, and $SE_{SD(B)}$ = 0.295.

**Step 5: Autocorrelation estimation.** Before estimating the lag-1 autocorrelation coefficient, $r_1$, the regression residuals are standardized. Standardizing a time series does not alter its serial dependency. However, once standardized, autocorrelation can be estimated across both phases; increasing the length of the time series (by combining A and B phases) improves the precision of the $r_1$ estimate, which is known to be imprecise in small samples. The standardized error residuals of the example data are

$z_\varepsilon$: -0.140, 1.334, -0.060, -0.499, 0.497, -0.897, -0.380, 2.051, 0.888, -0.148, 2.367, 0.148, 0.296

The standard estimator for lag-1 autoregressive error is

$$r_1 = \frac{\sum_{t=2}^{N}(e_t)(e_{t-1})}{\sum_{t=1}^{N} e_t^2}$$

where $N$ is the number of observations in the time-series, and $e_t$ is the error term at time $t$. However, this estimator is biased in small samples (Anderson, 1942; Matyas & Greenwood, 1991). There are several ways to correct for the small sample bias in autocorrelation estimation (Solanas et al., 2010). The method recommended by Huitema and McKean (1991; Ferron, 2002) was selected due to its simplicity and empirical validation. Lag-1 autocorrelation is therefore calculated as

$$r_1 = \frac{\sum_{t=2}^{N}(e_t)(e_{t-1})}{\sum_{t=1}^{N} e_t^2} + \frac{P}{N}$$

where $P$ is the number of estimated parameters in the regression model used to extract residuals; in this case, $P$ = 4. For the example data standardized error residuals, $z_\varepsilon$, the

estimated autocorrelation is $r_1 = 0.133$.

**Step 6: Autocorrelation standard error.** The standard error for the autocorrelation estimator (Huitema & McKean, 2000; Moran, 1948) is calculated as

$$SE_{r_1} = \frac{r_1}{\sqrt{\frac{(N-2)^2}{(N^2)(N-1)}}}$$

For the example data, $SE_{r_1} = 0.244$.

**Stage II: Time-Series Simulation**

The parameter estimation stage (Stage I) yields seven parameter estimates and their standard errors: A phase level (intercept), A phase trend (slope), A phase standard deviation, B phase level (intercept), B phase trend (slope), B phase standard deviation, and cross-phase autocorrelation. The A phase estimates are referred to as the *null effect model.* The B phase estimates are referred to as the *experimental effect model.* These coefficients may be reported in a table such as Table 2. The distributions of all coefficients are assumed to be normal (Anderson, 1942; Cox, 1966; Mann, 1945; Sen, 1968). The distributions are assumed to be independent.

**Step 7: Randomly sample model coefficients.** Seven coefficients are randomly drawn from each of the seven parameter estimates' distributions—essentially, a data point is selected at random from a seven-dimensional multivariate normal distribution. For the example data, one hypothetical sample of model coefficients is presented in Table 10.

Table 10

*Randomly Sampled Parameter Estimates*

| | | **Coefficient** |
|---|---|---|
| **Null Model** | intercept | 6.762 |
| | slope | -0.698 |
| | *s* | 1.471 |
| **Exp. Model** | intercept | 1.198 |
| | slope | 0.608 |
| | *s* | 1.271 |
| **Autocorrelation** | $r_1$ | 0.208 |

**Step 8: Simulate autocorrelated time-series.** A time-series of length $n_B$ with lag-1 autoregressive error equal to the randomly sampled $r_1$ value is generated using the method of Manolov and Solanas (2008, 2009, 2012, 2013). The time-series is then standardized (this does not affect the autocorrelation value). Using the sampled coefficients in Table 10, two simulated time series could be

$z_{\varepsilon(1)}$: -0.082, -0.205, 1.764, 0.495, -0.907

$z_{\varepsilon(2)}$: 0.029, 1.603, -0.636, 0.266, 1.616

**Step 9: Add variance to simulated time-series.** The simulated time-series are multiplied by the simulated standard deviation terms to give residuals for the null model ($\varepsilon_{NULL}$) and experimental model ($\varepsilon_{EXP}$). Using the example time-series from Step 8 for the null model and experimental model, respectively, the time-series would be

$\varepsilon_{NULL}$: -0.121, -0.302, 2.595, 0.728, -1.334

$\varepsilon_{EXP}$: 0.037, 2.037, -0.808, 0.338, 2.054

**Step 10: Add slope and intercept to simulated time-series.** Simulated null model observations ($Y_{NULL}$) and experimental model observations ($Y_{EXP}$) are calculated using the regression equation in Step 3 and the simulated intercept and slope terms. Recall that simulated time series are generated for the time interval of the B phase. Under the null model coefficients, the time variable, $T_t$, is $n_A + 1$, $n_A + 2\ldots$, $n_A + n_B$, which corresponds to the original B phase time values. For the experimental model coefficients, which were estimated from re-centered data, the time variable, $T_t$, is *1, 2…, $n_B$*. For the example data, the simulated time series are

$Y_{NULL}$ = 0.359, -0.520, 1.679, -0.886, -3.646

$Y_{EXP}$ = 1.843, 4.451, 2.214, 3.968, 6.292

**Step 11: Calculate means of simulated time-series.** For the example data, $\overline{Y}_{NULL}$ = -0.603, and $\overline{Y}_{EXP}$ = 3.754.

**Step 12: 100,000 iterations of Steps 7 through 11.** Repeating the simulation procedures will yield 100,000 null model time-series means ($N_i$), and 100,000 experimental model time-series means ($E_i$). Per the central limit theorem, the two distributions of means will be approximately normal.

**Stage III: Effect Size Calculation**

**Step 13: Unstandardized effect.** Treatment effects are calculated from the means and standard deviations of the null model and experimental model mean distributions. The unstandardized treatment effect, *D*, is the difference of the distribution

means, i.e., the average change in participant performance over all 100,000 plausible simulated time-series.

$$D - \bar{E}_i - \bar{N}_i$$

For the example data, $D$ = 0.000 - 4.453 = -4.453.

**Step 14: Standardized effect.** The standardized mean difference, $d$, is calculated as

$$d = \frac{\bar{E}_i - \bar{N}_i}{S_{within}}$$

where

$$S_{within} = \sqrt{\frac{(n_A - 1)s_{\bar{N}_i}^2 + (n_B - 1)s_{\bar{E}_i}^2}{n_A + n_B - 2}}$$

For the example data, $d$ = -1.584. The standardized mean difference, $d$, can also be converted to a Pearson $r$ correlation with the equation

$$r = \frac{d}{\sqrt{d^2 + a}}$$

where

$$a = \frac{(n_A + n_B)^2}{n_A n_B}$$

For the example data, $r$ = -0.611.