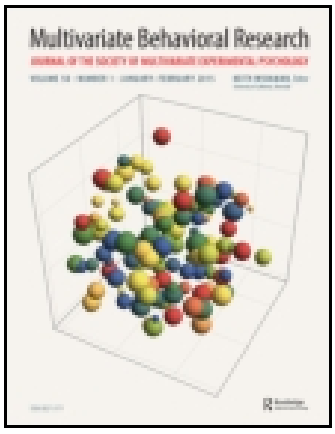
2015

Comparing Visual and Statistical Analysis in Single-Case Studies Using Published Studies

Magadalena Harrington
University of Rhode Island

Wayne Velicer
University of Rhode Island, wvelicer@uri.edu

Citation/Publisher Attribution
Harrington, M., & Velicer, W. F. (2015). Comparing Visual and Statistical Analysis in Single-Case Studies Using Published Studies. *Multivariate Behavioral Research, 50*(2), 162-183.
Available at: http://dx.doi.org/10.1080/00273171.2014.973989



Multivariate Behavioral Research
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hmbr20

Comparing Visual and Statistical Analysis in Single-Case Studies Using Published Studies

Magadalena Harrington[a] & Wayne F. Velicer[a]
[a] University of Rhode Island Cancer Prevention Research Center
Published online: 16 Apr 2015.

Multivariate Behavioral Research, 50:162–183, 2015

ISSN: 0027-3171 print / 1532-7906 online
DOI: 10.1080/00273171.2014.973989

Routledge
Taylor & Francis Group

Comparing Visual and Statistical Analysis in Single-Case Studies Using Published Studies

Magadalena Harrington and Wayne F. Velicer
University of Rhode Island Cancer Prevention Research Center

Little is known about the extent to which interrupted time series analysis (ITSA) can be applied to short, single-case study designs and whether those applications produce results consistent with visual analysis (VA). This article examines the extent to which ITSA can be applied to single-case study designs and compares the results based on two methods: ITSA and VA, using papers published in the *Journal of Applied Behavior Analysis* in 2010. The study was made possible by the development of software called UnGraph®, which facilitates the recovery of raw data from the graphs. ITSA was successfully applied to 94% of the examined graphs with the number of observations ranging from 8 to 136. Moderate to high lag-1 autocorrelations (>.50) were found for 46% of the data series. Effect sizes similar to group-level Cohen's *d* were identified based on the tertile distribution. Effects ranging from 0.00 to 0.99 were classified as small, those ranging from 1.00 to 2.49 as medium, and large effect sizes were defined as 2.50 or greater. Comparison of the conclusions from VA and ITSA had a low level of agreement (Kappa = .14, accounting for the agreement expected by chance). The results demonstrate that ITSA can be broadly implemented in applied behavior analysis research. These two methods should be viewed as complementary and used concurrently.

Group-level and single-case research designs are two methodological models employed for analyzing longitudinal research. The first model is based on data obtained from a large number of individuals and provides average estimates of longitudinal trajectories of behavior change based on group-level data, emphasizing between-subject variability. A significant limitation of group-level designs, also known as nomothetic designs, is the inability to capture high levels of variability and heterogeneity within the studied populations (Molenaar, 2004). Further, group-level designs emphasize central tendencies of the population and consequently obscure natural patterns of behavior change, their multidimensionality and unique variability within each individual (Molenaar & Campbell, 2009).

The second methodological approach employed in longitudinal research is based on data obtained from one individual or unit ($N = 1$) through intensive data collection over time. Single-case designs, also known as idiographic designs, examine individual-level data, which allows for highly accurate estimates of within-subject variability and longitudinal trajectories of each individual's behavior. Idiographic methodology characterizes highly heterogeneous processes, which consequently allow for more accurate inferences about the nature of behavior change specific to an individual (Velicer & Molenaar, 2013). Single-case designs address the limitations of group-level designs and present several advantages. They allow for a highly accurate assessment of the impact of the intervention for each individual while group-level designs provide information about the effectiveness of the intervention for an "average" person, rather than any person in particular (Velicer & Molenaar, 2013).

In addition, single-case research allows studying longitudinal processes of change with much better precision than group-level designs, due to a higher number of data points and better controlled variability of the data. Also, it can be applied to populations that are otherwise difficult to recruit in numbers large enough to allow for a group-level design (Barlow, Nock, & Hersen, 2009; Kazdin, 2011).

Ergodic Theorems

The discrepancies between results from cross-sectional nomothetic data and those from longitudinal idiographic data

Correspondence concerning this article should be addressed to Wayne F. Velicer, Ph.D., University of Rhode Island Cancer Prevention Research Center, 130 Flagg Road, Kingston, RI 02881. E-mail: velicer@uri.edu

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/hmbr.

can be understood through the ergodic theorems (Choe, 2005; Molenaar, 2008). Equivalent results will only occur if the two conditions specified by the ergodic theorems are met: (1) Each individual trajectory has to obey the same dynamic laws, and (2) Each individual trajectory must have equal mean levels and serial dependencies. If these conditions are not met, then results from nomothetic analyses will not capture the processes of the individuals that make up a sample. Inappropriately inferring from a group to an individual is known as an ecological fallacy, and is a common issue with nomothetic methods.

The ergodic theorems are based on a general theory about the relationships between *inter*system and *intra*-system variability (where a system can be any unit: person, family, school, etc.). Ergodic theory addresses the relationships between individual units and groups of those units in the most general setting possible, namely for all measurable processes and using different metrics. A very special case is the set of Gaussian processes. Ergodicity for Gaussian processes is associated with the two conditions specified by the ergodic theorems. While these two criteria are sufficient for Gaussian processes, they are necessary (but not sufficient) for non-Gaussian processes. These fundamental issues are rarely evaluated by researchers, often due to too few data points (2–3 per unit is common and 8–10 in single-subject research).

Visual Analysis

Visual analysis (VA) is a descriptive method, widely used in applied behavior analysis research. The most basic experimental model used in single-case design is an AB model with a well-defined target behavior that is examined before and after the intervention. The first phase (A) of the design consists of multiple baseline observations that assess the pre-intervention characteristics of the behavior. In the second phase (B) of the design, the treatment component of the experiment is introduced and changes in behavior are examined (Barlow et al., 2009; Kazdin, 2011; Parsonson, & Baer, 1978). The most common form of AB design is a multiple baseline design (Shadish & Sullivan, 2011) where the timing of the intervention is staggered across cases, across dependent variables, or across settings.

The VA of the graph, performed by a judge or a rater, is based on a set of criteria that evaluate and compare the characteristics of Phase A and B and examine whether behavior changes observed in Phase B are caused by the intervention. The baseline (A) phase provides information about the descriptive and predictive aspects of the target behavior, such as stability and variability. Stable behavior, characterized by the absence of a trend or slope in the data, indicates that the targeted behavior neither increases nor decreases on average over time during the baseline phase (Kazdin, 2011). Variability of the data is characterized by the changes in the behavior within the range of possible low and high levels, and it is widely acknowledged that substantial variability of the behavior in the baseline phase can significantly impair the conclusions regarding the effects of the intervention (Barlow et al., 2009). Single-case experiments are evaluated based on magnitude and rate of change between Phase A and B. The magnitude of change is based on variability in level and slope of the data. Changes in level refer to average changes in the frequency of target behavior, whereas changes in slope refer to shifts in direction of the behavior across different phases. The mean is the average for all data in a particular phase. If the series is stable, the level will equal the slope. Changes in level and slope are independent from each other. Rate of change is based on changes in trend or slope of the data and latency of change. Trend analysis provides information on systematic increases or decreases in the behavior across phases, whereas latency of change refers to the amount of time between the termination of one phase and changes in behavior (Kazdin, 2011).

Although the above criteria are well established in the literature, they are rarely used in practice. Often, conclusions regarding the intervention effects instead of being based on a systematic and criterion based review, they are driven by a researcher's subjective evaluation. Applied behavior analysis researchers argue that large intervention effects are evident and provide unequivocal conclusions that can be easily observed by independent judges. Further, they state that the subjective evaluation of intervention effects has a minimal impact on reliability and validity of the conclusions drawn from the graphs presenting large and therefore easily observable treatment effects, because only those are considered to have significant clinical implications (Baer, 1977; Kazdin, 2011).

Proponents of VA acknowledge that certain characteristics of the data can significantly impair the ability to accurately evaluate intervention effects. The presence of slope in the baseline phase of the experiment may negatively affect the evaluation of the experiment, especially when the trend of the targeted behavior is moving in the same direction as potential treatment effects. High variability of the data may also interfere with the validity of the conclusions. However, advocates of this method state that the conservative approach to evaluating intervention effects guarantees highly accurate and consistent conclusions across independent judges, as well as reduces unknown probability of Type I error rate and consequently increases the probability of Type II error rate (Baer, 1977; Kazdin, 2011).

In the recent literature, some of the VA supporters have discussed the problem of the lack of effect size estimation, which results in an inability to perform meta-analytic reviews of single-case experiments. As stated by Kazdin (2011), the single-case research field would benefit from the ability to integrate a large number of studies in a systematic way that would allow drawing broader conclusions. However, to date there is no consensus regarding guidelines for interpreting effect sizes calculated based on methods that supplement VA. Brossart, Parker, Olson, and Mahadevan (2006) compared five analytic techniques frequently used in single-case research by applying them to the same data. They concluded

that each analytical approach was strongly influenced by serial dependency, and the obtained results based on each method varied so much that it prohibited the development of any reliable effect size interpretation guidelines. A noteworthy study by Hedges, Pustejovsky, and Shadish (2012) proposed a new effect size that is comparable to Cohen's *d*, frequently used in group-level designs. It assumes normality and no trend in the data, and it can be applied across studies with at least three independent cases. This new approach can be applied in meta-analytic research and warrants further examination.

Several studies examined agreement rates among judges and showed that VA led to inconsistent conclusions about the intervention effects across different raters. The inter-rater agreement among judges who reviewed the same graphs was relatively poor, ranging on average from .39 to .61 (Jones, Weinrott, & Vaught, 1978; DeProspero & Cohen, 1979; Ottenbacher, 1990). Higher complexity of the data and experimental design resulted in less consistent conclusions. Factors like high variability of the data, inconsistent patterns of behavior over time, changes in slope, and small changes in level of the data were associated with lower agreement rates across judges (DeProsper & Cohen, 1979; Ottenbacher, 1990). One study by Jones et al. (1978) showed that the highest level of agreement between the two methods was found when there were non-statistically significant changes in the behavior, and the lowest agreement occurred when there were significant effects of the intervention. In addition, a number of studies demonstrated that higher levels of serial dependency in the data lead to higher rates of disagreement between visual and statistical analysis (Bengali & Ottenbacher, 1998; Jones et al., 1978; Matyas & Greenwood, 1990). Particularly, Matyas and Greenwood (1990) showed that positive autocorrelation and high variability in the data tend to increase Type I error rates. Overall, the above findings suggest that advantages of the conservative approach of VA are overstated and do not guarantee the reduction of Type I error rate. In addition, the effects of high autocorrelation on single-case data have been shown to negatively impact other analytical techniques such as inferential precision (Smith, Borckardt, & Nash, 2012) and effect size estimation (Manolov & Solanas, 2008).

Interrupted Time Series Analysis

Interrupted time series analysis (ITSA)[1] is a statistical method used to examine intervention effects of single-case study designs. It was initially developed by Box and Tiao (1965; Box & Jenkins, 1976) and introduced to the behavioral sciences by Glass, Willson, and Gottman (1975/2008). Although ITSA is widely used in areas such as econometrics, it has not reached saturation in the behavioral and social sciences to the same degree where there is little consensus on the appropriate method. Other methods that have been proposed for the same task, including multiple regression (e.g., Huitema, 2011; Maggin et al., 2011), multilevel modeling (e.g., Van den Noortgate & Onghena, 2003a, 2003b, 2007, 2008), and the overlap statistics proposed by Parker and others (e.g., Parker et al., 2011). However, the autocorrelation structure of the data is sometimes ignored. For example, multiple regression is a special case of time series analysis when the autocorrelations are all equal to zero. As noted below, having all autocorrelations equal to zero is unlikely to occur, and ignoring the dependency in the data can lead to very inaccurate parameter estimates. An important general problem is that these methods do not directly address violations of the Ergodic Theorems (Molenaar, 2007; Molenaar, 2008; Velicer, Babbin, & Palumbo, 2014). The Ergodic Theorems represent a critical distinction between nomothetic and idiographic approaches and must be addressed before combining multiple idiographic studies.

An inherent property of time series data is serial dependency that reflects the impact of previous observations on the current observation and violates the assumption of independence of errors, which can significantly affect the validity of the statistical test. Serial dependency, examined by the magnitude and direction of autocorrelations between observations spaced at different time intervals (lags), directly impacts error term estimation and validity of the statistical test. Negative autocorrelations produce an overestimation of the error variance, which leads to conservative bias and increases Type II error rate, whereas positive autocorrelations lead to underestimation of the error variance, and cause liberal bias and increase Type I error rates (see Velicer & Molenaar, 2013, for an illustration).

Time series analysis may be expressed as a generalized least squares problem, i.e.,

$$\mathbf{b} = (\mathbf{X}'\mathbf{T}'\mathbf{T}\mathbf{X})^{-1}\mathbf{X}'\mathbf{T}'\mathbf{T}\mathbf{Z} = (\mathbf{X}^{*'}\mathbf{X}^{*})^{-1}\mathbf{X}^{*'}\mathbf{Y} \qquad (1)$$

where the parameters of interest are contained in the vector **b**, X is a design matrix, Z is the vector of observed data, and **T** is a lower triangular transformation matrix where the dependency is removed from the data. For an interrupted time series analysis, there are typically four parameters of interest in the **b** vector, the Level of the series (L), the Slope of the series (S), the Change in Level (DL), and the Change in Slope (DS). The slope parameters represent one of the other unique characteristics of a longitudinal design, the pattern of change over time. Investigating the pattern of change over time represents one of the real advantages of employing a longitudinal design. If the transformation matrix $\mathbf{T} = \mathbf{I}$, the identity matrix, there is no dependency in the data, and the parameter estimates are provided by the standard general linear model.

[1]ITSA is a term descriptive of a method of analyzing idiographic data widely used in many disciplines. It should not be confused with a computer program ITSE developed by Williams and Gottman (1982) which was shown to be inaccurate (Harrop & Velicer, 1990) or the later version ITSACORR (Crosbie, 1993) which is also fatally flawed (Huitema, Bradley, McKean, & Laraway, 2007).

The ITSA method is able to measure the degree of the serial dependency in the data and statistically remove it from the series, allowing for an unbiased estimate of the changes in level and trend across different phases of the experiment (Glass et al., 1975/2008). In addition, after accounting for serial dependency in the data, ITSA facilitates an estimate of the single-case effect size, by accounting for a within-individual variance and evaluating the differences between experimental phases of the study. This type of effect size is similar to group-level Cohen's *d* effect size (Cohen, 1988), which is the most commonly used measure of intervention effects in behavioral sciences research with widely implemented interpretative guidelines.

ITSA Model Identification

Identification of the correct autoregressive moving average model (ARIMA), i.e., determining the specific transformation matrix **T**, is an essential element of ITSA, because model identification, as well as sample size, directly impact the accuracy of the parameter estimation. Proposed by Glass et al. (1975/2008) method for ARIMA estimation is computationally very complex, therefore not accessible to the average researcher, and it requires a large number of observation (minimum 100 data points). Nevertheless, Velicer and Harrop (1983) showed that identifying the correct ARIMA is often unreliable, even with the recommended number of data points, leading to model misidentification. To address the limitations of the Glass et al. (1975/2008) method, the general transformation model that does not require specification of a particular model, was proposed (Velicer & McDonald, 1984). While the Glass et al. (1975/2008) method requires a two-step approach: (1) identification of the ARIMA (*p, d, q*), which requires large number of data points and has been shown to be unreliable, and (2) estimation of the parameters after correct transformation of the data, the general transformation method replaces the specific transformation matrix by a generalized transformation matrix and avoids the problematic model identification step (Velicer & McDonald, 1984). Harrop and Velicer (1985) compared the results of ITSA using: (1) the model developed by Glass et al. (1975/2008); (2) a priori specified (1, 0, 0) model proposed by Simonton (1977); and (3) an assumed (3, 0, 0) model as an approximation to the recommended (5, 0, 0) general transformation model. The findings led to the conclusion that the model identification step can be eliminated and replaced with the assumed (1, 0, 0) and general transformation model, even for time series data with as few as 40 data points. However, the general transformation model instead of the assumed (1, 0, 0) model is recommended for higher order models.

ITSA Limitations

Although the accuracy of the assumed (1, 0, 0) model and the general transformation model has been shown for data with at least 40 observation, it has not been tested on a very short time series with less than 40 observations, which are very common in single-case study designs. Shadish and Sullivan (2011) found that among single-case studies, the median number of observations is 20, and 90.6% have less than 50 data points. Therefore, the accuracy of the parameter estimation based on time series with less than 40 observations should be considered cautiously. Another limitation is that information about the outcome distribution of the measures is typically lacking, and there may be more appropriate methods for alternative distributions. For example, a Poisson distribution is usually used to model counts, and that type of data is common in JABA reports. Alternative analytic methods based on the Poisson distribution are under development (Jazi, Jones & Lai, 2012) and represent an alternative choice for future analysis of these types of data.

Study Aims

The aim of this study is to examine the level of agreement between the conclusions drawn from VA of graphically presented data with the findings based on ITSA of the same data. The study uses graphical data based on single-case studies published in the *Journal of Applied Behavior Analysis* in 2010. This journal was selected because it is a leading journal on the topic used by applied researchers, and it strongly promotes the use of VA rather than quantitative analysis methods (Shadish & Sullivan, 2011; Smith, 2012). In a related study, all the studies published in a leading textbook (Kazdin, 2011) were evaluated in the same way (Harrington & Velicer, 2015). The study will also provide estimates of the degree of autocorrelation and estimate the effect size for each study.

METHOD

Sample

Graphical data was obtained from the research papers published in the *Journal of Applied Behavior Analysis* (JABA) in 2010. For a graph to be included in this study, it was required to meet the following inclusion criteria: (1) present actual data (not simulated), (2) present interrupted time series data, (3) present a minimum of three observations in each phase of the design in order to estimate a full four-parameter model, (4) present baseline and treatment phases of an experimental design, (5) include corresponding description of the conclusions drawn from the VA of the graph, and (6) present well-defined data points (observations) in the graph. Any study design (e.g., ABAB, ABCA, etc.) presenting graphs that met the above eligibility criteria was included in the study. Graphs presenting cumulative data or alternating-treatment designs were not eligible.

Procedure

Eligible graphs were scanned and electronically imported into UnGraph® software version 5.0 (Biosoft, 2004), and the data was extracted using the UnGraph® software's function of a coordinate system. Then, sequentially ordered data recorded in a time series data format was exported into a Microsoft Excel® spreadsheet.

Validity and Reliability of UnGraph® Software

UnGraph® software has been previously examined for its validity and reliability when extracting data from graphs representing single-case designs (Shadish et al., 2009). Results of this study indicated high validity and reliability of the extracted data from graphs, with .96 as an average correlation coefficient between two raters.

Analysis

ITSA was used to evaluate intervention effects of each single-case study. ITSA parameters were estimated using the assumed ARIMA (1, 0, 0) (Simonton, 1977) and the general transformation ARIMA (5, 0, 0) (Velicer & McDonald, 1984; Harrop & Velicer, 1985), that do not require the model identification step. First, the assumed ARIMA (1, 0, 0) (Simonton, 1977) was applied and a chi-square test for the residuals was used to examine whether the residuals were uncorrelated ("white noise") or contained additional information that required a higher order model (Glass et al., 1975/2008). If the residuals were correlated, then the general transformation ARIMA (5, 0, 0) (Velicer & McDonald, 1984; Harrop & Velicer, 1985) was applied.

Once the best fitting ARIMA was identified, parameters such as trend, change in trend, level, change in level, and mean and variability of the data series were evaluated. Intervention effects were examined based on changes in slope and level across the experimental phases of the design. An effect size similar to Cohen's *d* ($d = \Delta$ Level /*s*), where the numerator represents change in level at the interruption point and the denominator represents within-case standard deviation, was calculated to examine the magnitude of the behavior change due to the intervention. The measure of effect size based on within-case standard deviation is expected to be inflated relative to between groups Cohen's *d*. An effect size was calculated only for studies where no significant slope or change in slope was present. Analyses were performed in SAS version 9.2.

Description of the VA of the graphs presented in the publications published in JABA was used to perform the comparison of the findings. These comparisons were based on conclusions made in regards to trend, change in trend, variability, and change in level across different experimental phases of the experiment.

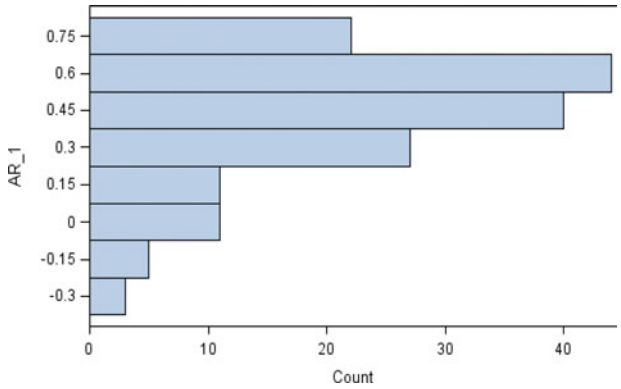


FIGURE 1 Distribution of lag-1 autoregressive coefficients in eligible time series data ($K = 163$).

Illustration of ITSE Application and VA Conclusions

To illustrate the application of the ITSA method in the analysis of single-case studies and comparison with the conclusions drawn on VA, three examples were selected from the experiments presented in Table 1.

Example 1

The first example is based on a study that examined the effects of providing praise and preferred edible items based on variable-time schedule in order to reduce problem behavior. In addition, effects of variable-time schedule on compliance were also evaluated. The study was based on a reversal design (ABAB) and included three participants (Lomas, Fisher, & Kelly, 2010). In the current example, data for one of the participants is provided. Sam was an 8-year-old boy diagnosed with Asperger syndrome and attention deficit hyperactivity disorder. Data displaying frequency of problem behavior and percentage of compliance in each phase of the design are presented in Figure 4 and Figure 5, respectively. Conclusions based on VA of the data suggested that the variable-

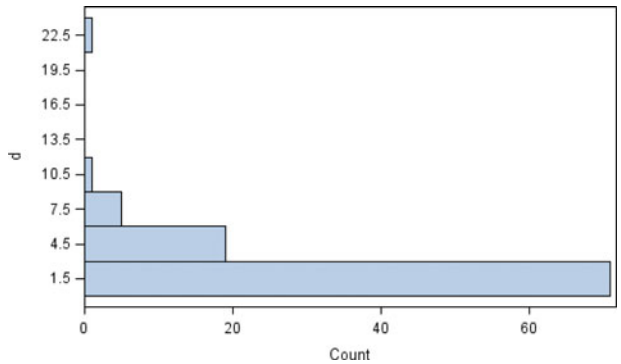


FIGURE 2 Distribution of the Cohen's d effect size estimates for eligible time series data ($k = 98$). *Note.* r_1 = percent of cases where first autocorrelations were greater than .40; m_d = mean *d* (effect size estimate).

TABLE 1
Summary of Visual Analysis and Interrupted Time Series Analysis (ITSA) Based on Eligible Studies Published in the *Journal of Applied Behavior Analysis* in 2010

Figure	*N* BL	*N* TX	ARIMA	AR 1	Level	Error σ	Slope	Δ Slope	Δ Level	*d*
St. Peter Pipkin, Vollmer, and Sloman (2010)										
Figure 6. Top panel (ABCDEFBFEDC)										
First sequence of conditions										
"DRA lost its efficacy when implemented at less than 50% integrity with combined omission and commission errors" (p. 60).										
On task (B (EF))	9	17	(1, 0, 0)	.48*	−11.22	27.58	4.16*	−5.52*	−1.04	—
Off task (B (EF))	9	17	(1, 0, 0)	.51*	102.47*	30.37	−3.55*	5.01*	0.46	—
Second sequence of conditions										
"The condition sequence did not influence Helena's behavior strongly during the integrity failure phases, insofar as her behavior during the replications matched the results obtained from the initial exposures" (p. 62).										
On task (B (FE))	5	16	(1, 0, 0)	.29	−28.61	37.38	2.33*	−2.41*	−1.88	—
Off task (B (FE))	5	16	(1, 0, 0)	.31	123.61*	37.23	−2.33*	2.41*	2.01	—
Figure 7. Bottom panel (ABABCACBABCACAB)										
"During subsequent DRA phases that followed baseline, aggression decreased to low rates. .., and appropriate behavior increased to moderate rates. .." (p. 65).										
Aggression (ABABABAB)[c]	**10/5/13/8**	**22/12/10/46**	**(5, 0, 0)**	**.73***	**8.87***	**3.50**	**1.08**	**−1.88**	**−1.38**	**0.83**
Appropriate behavior (ABABABAB)[d]	**10/5/13/8**	**22/12/10/46**	**(5, 0, 0)**	**.78***	**0.39**	**1.28**	**0.66**	**1.84**	**−1.13**	**0.86**
"During the 50% integrity phases that followed DRA, a mixture of aggression and greetings occurred, with some bias toward aggression" (p. 65).										
Aggression (BCBC)	12/10	11/36	(5, 0, 0)	.45*	2.42*	2.24	0.49	−0.24	2.67*	1.14
Appropriate behavior (BCBC)[b]	12/10	11/36	(5, 0, 0)	.38*	5.22*	1.22	0.37	−1.90	−0.03	0.02
"During the integrity failures following baseline, rates of greetings remained low or near zero . . ., and rates of aggression remained high and stable. .." (p. 65).										
Appropriate behavior (ACAC)[a]	11/18	8/5	(5, 0, 0)	.70*	3.30*	0.89	−3.13*	0.27	−0.35	—
Aggression (ACAC)	11/18	8/5	(5, 0, 0)	.25	9.97*	2.82	−0.27	−0.28	1.59	1.30
Lee, Yu, Martin, and Martin (2010)										
Figure 1. (ABAB)										
"For all stimuli, higher rates of responding were observed in the reinforcement condition than in baseline" (p. 97).										
Lynn										
Goldfish Crackers (ABAB)	3/3	3/3	(1, 0, 0)	.45	2.27	1.75	−0.55	1.24	8.58*	8.77
Pretzel (ABAB)[†]	4/3	4/3	(1, 0, 0)	.57*	1.37	1.41	−1.29	4.61*	5.53*	—
Popcorn Twist (ABAB)	3/8	4/7	(1, 0, 0)	.50*	1.56	2.12	0.51	2.13*	0.48	**—**
Cereal (ABAB)	3/4	8/3	(1, 0, 0)	.03	1.30	1.71	0.18	−1.28	2.92*	2.01
Jell-O (ABAB)	3/3	3/3	(1, 0, 0)	.50	1.24*	0.60	−0.19	−1.79	2.60*	3.53
James										
Orange Juice (ABAB)	6/3	5/5	(1, 0, 0)	.22	0.34	1.81	0.39	−0.26	3.95*	2.22
Smarties (ABAB)	9/3	3/5	(1, 0, 0)	.24	1.92*	1.08	−1.66	−0.19	3.91*	2.82
Pretzel (ABAB)	**3/3**	**9/4**	**(1, 0, 0)**	**.08**	**1.51***	**1.11**	**−1.14**	**1.47**	**0.93**	**0.62**
Mini Cookies (ABAB)	4/3	4/7	(1, 0, 0)	−.11	1.28*	0.76	−1.12	2.22*	−0.46	—
Apple Sauce (ABAB)	**3/7**	**3/6**	**(1, 0, 0)**	**−.09**	**1.67***	**0.85**	**−1.14**	**1.75**	**−1.04**	**0.80**
Popcorn Twist (ABAB)	3/3	8/3	(1, 0, 0)	−.16	1.03*	0.42	−1.51	−0.73	2.77*	1.67
Groskreutz, Karsina, Miguel, and Groskreutz (2010)										
Figure 1. (ABC)										
"Posttest performances indicated conditional relations were evident for all stimuli tested. .." (p. 134).										
Lyle (AC)	**4**	**4**	**(1, 0, 0)**	**.54***	**−6.91**	**7.06**	**8.77***	**−5.08***	**1.63**	**—**
Derrick (AC)	6	6	(1, 0, 0)	.71*	15.83*	4.89	−2.88*	−0.00	9.15*	—
Roy (AC)	6	6	(1, 0, 0)	.73*	35.60*	8.40	−2.83*	2.66*	9.35*	—
Keith (AC)	6	6	(1, 0, 0)	.62*	5.92	12.46	0.39	−0.55	5.46*	6.74
Waller and Higbee (2010)										
Figure 1. (ABAB)										
"Brent's disruption rapidly decreased when treatment was introduced. .. " (p. 152).										
Brent: disruption (ABAB)[a]	12/3	12/30	(5, 0, 0)	.58*	50.99*	14.62	−0.67	0.42	−6.52*	2.68
"David's disruption decreased to low levels. .. during treatment" (p. 152).										
David: disruption (ABAB)[a]	7/3	12/21	(1, 0, 0)	.58*	23.42*	5.55	3.07*	−3.11*	−8.26*	—
David: academic beh. (ABAB)[a]	7/3	12/21	—	.67*			Model did not converge			

(Continued on next page)

TABLE 1
Summary of Visual Analysis and Interrupted Time Series Analysis (ITSA) Based on Eligible Studies Published in the *Journal of Applied Behavior Analysis* in 2010 *(Continued)*

Figure	*N* BL	*N* TX	ARIMA	AR 1	Level	Error σ	Slope	Δ Slope	Δ Level	*d*
Toussaint and Tiger (2010)										
Figure 1. (AC)										
"Correct responding. .. increased and was maintained after instruction for the AB relation. .. during postinstruction probes" (p. 187).										
Fred: (CA) set 1	4	7	(1, 0, 0)	.65*	82.52*	7.60	−4.22*	5.99*	4.24*	—
Fred: (CA) set 2	5	6	—	.69*			Model did not converge			
Fred: (CA) set 3	**6**	**5**	**(1, 0, 0)**	**.14**	**52.49***	**25.58**	**−0.70**	**1.73**	**0.42**	**0.47**
Fred: (CA) set 4	**7**	**4**	**(1, 0, 0)**	**.28**	**38.60**	**24.12**	**0.60**	**0.11**	**0.81**	**1.13**
". . . correct responding. .. increased to and was maintained at high levels following the AB instruction. .." (p. 187)										
Fred: (AC) set 1	4	6	(1, 0, 0)	.51*	11.21	14.60	1.42	−1.60	3.62*	3.62
Fred: (AC) set	5	5	—	.70*			Model did not converge			
Fred: (AC) set 3	6	4	(1, 0, 0)	.65*	7.67	11.43	−0.65	2.09	4.75*	5.74
Fred: (AC) set 4	7	3	(1, 0, 0)	.62	5.02	8.52	0.75	−0.57	7.24*	11.11
Figure 2. (AC)										
"For the BA relation, mean correct responding. .. increased. .. following AB instruction. .." (p. 187).										
Jeremy: (BA) set 1	3	5	(1, 0, 0)	.70*	60.87*	8.29	−2.63	3.43*	6.37*	—
Jeremy: (BA) set 2	4	4	—	.42			Model did not converge			
"For the CA relation, mean correct responding. .. increased. .. following AB training. .. "(p. 187).										
Jeremy: (CA) set 1	3	5	—	.43			Model did not converge			
Jeremy: (CA) set 2	4	4	—	.45			Model did not converge			
Figure 3. (AC)										
". . . correct responding. .. increased to high levels. .. after AB instruction" (p. 190).										
Danielle: (BA) set 1	3	6	—	.61			Model did not converge			
Danielle: (BA) set 2	**5**	**4**	**(1, 0, 0)**	**.67***	**49.67***	**7.18**	**3.16***	**−0.04**	**0.18**	**—**
"Mean correct responding for the CA transitive relation increased. .. after instruction" (p. 190).										
Danielle: (CA) set 1	3	6	(1, 0, 0)	.34	16.17	14.23	−0.02	−0.26	5.54*	6.15
Danielle: (CA) set 2	**5**	**4**	**(1, 0, 0)**	**.65***	**45.75***	**2.99**	**15.49***	**−9.53***	**1.92**	**—**
"Correct responding was low in both letter sets. .. during baseline for the AC relation and increased. .. during postinstruction probes" (p. 190).										
Danielle: (AC) set 1	3	6	(1, 0, 0)	.27	19.62	15.98	0.02	−0.18	4.46*	5.06
Danielle: (AC) set 2	**5**	**4**	**(2, 0, 0)**	**.50**	**1.18***	**12.36**	**3.32***	**−2.21**	**2.66**	**—**
Kuhn, Chirighin, and Zelenka (2010)										
Figure 2. (ABAB)										
"After the introduction (and reintroduction) of FCT+ EXT, immediate reductions in problem behavior were observed for Angela and Greg. .." (p. 256).										
Angela (ABAB)	**4/6**	**6/6**	**(1, 0, 0)**	**.23**	**5.12***	**2.92**	**0.54**	**−0.82**	**−1.66**	**1.28**
Greg (ABAB)[a]	**3/7**	**7/7**	**(1, 0, 0)**	**.01**	**2.27***	**1.20**	**0.55**	**−1.23**	**−2.02**	**1.20**
Figure 4. (ABC)										
"After the introduction of the DFCT contingency, head banging increased slightly for Angela in Pairs 1 and 2. .., whereas problem behavior persisted at low levels for Greg in both Pair 1 and Pair 2. .. " (p. 259).										
Angela: Pair 1 (AB)	6	19	(1, 0, 0)	.18	0.00	1.60	0.08	−0.06	1.12	0.98
Angela: Pair 2 (AB)	**12**	**28**	**(5, 0, 0)**	**.25**	**2.79***	**1.56**	**−2.29***	**1.46**	**4.02***	**—**
Greg: Pair 1 (AB)	3	21	(1, 0, 0)	−.23	−0.01	0.73	0.16	0.18	0.69	0.66
Greg: Pair 2 (AB)	10	21	(1, 0, 0)	−.15	−0.06	1.12	0.19	−0.68	1.67	1.04
"When the therapist provided Angela with noncontingent access to preferred toys (i.e., bumble ball, massager), head banging decreased to near-zero levels across both pairs. .." (p. 259).										
Angela: Pair 1 (BC)	19	23	(5, 0, 0)	.35*	1.43*	1.14	1.54	−3.05*	−2.91*	—
Angela: Pair 2 (BC)	**28**	**8**	**(5, 0, 0)**	**.39***	**3.24***	**1.49**	**−3.08***	**0.41**	**−0.40**	**—**
"In addition, as shown in the DFCT with observing behavior condition (Figure 4, bottom two panels), rates of problem behavior persisted at near-zero levels for Pair 1 and Pair 2 activities. .." (p. 259).										
Greg: Pair 1 (BC)	21	12	(5, 0, 0)	−.07	0.71*	0.62	−0.48	−0.64	−0.25	0.15
Greg: Pair 2 (BC)	21	5	(1, 0, 0)	−.17	1.30*	1.25	−1.42	−0.10	0.18	0.18
Digennaro-Reed, Codding, Catania, and Maguire (2010)										
Figure 1. (ABC)										
"Percentage correct increased immediately during the IVM condition for all participants. .." (p. 295).										
Kelly (AB)[†]	3	5	(1, 0, 0)	.52	28.28	16.74	0.83	−0.46	3.67*	3.36
Lauren (AB)	5	7	(1, 0, 0)	.60*	44.83*	8.34	0.65	−0.86	5.49*	4.89
Shannon (AB)	**7**	**6**	**(1, 0, 0)**	**.25**	**29.44**	**22.56**	**1.26**	**−0.89**	**0.41**	**0.47**

(Continued on next page)

TABLE 1
Summary of Visual Analysis and Interrupted Time Series Analysis (ITSA) Based on Eligible Studies Published in the *Journal of Applied Behavior Analysis* in 2010 *(Continued)*

Figure	*N* BL	*N* TX	ARIMA	AR 1	Level	Error σ	Slope	Δ Slope	Δ Level	*d*
Dolezal and Kurtz (2010)										
Figure 1 Bottom panel (AB)										
"During FCT treatment in the demand and diverted- attention condition, rate of problem behavior decreased. .." (p. 312).										
Problem behavior (AB)	6	8	(1, 0, 0)	.54*	0.81*	0.17	−2.00	1.88	−2.62*	2.29
Van Houten, Malenfant, Reagan, Sifrit, Compton, and Tenenbaum (2010)										
Figure 2. (ABA and ABCA)										
"The top panel shows data from a driver who demonstrated an increase in seat belt use following the 8-s delay and a decline when the delay was removed" (p. 377).										
Top panel (AB)[d]	22	61	(5, 0, 0)	.56*	27.33*	12.74	1.28	−0.89	8.59*	2.78
Top panel (BA)[d]	30	61	(5, 0, 0)	.76*	70.61*	11.54	1.29	−2.19*	−8.58*	—
"The second panel shows the data from a driver who demonstrated an increase following the introduction of the delay and maintenance following its removal. .." (p. 377).										
Second panel (AB)[d]	22	67	(5, 0, 0)	.68*	35.46*	19.18	−1.08	1.32	4.93*	3.14
Second panel (BA)	23	67	(5, 0, 0)	.37*	74.70*	18.39	1.61	−0.23	−0.20	0.13
". . . the third panel shows data from a driver for whom there was no effect when the delay was introduced or increased from 8 to 16 s." (p. 377).										
Third panel (ABCA)[b]	26/27	60/23	—	.50*			Model did not converge			
"The bottom panel shows the data of a participant who initially showed an increase in seat belt use following the introduction of the 8-s delay followed by a gradual decline in seat belt use. After the 16-s fixed delay was introduced, seat belt use improved" (p. 377).										
Bottom panel (AB)[d]	26	60	(5, 0, 0)	.74*	35.72*	18.58	0.27	−3.43*	6.85*	—
Bottom panel (BC)[d]	60	23	(5, 0, 0)	.81*	107.92*	16.74	−13.01*	2.94*	8.05*	—
Lomas, Fisher, and Kelley (2010)										
Figure 2. (ABAB)										
"Variable-time delivery of food and praise. .. greatly reduced problem behavior for all three children. .." (p. 431).										
Sam: problem behavior (ABAB)	7/5	5/5	(1, 0, 0)	.40	2.56*	1.11	−1.54	0.94	−2.39*	1.85
Aaron: problem behavior (ABAB)	6/7	4/4	(1, 0, 0)	.19	4.60*	2.37	−0.43	0.18	−2.37*	1.79
Mark: problem behavior (ABAB)	7/4	5/6	—	.41			Model did not converge			
"Levels of compliance were only slightly higher during treatment with VT food and praise for Sam. .. and Mark. .." (p. 431).										
Sam: compliance (ABAB)	**7/5**	**5/5**	**(1, 0, 0)**	**.13**	**17.02**	**23.31**	**0.63**	**−1.54**	**2.43***	**1.76**
Mark: compliance (ABAB)	7/4	5/6	(1, 0, 0)	.37	37.27*	18.10	0.21	−0.68	1.94	1.93
"Aaron's compliance was maintained at higher and more stable levels during VT food and praise. .." (p. 431).										
Aaron: compliance (ABAB)	**6/7**	**4/4**	**(1, 0, 0)**	**.23**	**36.03**	**31.87**	**−0.43**	**0.56**	**1.02**	**0.88**
Stokes, Luiselli, Reed, and Fleming (2010)										
Figure 1. (ABC)										
"Descriptive feedback alone did not improve pass blocking" (p. 469).										
Dan (AB)[†]	5	3	(1, 0, 0)	.25	40.38*	12.53	−1.19	2.22	0.85	0.76
Steve (AB)	6	3	(1, 0, 0)	−.32	49.88*	4.84	−0.92	0.51	0.38	0.67
Logan (AB)	**7**	**5**	**(1, 0, 0)**	**.34**	**38.33***	**3.17**	**5.27***	**−7.28***	**7.59***	**—**
Matt (AB)	**9**	**7**	**(1, 0, 0)**	**.52***	**65.22***	**6.05**	**−2.25***	**2.05**	**3.10***	**—**
Russ (AB)	12	7	(1, 0, 0)	.18	32.51*	10.23	0.79	−1.02	0.52	0.63
"The descriptive and video feedback condition was demonstrated to be effective in improving correct pass blocking for all five participants" (p. 469).										
Dan (AC)	5	6	(1, 0, 0)	.65*	45.97*	6.91	−0.96	2.92*	2.76*	—
Steve (AC)	6	7	(1, 0, 0)	.80*	50.67*	6.73	−0.66	1.96	3.61*	4.50
Logan (AC)	7	4	(1, 0, 0)	.72*	39.27*	5.44	2.09	−0.03	5.63*	6.79
Matt (AC)	9	7	(1, 0, 0)	.73*	64.89*	5.13	−2.63*	2.48*	6.93*	—
Russ (AC)	**12**	**5**	**(1, 0, 0)**	**.50***	**33.19***	**10.42**	**0.73**	**0.48**	**1.21**	**1.62**
"Video feedback combined with descriptive feedback was consistently superior to descriptive feedback alone in improving pass blocking" (p. 469).										
Dan (BC)	**3**	**6**	**(1, 0, 0)**	**.41***	**−54.43**	**6.57**	**3.63***	**−2.55**	**−0.14**	**—**
Steve (BC)[†]	3	7	(1, 0, 0)	.76*	50.01*	7.56	0.57	0.28	2.65*	2.65
Logan (BC)	5	4	(1, 0, 0)	.58*	72.94*	5.27	−3.61*	2.36	6.30*	—
Matt (BC)	**7**	**7**	**(1, 0, 0)**	**.36**	**67.99***	**6.53**	**0.88**	**0.04**	**0.91**	**0.92**
Russ (BC)	**7**	**5**	**(1, 0, 0)**	**.59***	**46.73**	**5.99**	**−1.04**	**1.58**	**2.24**	**3.67**

(Continued on next page)

TABLE 1
Summary of Visual Analysis and Interrupted Time Series Analysis (ITSA) Based on Eligible Studies Published in the *Journal of Applied Behavior Analysis* in 2010 *(Continued)*

Figure	*N* BL	*N* TX	ARIMA	AR 1	Level	Error σ	Slope	Δ Slope	Δ Level	*d*
Stokes, Luiselli, and Reed (2010)										
Figure 1. (AB)										
"His correct tackling also increased with intervention. .. (p. 511).										
Mike (AB)[a]	12	10	(1, 0, 0)	.74*	29.01*	7.22	0.08	4.21*	0.68	—
Falcomata, Roane, Feeney, and Stephenson (2010)										
Figure 1. Top panel (ABAB)										
"Rates of elopement were elevated during the free-access condition. .. relative to rates during the blocking condition" (p. 515).										
Elopement (ABAB)	3/8	5/12	(1, 0, 0)	.31*	1.50*	0.43	−0.65	1.00	−3.66*	2.63
Raiff and Dallery (2010)										
Figure 1. (ABA)										
"When the intervention was introduced, an increase in the frequency of testing occurred" (p. 489).										
Talia (AB)	5	5	(1, 0, 0)	.55	0.80	1.22	1.47	−1.33	3.41*	3.98
Bonita (AB)	5	5	(1, 0, 0)	.30	1.45	1.33	0.65	−0.69	2.61*	2.68
Edward (AB)	**5**	**5**	**(1, 0, 0)**	**.27**	**2.78**	**0.28**	**2.23**	**−1.26**	**1.12**	**1.21**
Andrea (AB)	**5**	**5**	**(1, 0, 0)**	**.51**	**1.02**	**1.30**	**0.10**	**1.80**	**1.78**	**2.00**
"Removing the intervention resulted in a decrease in the frequency of testing. .." (p. 489).										
Talia (BA)	**5**	**5**	**(1, 0, 0)**	**.51**	**6.88**	**1.38**	**−0.20**	**−0.94**	**−0.73**	**1.02**
Bonita (BA)	**5**	**5**	**(1, 0, 0)**	**−.05**	**5.84**	**1.21**	**−0.01**	**−0.39**	**−1.58**	**1.67**
Edward (BA)	5	5	(1, 0, 0)	.35	3.98*	0.22	0.04	4.15*	−5.44*	—
Andrea (BA)	5	5	(1, 0, 0)	.51	0.35	0.96	2.80*	−2.46	−3.99*	—
Leon, Hausman, Kahng, and Becraft (2010)										
Figure 1. (ABCD)										
"The implementation of differential reinforcement during nonbusy activities resulted in an increase in appropriate responding during nonbusy activities in Pair 1. .." (p. 527).										
Pair 1: Communication (ABCD)[c]	**4**	**50/10/21**	**(3, 0, 0)**	**.41***	**73.84***	**20.55**	**−0.25**	**0.31**	**−0.34**	**0.31**
Carter (2010)										
Figure 1. Middle panel (ABABC)										
". . . presentation of a high-preference edible item contingent on compliance increased compliance and reduced destructive behavior. .." (p. 545).										
Compliance (ABAB)	5/4	8/3	(1, 0, 0)	.54*	33.44*	7.92	−0.61	1.57	5.16*	3.51
Destructive behavior (ABAB)	5/4	8/3	(1, 0, 0)	.47*	52.38*	9.25	0.38	−3.61*	−1.81	—
". . . the provision of a 30-s break from the tasks for both compliance and destructive behavior produced levels of responding similar to those observed during baseline. . ." (p. 545).										
Compliance (AAC)	**5/4**	**6**	**(1, 0, 0)**	**.14**	**33.58***	**8.37**	**−0.61**	**−2.46***	**3.90***	**—**
Destructive behavior (AAC)	5/4	6	(1, 0, 0)	.46	53.12*	8.01	0.15	−0.06	−1.94	2.24
Grauvogel-Macaleese and Wallace (2010)										
Figure 2.										
"When peers implemented differential reinforcement, off-task behavior immediately decreased for all three participants. .." (p. 549).										
Scott (ABAB)	3/3	7/4	(1, 0, 0)	.45*	63.44*	14.10	2.40*	−1.75	−4.54*	—
Zane (AB)	5	9	(1, 0, 0)	.55*	46.43*	12.55	3.10*	−3.55*	−6.83*	—
Drew (AB)	7	12	(1, 0, 0)	.76*	51.01*	11.15	1.18	−0.99	−5.53*	5.28
Athens and Vollmer (2010)										
Figure 3.										
". . . for both participants, the relative rates of problem behavior and appropriate behavior were sensitive to the reinforcement duration. .." (p. 578).										
Justin (ABCACA)										
Problem behavior (ACACA)[b]	**4/10/14**	**14/20**	**(5, 0, 0)**	**.58***	**2.40***	**0.80**	**−3.04***	**0.17**	**−0.96**	**—**
Compliance (ACACA)	**4/10/14**	**14/20**	**(5, 0, 0)**	**.30***	**0.24**	**0.64**	**1.93**	**0.56**	**−0.30**	**0.17**
Lana (ABAB)										
Problem behavior (ABAB)[a]	**6/9**	**13/11**	**(5, 0, 0)**	**.68***	**1.67***	**0.68**	**0.12**	**−0.58**	**−0.77**	**0.74**
Mand (ABAB)[b]	**6/9**	**13/11**	**(5, 0, 0)**	**.76***	**0.11**	**0.45**	**1.90**	**−0.43**	**0.47**	**0.51**

(Continued on next page)

TABLE 1
Summary of Visual Analysis and Interrupted Time Series Analysis (ITSA) Based on Eligible Studies Published in the *Journal of Applied Behavior Analysis* in 2010 *(Continued)*

Figure	*N* BL	*N* TX	ARIMA	AR 1	Level	Error σ	Slope	Δ Slope	Δ Level	*d*
Figure 4.										
Justin (ABCAC)										
"In the 1 HQ/1 LQ condition, rates of problem behavior decreased, and appropriate behavior increased" (p. 579).										
Problem behavior (AB)	**5**	**14**	**(1, 0, 0)**	**−.05**	**5.31***	**1.46**	**−2.02**	**2.18***	**−0.46**	**—**
Compliance (AB)	5	14	(1, 0, 0)	.13	3.39*	1.21	−2.22*	1.52	4.24*	—
"Problem behavior decreased, and appropriate behavior increased to high levels during the return to the 3 HQ/1 LQ condition" (p. 580).										
Problem behavior (AC)	**9**	**7**	**(1, 0, 0)**	**−.29**	**2.16***	**0.70**	**−3.46***	**−0.51**	**2.21***	**—**
Compliance (AC)	**9**	**7**	**(1, 0, 0)**	**.06**	**5.23***	**1.34**	**−0.78**	**1.61**	**−1.52**	**1.35**
"In summary, results of the quality analyses indicated that. .. the relative rates of both problem behavior and appropriate behavior were sensitive to the quality of reinforcement available for each alternative" (p. 581).										
Problem behavior (ABCAC)	5/9	14/10/7	(1, 0, 0)	−.07	2.92*	1.20	−2.88*	1.41	−2.91*	—
Compliance (ABCAC)[b]	**5/9**	**14/10/7**	**(5, 0, 0)**	**.44***	**1.00**	**1.44**	**3.49***	**−1.98**	**0.70**	**—**
Kenneth (ABABACBC)										
"In the 1 HQ/1 LQ condition, rates of problem behavior decreased, and appropriate behavior increased" (p. 580).										
Problem behavior (AB)	**6**	**15**	**(1, 0, 0)**	**.57***	**4.54***	**1.75**	**0.21**	**−0.71**	**−0.21**	**0.23**
Mand (AB)	**6**	**15**	**(1, 0, 0)**	**.54***	**−0.05**	**0.51**	**0.28**	**1.45**	**−2.14***	**1.67**
". . . we conducted the 3 HQ/1 LQ condition, and problem behavior decreased to rates lower than observed in previous conditions and appropriate behavior increased to high rates" (p. 580).										
Problem behavior (ABABAC)[a]	**6/15/5/4/10**	**19**	**(5, 0, 0)**	**.67***	**4.34***	**1.41**	**−2.21***	**0.32**	**−0.97**	**—**
Mand (ABABAC)[c]	**6/15/5/4/10**	**19**	**(5, 0, 0)**	**.64***	**−0.17**	**0.62**	**5.97***	**−1.14**	**−0.53**	**—**
"In summary, results of the quality analyses indicated that. .. the relative rates of both problem behavior and appropriate behavior were sensitive to the quality of reinforcement available for each alternative" (p. 581).										
Problem behavior (ABABACBC)[c]	6/5/10	15/4/19/8/ 22	(5, 0, 0)	.58*	4.99*	1.54	−1.35	0.94	−2.19*	1.36
Mand (ABABACBC)[d]	**6/5/10**	**15/4/19/8/ 22**	**(5, 0, 0)**	**.57***	**0.17**	**0.72**	**1.59**	**−1.05**	**1.84**	**1.14**
Figure 5.										
Corey (ABCAC)										
"In summary, results of the delay analysis indicate that the relative rates of problem behavior and appropriate behavior were sensitive to the delay to reinforcement following each alternative" (p. 582).										
Problem behavior (ABCAC)	**23/6**	**21/17/44**	**(5, 0, 0)**	**.27***	**3.28***	**2.23**	**−0.14**	**−0.77**	**−0.46**	**0.16**
Mand (ABCAC)[a]	**23/6**	**21/17/44**	**(5, 0, 0)**	**.17**	**1.60***	**1.12**	**−0.14**	**−0.09**	**0.00**	**0.00**
Henry (ABACABAC)										
"In a reversal to 0-s/0-s delay baseline, there was a slight increase in problem behavior from the previous condition and a decrease in appropriate behavior" (p. 582).										
Problem behavior (BA)	**6**	**8**	**(1, 0, 0)**	**.22**	**0.46**	**0.79**	**1.55**	**0.83**	**−1.60**	**1.75**
Mand (BA)	**6**	**8**	**(1, 0, 0)**	**.00**	**1.87***	**0.75**	**−2.45***	**−0.79**	**2.28***	**—**
"During the 0-s/60-s delay condition, there was a decrease in problem behavior to zero rates and an increase in appropriate. .. (p. 582).										
Problem behavior (AB)	6	11	(1, 0, 0)	.59*	0.48	0.75	2.21*	−3.77*	0.00	—
Mand (AB)	6	11	(1, 0, 0)	.47	1.67*	0.64	−2.25*	2.79*	2.25*	—
"In summary, results of the delay analysis indicate that the relative rates of problem behavior and appropriate behavior were sensitive to the delay to reinforcement following each alternative" (p. 582).										
Problem behavior (ABACABAC)	**4/6/12/4**	**8/11/11/16**	**(5, 0, 0)**	**.55***	**1.50***	**0.80**	**−0.20**	**−0.35**	**0.18**	**0.13**
Mand (ABACABAC)	**4/6/12/4**	**8/11/11/16**	**(5, 0, 0)**	**.38***	**0.74**	**0.82**	**−0.60**	**1.14**	**0.63**	**0.35**
Figure 6.										
George (ABAB)										
"In summary, results of the combined analyses indicate that for these participants the relative rates of problem behavior and appropriate behavior were sensitive to a combination of the quality, delay, and duration of reinforcement following each alternative" (p. 584).										
Problem behavior (ABAB)	7/6	7/10	(5, 0, 0)	.45*	3.60*	1.39	−1.89	1.37	−2.63*	1.94
Mand (ABAB)[a]	7/6	7/10	(1, 0, 0)	−.06	0.03	0.47	1.17	0.19	3.92*	1.55
Clark (ABAB)										
"In summary, results of the combined analyses indicate that for these participants the relative rates of problem behavior and appropriate behavior were sensitive to a combination of the quality, delay, and duration of reinforcement following each alternative" (p. 584).										
Problem behavior (ABAB)[a]	6/5	8/10	(1, 0, 0)	.77*	2.91*	0.61	−1.63	1.44	−4.09*	4.07
Mand (ABAB)[a]	6/5	8/10	(1, 0, 0)	.51*	1.13*	0.53	−2.43*	3.79*	0.59	—

(Continued on next page)

TABLE 1
Summary of Visual Analysis and Interrupted Time Series Analysis (ITSA) Based on Eligible Studies Published in the *Journal of Applied Behavior Analysis* in 2010 *(Continued)*

Figure	*N* BL	*N* TX	ARIMA	AR 1	Level	Error σ	Slope	Δ Slope	Δ Level	*d*
Wilder, Allison, Nicholson, Abellon, and Saulnier (2010)										
Figure 1.										
Ricky (ABACABAC)										
"For Ricky, compliance improved when the guided compliance procedure was conducted" (p. 606).										
Compliance (AACAAC)	3/3/3/3	6/8	(1, 0, 0)	.58*	6.91	23.11	−0.77	1.87	3.23*	1.99
Ian (ABACABADAD)										
"For Ian, contingent access to preferred edible items initially appeared to be effective in increasing compliance, but compliance decreased toward the end of this phase" (p. 606).										
Compliance (AC)	**5**	**6**	**(1, 0, 0)**	**.37**	**1.85**	**24.35**	**−0.14**	**−1.21**	**3.71***	**3.78**
"Therefore, a response-cost component was added, which increased compliance to high levels" (p. 607).										
Compliance (ADAD)	7/3	3/5	(1, 0, 0)	.50*	3.47	18.75	0.38	−0.76	5.70*	5.08
Andy (ABACADABACAD)										
". . . contingent access to preferred edible items was immediately effective in increasing compliance" (p. 607).										
Compliance (AD)	3	6	(1, 0, 0)	.66*	4.03	16.49	−0.24	0.43	4.05*	5.03
Figure 2.										
Ricky (ABACABAC)										
"For Ricky, problem behavior occurred exclusively during the guided compliance conditions, but appeared to subside during each implementation" (p. 607).										
Problem behavior (AACAAC)	3/3/3/3	6/8	(1, 0, 0)	.07	−1.58	12.45	0.36	−2.16*	3.94*	—
Ian (ABACABADAD)										
"Ian exhibited most of his problem behavior during rationale conditions" (p. 608).										
Problem behavior (ABAB)	**5/9**	**8/5**	**(1, 0, 0)**	**.46***	**32.75**	**26.72**	**−0.28**	**0.22**	**−0.07**	**0.07**
Figure 3.										
Ed (ABACABAC)										
"For Ed, compliance improved when he received access to his preferred edible item contingent on compliance" (p. 609).										
Compliance (ACAC)	**3/4**	**6/7**	**(1, 0, 0)**	**.62***	**5.15**	**22.66**	**0.49**	**1.16**	**1.93**	**1.71**
Carl (ABACABAC)										
"For Carl, contingent access to preferred edible items was also effective in increasing compliance" (p. 609).										
Compliance (ACAC)	3/3	3/16	(1, 0, 0)	.57*	1.06	4.31	−0.42	0.53	37.44*	22.74
Sam (ABACABAC)										
"For Sam, contingent access to preferred edible items was also effective in increasing compliance" (p.609).										
Compliance (ACAC)	3/3	3/25	(1, 0, 0)	.42*	18.73	17.54	−0.84	0.60	8.29*	4.85
Carbone, Sweeney-Kerwin, Attanasio, and Kasper (2010)										
Figure 1. (AB)										
"Tony's mean responding showed a threefold increase in unprompted vocal responding. .." (p. 707).										
Tony (AB)[b]	10	21	(5, 0, 0)	.66*	9.58*	6.27	0.16	0.76	3.09*	2.30
"Both Ralph's and Nick's manual sign mands were accompanied by very few vocal responses during baseline, but demonstrated substantial increases in unprompted vocalizations during treatment" (p. 707).										
Ralph (AB)	**17**	**10**	**(1, 0, 0)**	**.67***	**1.18**	**5.51**	**−0.12**	**0.87**	**1.51**	**1.57**
Nick (AB)	**21**	**7**	**(1, 0, 0)**	**.59***	**1.13**	**1.46**	**0.11**	**−0.31**	**−0.33**	**0.40**
Ulke-Kurkcuoglu and Kircaali-Iftar (2010)										
Figure 1. (ABABA)										
"All participants except Yavuz consistently displayed higher levels of on-task behaviors during choice conditions than during baseline" (p. 719).										
Utku (ABABA)	4/4/4	4/4	(1, 0, 0)	.47*	62.45*	5.52	4.13*	−2.01	5.91*	—
Alp (ABABA)	4/4/4	4/4	(1, 0, 0)	.53*	65.04*	3.52	6.07*	−0.79	6.34*	—
Selim (ABABA)	4/4/4	4/4	(1, 0, 0)	.57*	70.97*	3.62	3.10*	−0.02	4.64*	—
Yavuz (ABABA)	**4/4/4**	**4/4**	**(1, 0, 0)**	**.65***	**66.03***	**2.51**	**16.08***	**−6.11***	**13.31***	**—**
"Yavuz's on-task behavior during the last baseline condition was similar to his on-task behavior in the choice conditions" (p. 719).										
Yavuz (BBA)	**4**	**4/4**	**(1, 0, 0)**	**.23**	**95.83***	**1.60**	**1.62**	**0.43**	**−3.13***	**3.85**

(Continued on next page)

TABLE 1
Summary of Visual Analysis and Interrupted Time Series Analysis (ITSA) Based on Eligible Studies Published in the *Journal of Applied Behavior Analysis* in 2010 *(Continued)*

Figure	*N* BL	*N* TX	ARIMA	AR 1	Level	Error σ	Slope	Δ Slope	Δ Level	*d*
Roscoe, Kindle, and Pence (2010)										
Figure 1. Bottom panel (ABAB)										
"During the first FCT intervention phase [as well as return to the FCT intervention], she did not exhibit aggression and emitted the communication response at mostly short latencies and at a high frequency. . ." (p. 726).										
Aggression (ABAB)	3/3	5/9	(1, 0, 0)	.46*	42.94	65.99	−0.73	0.51	5.45*	3.80
Communication (ABAB)	3/3	5/9	(1, 0, 0)	.30	304.88*	95.74	−1.11	0.53	−3.17*	2.20
Travis and Sturmey (2010)										
Figure 1. Bottom panel (ABAB)										
"The immediate success of this intervention. .. "(p. 748).										
Nondelusional statements (ABAB)	4/4	5/4	(1, 0, 0)	.55*	0.61*	0.15	−2.29*	5.02*	6.61*	—
Delusional statements (ABAB)	4/4	5/4	(1, 0, 0)	.58*	1.45*	0.16	−0.07	−1.66	−7.95*	6.38
Wilder, Nicholson, and Allison (2010)										
Figure 1.										
Top panel (ABABACAC)										
"Ralph's compliance was generally low during baseline. . . . However, when physical guidance was added, his compliance increased and remained at high levels" (p. 753).										
Compliance (ACAC)	**3/6**	**10/5**	**(1, 0, 0)**	**.64***	**11.83**	**31.61**	**0.29**	**1.35**	**−0.21**	**0.21**
Middle panel (ABABACADACAD)										
"During the first advance notice plus physical guidance phase, compliance remained relatively low. .." (p. 753).										
Compliance (AC)	4	7	(1, 0, 0)	.37	−1.21	14.99	0.13	−0.21	0.50	0.62
"During the physical guidance only phase, compliance increased and remained at high levels. .." (p. 753).										
Compliance (AD)	**3**	**11**	**(1, 0, 0)**	**.62***	**4.21**	**25.27**	**−0.24**	**0.73**	**−0.07**	**0.08**
Compliance increased again during the second advance notice plus physical guidance phase. . ." (p.753).										
Compliance (AC)	**3**	**8**	**(1, 0, 0)**	**.41***	**48.89**	**29.74**	**−1.03**	**1.05**	**2.26**	**2.49**
Compliance. .. increased to high, stable levels during the second physical guidance phase. .." (p. 753).										
Compliance (AD)	9	9	(1, 0, 0)	.04	60.76*	25.15	−2.20*	−0.23	5.27*	—
Bottom panel (ABABACACAD)										
"When physical guidance was added, compliance increased. .." (p. 753).										
Compliance (ACAC)	**3/4**	**8/10**	**(1, 0, 0)**	**.52***	**12.84**	**24.74**	**−0.31**	**0.69**	**0.90**	**0.82**
"During the last phase, advance notice was removed and physical guidance alone was implemented. Compliance improved. .. and remained at high levels. .. during this phase" (p. 753).										
Compliance (AD)	4	7	(1, 0, 0)	.70*	−3.05	14.24	0.31	3.56*	−2.29*	—
Miller, Lerman, and Fritz (2010)										
Figure 1. (ABAB)										
". . . the percentage of trials with reprimands decreased during the first extinction phase. .. Cindy's responding was similar to that during her first extinction phase, although suppression was less pronounced" (p. 771).										
Cindy (ABAB)	**3/4**	**4/8**	**(1, 0, 0)**	**.38**	**71.41***	**30.05**	**1.19**	**−1.23**	**−0.55**	**0.49**

Note. The following information is included in the first column: authors of the publication, figure label as it is presented in the publication, experimental design presented using capital letters in the parenthesis. Unless otherwise indicated with the superscript (†), each ITSA model was determined based on four parameters: level, slope, change in slope, and change in level. *N* BL = number of observations in the baseline or reference phase; *N* TX = number of observations in the treatment phase; ARIMA = autoregressive moving average model; AR 1 = autoregressive term 1; Level = intercept; Error σ = standard error estimate; Slope = *t* test statistic for linear trend of the time series; Δ Slope = *t* test statistic for change in slope at the interruption point; Δ Level = *t* test statistic for change in level at the interruption point; *d* = Cohen's *d* effect size; Cohen's *d* effect size is not available for time series with significant slope or change in slope. The quotes in the Table are the interpretation of a significant effect as presented in the original paper. The comparison that the quote refers to is indicated by the bolding below the quote.

[a]significant AR 2.

[b]significant AR 2 and AR 3.

[c]significant AR 2, AR 3, and AR 4.

[d]significant AR 2, AR 3, AR 4, and AR 5.

[†]ITSA model estimated separately for slope and change in slope due to small number of observation that affected model's stability.

*$p < .05$

time schedule reduced problem behavior, but did not increase compliance for Sam. Lomas et al. (2010) stated that "levels of compliance were only slightly higher during treatment with VT food and praise for Sam. .." and that "variable-time delivery of food and praise superimposed on a demand baseline (in which problem behavior continued to produce escape) greatly reduced problem behavior. . ." (p. 431).

ITSA was implemented to evaluate the effect of variable-time delivery on problem behavior and compliance. The ARIMA (1, 0, 0) was applied to both behaviors to estimate 4 parameters: level, change in level, slope, and change in slope.

For problem behavior, lag-1 autocorrelation was .40. The analysis for slope and change in slope yielded nonsignificant findings, whereas change in level in the variable-time delivery phase indicated significant decrease in problem behavior ($t\,(18) = -2.39, p < .05$) with medium effect size ($d = 1.85$) based on tertile distribution. The findings based on statistical analysis confirm conclusions drawn from VA, indicating decrease in problem behavior due to variable-time delivery of preferred food and praise.

For compliance, lag-1 autocorrelation was .13. The analysis for slope and change in slope yielded nonsignificant findings, whereas change in level in the variable-time delivery phase indicated significant increase in compliance ($t\,(18) = 2.43, p < .05$) with medium effect size ($d = 1.76$). The findings based on statistical analysis did not confirm the conclusions drawn from VA, which indicated only slight increases in compliance, while statistical findings show significant increases with large effect sizes. ITSA details are presented in Table 1.

Example 2

The second example is based on a study that examined the effectiveness of a device that prevents drivers from changing gears for up to 8 seconds unless the seatbelt is buckled. The study was based on an ABA reversal design and included 101 commercial drivers (Van Houten et al., 2010). Data for one driver is displayed in Figure 6. Based on the VA of the data presented in the top panel, Van Houten et al. (2010) concluded ". . . an increase in seat belt use following the 8-s delay and a decline when the delay was removed" (p. 377).

ITSA was implemented to evaluate the effect of the 8-s gearshift delay on seatbelt use. Two ARIMAs (5, 0, 0) were applied to test increases in seatbelt use following the 8-s delay (AB) and to test a decline in seatbelt use when the delay was removed (BA). Each model estimated 4 parameters: level, change in level, slope, and change in slope. For AB phase of the design, lag-1 autocorrelation was significant (AR 1 = .56). The analysis for slope and change in slope yielded nonsignificant findings, whereas change in level in the 8-s delay phase indicated significant increase in seatbelt use ($t\,(79) = 8.59, p < .05$) with large effect size ($d = 2.78$). The findings based on statistical analysis confirm conclusions drawn from VA, indicating an increase in seatbelt use due to 8-s gearshift delay. For the BA phase of the design, lag-1 autocorrelation was significant (AR 1 = .76) . The analysis for slope yielded nonsignificant findings; however, change in slope and change in level were significant and indicated a decrease in seatbelt use due to removal of the gearshift delay ($t\,(87) = -2.19, p < .05$; $t\,(87) = -8.58, p < .05$ for change in slope and change in level respectively). The findings based on statistical analysis confirm conclusions drawn from VA, indicating a decrease in seatbelt use following removal of the 8-s gearshift delay.

Example 3

The third example is based on a study that performed several experiments, one of which examined the effects of delivery of higher quality reinforcement following appropriate behavior and lower quality reinforcement following problem behavior on changes in behavior (Athens & Vollmer, 2010). The study participant reported in this example was a 7-year-old boy diagnosed with attention deficit hyperactivity disorder, and the experiment was based on ABCAC design. Based on the VA of data presented in Figures 7 and 8, Athens and Vollmer (2010) made several conclusions such as "in the 1 HQ/ 1 LQ condition, rates of problem behavior decreased, and appropriate behavior increased" (p. 579); "problem behavior decreased, and appropriate behavior increased to high levels during the return to the 3 HQ/ 1 LQ condition" (p. 580); and "in summary, results of the quality analyses indicated that. . . the relative rates of both problem behavior and appropriate behavior were sensitive to the quality of reinforcement available for each alternative" (p. 581).

ITSA was implemented to evaluate the effect of the quality reinforcement on problem behavior and appropriate behavior. Three ARIMAs, estimating 4 parameters (slope, change in slope, level, and change in level) were applied to test each of the conclusions made based on VA.

First, an ARIMA (1, 0, 0) was implemented to evaluate the effects of 1 HQ/ 1 LQ on problem behavior and appropriate behavior (AB phase of the experiment). The lag-1 autocorrelations were −.05 and .13, for problem behavior and compliance, respectively. For problem behavior, ITSA revealed nonsignificant slope, significant change in slope ($t\,(15) = 2.18, p < .05$), and nonsignificant change in level. These findings indicated an increase in problem behavior in the quality reinforcement phase and did not confirm conclusions based on VA that found a decrease in problem behavior. For appropriate behavior, ITSA indicated significant slope ($t\,(15) = -2.22, p < .05$), nonsignificant change in slope, and significant change in level ($t\,(15) = 4.24, p < .05$). These findings indicated an initial decreasing trend in baseline phase (A) followed by an increase in compliance as an effect of 1 HQ/ 1 LQ quality reinforcement. The statistical results are consistent with VA conclusions.

		Statistical Analysis		
		Significant	Not	Total
Graphical Analysis	Significant	79 73.4% r_1>.40 m_d = 4.25	52 65.4% r_1>.40 m_d = .95	131
	Not	8 25% r_1>.40 m_d = 3.13	15 13.3% r_1>.40 m_d = .81	23
	Total	87	67	**154**

FIGURE 3 Agreement between graphical analysis and statistical analysis.

Second, an ARIMA (1, 0, 0) was applied to examine the effect of the return to 3 HQ/ 1 LQ phase on problem and appropriate behavior (AC phase of the experiment). The lag-1 autocorrelations were −.29 and .06, for problem behavior and compliance, respectively. For problem behavior, ITSA revealed significant slope (t (12) = −3.46, p < .05), a nonsignificant change in slope, and significant change in level (t (12) = 2.21, p < .05). These findings indicate an initial decreasing trend in problem behavior; however, the change in level indicate an increase in problem behavior during the 3 HQ/ 1 LQ experimental phase. The statistical results are not consistent with VA that concluded a decrease in problem behavior during the return to the quality reinforcement phase. For appropriate behavior, ITSA revealed nonsignificant slope, change in slope, and change in level. These findings indicate that no significant change in compliance occurred as a result of the 3 HQ/ 1 LQ experimental phase. The statistical results are not consistent with VA that concluded a high increase in compliance as a result of quality reinforcement phase.

Third, an ARIMA (1, 0, 0) and (5, 0, 0), for problem and appropriate behavior, respectively, was applied to examine the overall effect of the quality reinforcement (ABCAC experimental design). The lag-1 autocorrelations were −.07 for problem behavior and significant .44, for compliance. For problem behavior, ITSA revealed significant slope (t (41) = −2.88, p < .05), a nonsignificant change in slope, and significant change in level (t (41) = −2.91, p < .05). These findings indicate an initial decreasing trend, as well as decrease in problem behavior during the quality reinforcement phases. These results are consistent with VA. For appropriate behavior, ITSA revealed an initial significant increase in trend (t (41) = 3.49, p < .05), a nonsignificant change in slope, and change in level, indicating that quality of reinforcement did not have an effect on compliance. These results are not consistent with VA that concluded effectiveness of experimental treatment on increasing appropriate behavior.

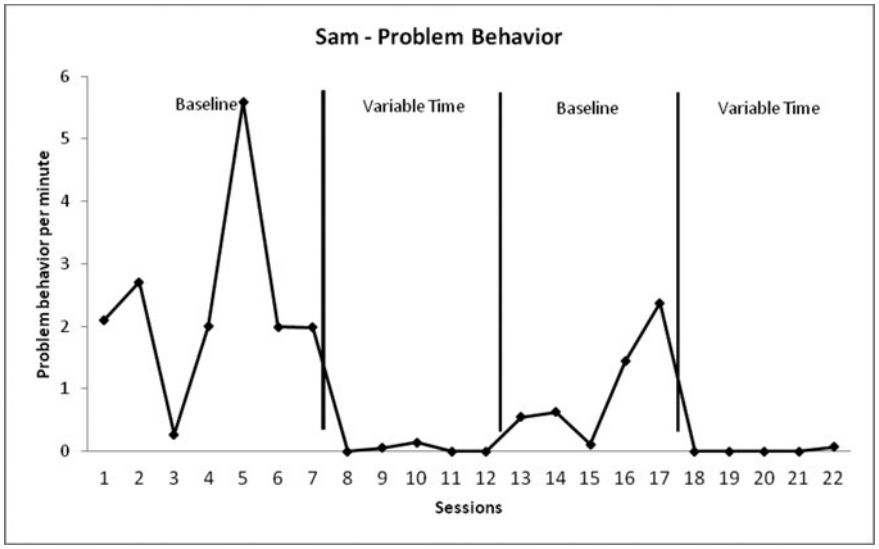


FIGURE 4 Graphical presentation of the data illustrated in the first example of interrupted time series analysis (ITSA) application. *Note.* Figure reproduced from the data extracted using UnGraph® software from Lomas, Fisher, and Kelly, 2010 (p. 430).

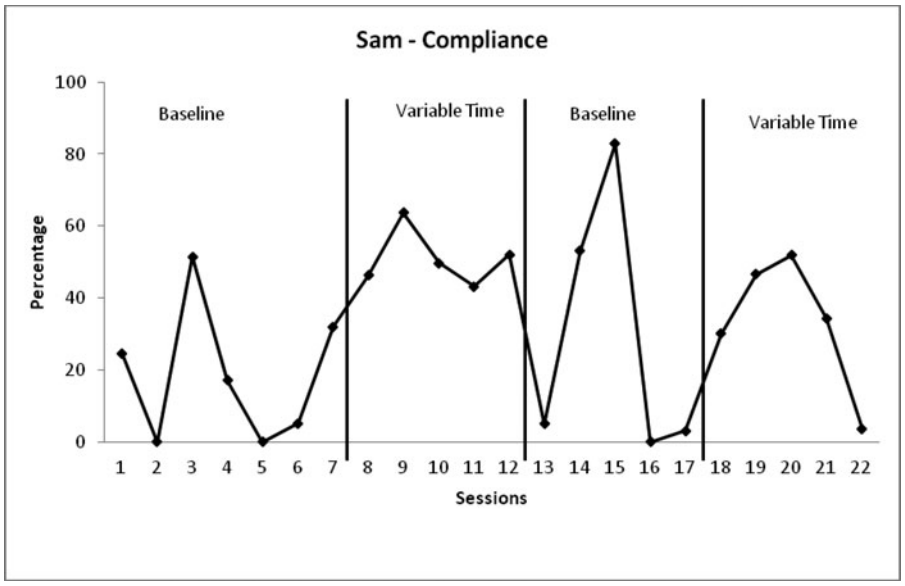


FIGURE 5 Graphical presentation of the data illustrated in the first example of interrupted time series analysis (ITSA) application. *Note.* Figure reproduced from the data extracted using UnGraph® software from Lomas, Fisher, and Kelly, 2010 (p. 430).

RESULTS

Sample

A total of 75 research papers were published in the JABA in 2010. After reviewing the content of the publications, 25 papers met eligibility criteria and were included in the study. Excluded publications did not present interrupted time series data (27), presented fewer than 3 observations in at least one phase of the design (4), presented cumulative data (3), or alternating-treatment designs (9). One study presented generated, hypothetical data, and one study presented a graph with insufficiently defined observations, which prevented data point extraction. Five studies were ineligible because presented descriptions of the findings based on the VA of the graphs were not possible to verify using ITSA (e.g., findings were generalized across all conducted experiments, rather than reported for each experiment separately). The eligible publications included one or more graphs. A total of 99 graphs presenting interrupted time series data with corresponding conclusions based on VA were included in the

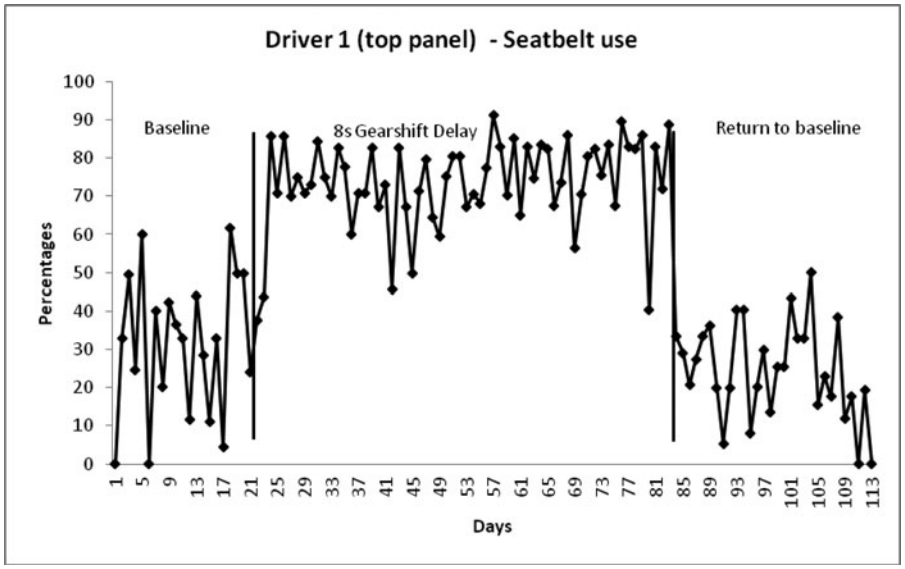


FIGURE 6 Graphical presentation of the data illustrated in the second example of interrupted time series analysis (ITSA) application. *Note.* Figure reproduced from the data extracted using UnGraph® software from Van Houten et al., 2010 (p. 377).

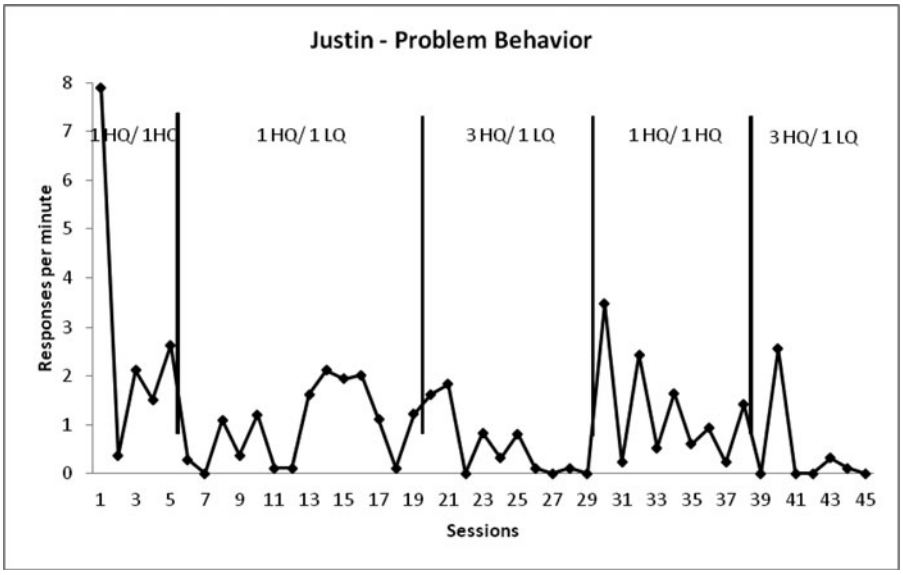


FIGURE 7 Graphical presentation of the data illustrated in the third example of interrupted time series analysis (ITSA) application. *Note.* Figure reproduced from the data extracted using UnGraph® software from Athens and Vollmer, 2010 (p. 580).

study. The graphs displayed a diversified range of single-case designs, such as AB design and its variations (e.g., ABA, ABAB), ABC design and its variations (e.g., ABCA, ABCACA, ABABACBC), and designs that included more than two different interventions (e.g., ABCD, ABCDEFBFEDC) (see Table 1 for details).

Based on 99 graphs, a total of 163 ITSA were performed, either because some graphs presented more than one interrupted time series data (e.g., two independent behaviors were plotted on a single graph) or multiple conclusions were made based on VA (e.g., conclusions were made based on different phases of the study). ITSA was applied to the data with the corresponding description of the findings formulated in a way that could be validated using statistical methods. To be certain that specific conclusions based on VA are directly comparable to findings based on ITSA, the key conclusions

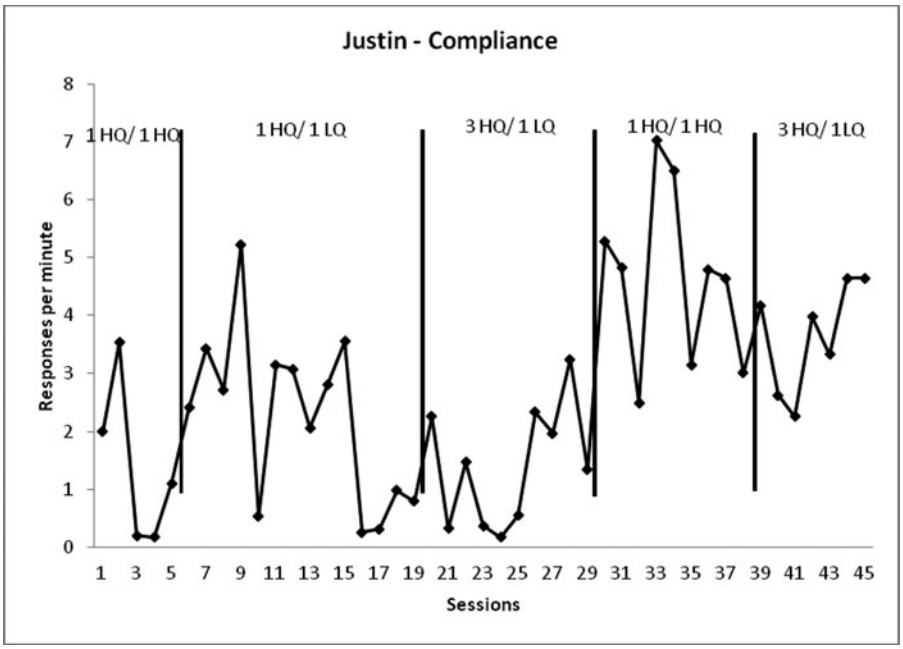


FIGURE 8 Graphical presentation of the data illustrated in the third example of interrupted time series analysis (ITSA) application. *Note.* Figure reproduced from the data extracted using UnGraph® software from Athens and Vollmer, 2010 (p. 580).

were identified and matched with specific study phases, so that ITSA can be computed only for those phases. To illustrate the comparison process, the first study presented in Table 1 (St. Peter Pipkin, Vollmer, & Sloman, 2010) is used as an example. The complete study design is presented in parenthesis, next to the figure number (Figure 6. Top panel (A**B**CD**EF**BFEDC)). The bolded and underlined phases are those comparisons made in the paper. In the row below, a conclusion based on VA of selected phases is cited: "DRA lost its efficacy when implemented at less than 50% integrity with combined omission and commission errors" (p. 60). In order to directly compare findings based on VA and ITSA, statistical analyses are performed only on data obtained from selected phases. The analyzed phases identified by letters are reported on the left side of the table in the same row as corresponding ITSA results, e.g. (B (EF)).

Descriptive Statistics

The number of observations in the analyzed experiments ranged from 8 to 136, with a minimum of 3 and maximum of 90 observations per phase. For 9 (5.52%) analyzed experiments, the interrupted-time series ARIMA did not converge. Six of those experiments came from one study that had multiple single-case data series characterized by low number of observations (<12) and low variability across observations; two experiments had higher number of observations (43 and 136) but low variability across observations; one experiment had high variability across 22 observations. The majority of the examples where the model did not converge typically did not meet the What Works Clearinghouse standards (Kratochwill & Levin, 2010; Smith, 2012).

An assumed ARIMA (1, 0, 0) (Simonton, 1977) was applied to 120 data series (77.92%). The general transformation ARIMA (5, 0, 0) (Velicer & McDonald, 1984; Harrop & Velicer, 1985) was applied to 32 data series (20.78%), all of which had 30 or more observations. ARIMAs (3, 0, 0) and (2, 0, 0) were applied to two experiments, after an assumed ARIMA (1, 0, 0) indicated correlated residuals and the general transformation ARIMA (5, 0, 0) (Velicer & McDonald, 1984; Harrop & Velicer, 1985) did not converge due to an insufficient number of observations (for details see Table 1). According to classification outlined by Jones et al. (1978), low lag-1 autocorrelations ranging from .00 to .50 were found for 75 (46.01%) time series data, moderate lag-1 autocorrelations ranging from .51 to .75 were found for 67 (41.10%) time series data, and high lag-1 autocorrelations over .75 were found for 8 (4.91%) time series data. Lag-1 autocorrelation less than .00 were found for 13 (7.98%) time series data and ranged from −.32 to −.05. Lag-1 autocorrelations were significant for 93 time series data, 28 of those time series data also had significant lag-2 autocorrelations. The findings show a high heterogeneity of the lag-1 autocorrelations that could be largely related to a small number of observations in some experiments as well as different study designs. It has been shown that for short data series autocorrelations are negatively biased and may underestimate the true autocorrelation, and correcting for small sample bias is suggested (Huitema & McKean, 2000; Shadish & Sullivan, 2011). Figure 1 presents the distribution of the lag-1 autocorrelations for the eligible studies, and details are presented in Table 1.

Changes in the target behavior such as changes in level and decreasing or increasing trend were evaluated for all 154 data series for which an ARIMA was established. Twenty-three experiments (14.94%) had significant slope, indicating that the target behavior was either decreasing or increasing in the baseline phase; 15 (9.74%) had significant change in slope due to experimental design, indicating that the target behavior was either decreasing or increasing in the intervention phase of the experiment; 18 (11.69%) had significant slope and change in slope, indicating that target behavior was either decreasing or increasing in both phases of the experiment. The nonlinearity of the slopes was not examined. Over 50% of the examined time series data ($k = 79$) had significant changes in level as a result of the experiment due to examined study design phase change.

An effect size estimate, based on a similar formula used to evaluate Cohen's d, was estimated for all experiments that did not have significant slope or change in slope, a total of 98 (63.64%). The effect sizes ranged from 0.00 to 22.74 (see Table 1 for details). Figure 2 presents the distribution of the effect size estimates for the eligible studies. Cohen's (1988) traditional classification of effect sizes cannot be used, as effects calculated for single-case designs are expected to be inflated relatively to Cohen's standards for between-group studies. Therefore, in this study we propose alternative classification based on the tertile distribution of the effect sizes, where effects ranging from 0.00 to 0.99 are classified as small, those ranging from 1.00 to 2.49 as medium, and large effect size are defined as 2.50 or greater.

ITSA and VA Comparison

Comparison of the findings based on VA and ITSA was performed for 154 data series. Consistent results were found for 94 (61.04%) data series, with most conclusions ($k = 79$, 84.04%) referring to significant changes between different phases of the experiment, and 15 (15.96%) referring to nonsignificant changes such as reversal to baseline. For the remaining 60 experiments (38.96%), the findings based on statistical analysis did not confirm the conclusions based on VA (bolded data in Table 1). Among the experiments that led to inconsistent findings between the two methods, 30% had significant slope, change in slope or both, and 53% had lag-1 autoregressive term greater than .40.

Fifty-two, out of 60 data series that were identifies ad inconsistent, were identified as significant based on the VA method, while ITSA did not confirm these results. For 49 of those experiments, VA indicated significant changes between

different phases of the study design, while statistical analysis did not reveal significant differences. No significant slope parameter or change in slope parameter was found for 38 of those experiments; 10 experiments had significant slope in the first phase with the trend of the targeted behavior moving in the same direction as hypothesized in the reference phase, and 1 experiment had significant slope in the treatment phase with the trend of the targeted behavior moving in the opposite direction than hypothesized based on VA analysis. For three experiments statistical analysis revealed significant findings. However those findings were in the opposite direction than conclusions reported based on VA. Two out of three experiments had significant slope in the first phase with the trend of the targeted behavior moving in the same direction as hypothesized in the reference phase. All three experiments were reported in the same research paper and are bolded and italicized in Table 1.

For eight experiments, nonsignificant findings based on VA were not confirmed by statistical analysis. Three of those experiments had no significant slope or change in slope parameter, four had significant slope in the first phase with the trend of the targeted behavior moving in the same direction as hypothesized in the reference phase, and one experiment had significant slope in the treatment phase with the trend of the targeted behavior moving in the opposite direction then hypothesized based on VA analysis.

The level of agreement between VA and ITSA was calculated based on the difference between the observed agreement and the expected agreement that would be present by chance alone. Kappa coefficient is a measure of this difference, ranging from –1 to 1, where 1 is a perfect agreement, 0 is an agreement by chance, and value < 0 would indicate an agreement less than expected by chance (Cohen, 1960). Figure 3 provides a summary of the agreement and disagreement between the two methods as well as the percent of cases with lag-1 autocorrelations greater than .40 for each cell. The overall level of agreement was low (Cohen's Kappa $= .14$) (Cohen, 1960). The VA results identified as significant had a very high percent of cases with lag-1 autocorrelations greater than .40. This is consistent with the potential bias that the positive autocorrelation can create the illusion of significance but decrease the apparent variability of the series. Figure 3 also presents the mean effect size estimate for each cell. As would be expected, the average effect size was higher for the significant ITSA results.

DISCUSSION

This study applied ITSA to 75 studies published in the JABA in 2010 and compared the conclusions authors reported based on VA with those obtained through ITSA. Issues such as autocorrelation, effect size estimation, and level of agreement between statistical and VA were addressed. Evaluated studies covered a wide range of single-case experiments that included different study designs, such as multiple-baseline, reversal, and multiple intervention designs. The experiments also differed in total number of observations in each study as well as within each phase of the design. ITSA model was estimated for all but nine of the eligible studies, indicating that this statistical method can be applied to a wide range of single-case experimental designs.

Agreement between Visual and Statistical Analysis

Comparison of the conclusions drawn from VA and ITSA revealed an overall low level of agreement (Kappa $= .14$). When graphical presentation of the intervention effects presents ideal or almost ideal data patterns, such as low variability, no trend, and large effect size, ITSA was in agreement with VA for 94 data series, including those with small numbers of observations. However, in 60 (38.96%) of the evaluated data series, the conclusions drawn based on VA did not agree with the statistical analysis. VA was more likely to imply significant effects when ITSA indicated nonsignificant findings. This is the opposite state of affairs expected by Baer (1977), who argued that visual analysis should be less likely to report significant findings than statistical analysis. Only for eight experiments, nonsignificant findings based on VA were not confirmed by statistical analysis, and for 3 experiments ITSA resulted in significant findings but in the opposite direction than indicated by VA. Among the experiments that led to inconsistent findings between the two methods, 30% had significant slope, change in slope or both, and 53% had lag-1 autoregressive term greater than .40.

If we view statistical analysis as a necessary but not sufficient condition for clinical significance, this result is discouraging. Moderate to high autocorrelation, present in most examples, is one potential explanation for the low agreement. Also, trend in the data, closely related to the autocorrelation and not easily observable, particularly in short series, may impact the accuracy of the conclusions based on VA. ITSA is able to account for trend in the data when examining intervention effects, as well as evaluate quantitatively trend and change in trend that may occur across different phases of the design.

Although the failure to detect a statistically significant effect occurred at a much smaller rate (5%), these errors have the potential to prematurely terminate the investigation of a potentially effective intervention. Initial studies of an intervention in a real world study typically represent an attempt to detect an effect in a very noisy environment, and effect sizes that are initially small can become much more important with additional controls.

Autocorrelation

Overall findings based on ITSA revealed high lag-1 autocorrelations for most of the evaluated data, including short time series of less than 20 observations. These results confirm findings based on earlier studies showing that serial dependency

is a common property of single-case data (Jones, Vaught, & Weinrott, 1977; Jones et al., 1978; Matyas & Greenwood, 1990; Barlow et al., 2009). With over 60% of the lag-1 autocorrelations at either moderate or high level, the assumption that autocorrelations can be ignored (Huitema & McKeon, 1998) seems to be indefensible. The effect of a positive autocorrelation is to decrease the apparent degree of variability. This would potentially affect both graphical analysis and any statistical analysis that ignores dependency in the data.

The autocorrelations can also help address another important research question, i.e., what is the nature of the generating function for the observed data. The autocorrelations also provide information about the extent to which the ergodic theorems are satisfied, a critical question for combining data across individuals (Molenaar, 2008; Velicer & Molenaar, 2013). In order to draw valid inferences from group-level data to the individual level, two ergodic theorem conditions must be met: (1) the individual trajectories must obey the same dynamic laws, and (2) must have equivalent mean levels and serial dependencies (Molenaar, 2008; Velicer, Babbin, & Palumbo, 2014). However, the small sample sizes available in the studies reviewed here do not permit these questions to be addressed.

Effect Size Estimation

The effect size estimates were predominately large with some very large effect sizes such as $d = 22.74$, an extremely large effect size for the behavioral sciences. The term "clinical significance" is largely undefined but can be viewed as analogous to a large effect size. (Statistical significance is typically viewed as a necessary but not sufficient condition for clinical significance.) Based on this interpretation, the effect size estimates observed in this set of studies support the contention that graphical methods focus on clinically significant effect sizes.

Advantages of Statistical Analysis

ITSA provides supplementary quantitative information such as degree of the serial dependency, trend, changes in trend and level across phases, and variability of the data, that are not available through visual inspection of the graphs. Evaluation of the serial dependency could provide information about the generating function of the examined behavior, such as the strength of relationships of the observations or cyclic patterns in the behavior that are not observable by visual inspection of the graph. Unbiased statistical evaluation of the graphs facilitates comparison of the intervention effects across different individuals within the same experiment or across different studies. This information is particularly useful when experiments are executed across multiple subjects or settings, allowing for a better understanding of the unique variability of the behavior across different subjects or settings.

ITSA facilitates an estimate of effect size similar to Cohen's *d* that enables systematic meta-analytic review of single-case experiments, as well as evaluation of the intervention effects for experiments with small numbers of observations. In this study, we used the effect size to examine the magnitude of the intervention effects within single-cases; for the application of Cohen's *d* effect size to between-cases see work by Hedges et al. (2012). Statistical significance tests are largely dependent on the sample size. For small sample sizes, the results may be insignificant due to insufficient statistical power. However effect size is independent of sample size, and meta-analysis can provide more accurate estimates of effect size based on multiple replications. The development of the new software such as UnGraph® (Biosoft, 2004), DataThief (Tummers, 2006), and a new function in R (Bulté & Onghena, 2012) permits extraction of the data from published graphs and reanalysis using ITSA. This would permit the inclusion of historical data based on single-case studies in meta-analytical studies.

Limitations

The results of this study have limited representativeness. The collected data is based on a set of single-case studies published in JABA in 2010. The characteristics of these studies may influence the findings, particularly the large effect sizes and high autocorrelations, which are likely a product of the design and interventions published in JABA and the journal's preference of publishing studies that are likely to show the "clinically meaningful" threshold.

In addition, the autocorrelations were not corrected for small sample bias, which could underestimate the true autocorrelation. Therefore replication of these results in other samples of the published studies within the applied behavior analysis field is needed. Another potential limitation of the analysis was that nonlinearity was not examined. There is some reason to believe nonlinearity is present in this type of data.

The sample size, defined in single-case study designs as number of observations in each phase rather than number of different individuals, is another limitation of the study. For the set of studies reviewed here, the numbers of observations was generally very small compared to idiographic studies reported in other disciplines or even other areas of behavioral science. The average number of data points was 28 (median = 19) for 163 data series. These findings are similar to those presented by Shadish and Sullivan (2011) in a comprehensive review of 21 journals that report on single-case studies. They found that the median number of data points was equal to 20, whereas average number of data points in JABA in the year of 2008 was 29.

Large effect sizes are necessary for any type of significance, given the small sample sizes. However, a power analysis was seldom performed to guide the choice of the number of observations. Given that these studies focus on four parameters (slope, change in slope, level, and change in level), the lack of statistical power produces very poor estimates

of the parameters of interest. Increasing the number of observations by even a small amount would greatly improve the quality of the research. There are times when obtaining additional observations is very difficult and expensive, but at other times a larger number of observations were collapsed for the graphical presentation of the data.

The number of observations is also related to the time between observations. Time is a core concept for idiographic studies, and we presently have very little information to guide researchers on how frequently observations should be taken. Advances from the information sciences are producing new measures that can greatly improve the quality and number of observations. A review of these methods, often labeled telemetrics, is provided by Goodwin, Velicer, and Intille (2008). Indeed, advances in telemetrics may shift the issue from not having many observations to having too many observations.

CONCLUSIONS

In conclusion, ITSA models can be applied to a large number of the published applied examples of single-case study designs. Moderate to high lag-1 autocorrelations (>.50) were found for 46% of the data series, and the majority of first order autocorrelations (more than 60%) were positive and at the moderate to high level (.41–.60 or >.60). Comparison of the conclusions drawn from VA and ITSA revealed an overall low level of agreement (Kappa = .14), and the results of the study support the conclusion that VA is prone to bias and should not be used as a stand-alone analytical method. When both methods produce discrepant results, the researcher should determine the basis for the discrepancy. Finally, ITSA provides important additional information such as effect size estimates, which permits the application of meta-analysis and the accumulation of knowledge.

ARTICLE INFORMATION

Conflict of Interest Disclosures: Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

Ethical Principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: This work was partially supported by Grant DA020112 from NIDA (PI: Velicer).

Role of the Funders/Sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Acknowledgments: The authors would like thank Dr. William Shadish, Dr. Colleen Redding, and Dr. Mathew Goodwin for their comments on prior versions of this manuscript. The ideas and opinions expressed herein are those of the authors alone, and endorsement by the authors' institution or NIDA is not intended and should not be inferred.

REFERENCES

Note: References marked with an asterisk (*) indicate studies included in the visual and interrupted time-series analysis comparison.

Aitken, A. C. (1934). On least squares and lineal combination of observations. *Proceedings of the Royal Society of Edinburg H*, *55*, 42–47.

*Athens, E. S., & Vollmer, T. R. (2010). An investigation of differential reinforcement of alternative behavior without extinction. *Journal of Applied Behavior Analysis*, *43*, 569–589.

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

Barlow, D. H., Nock, M. K., & Hersen, M. (2009). *Single case experimental designs: Strategies for studying behavior for change* (3rd ed.). Boston: Pearson Education.

Bengali, M. K., & Ottenbacher, K. J. (1998). The effects of autocorrelation on the results of visually analyzing data from single-subject designs. *Quantitative Research Series*, *52*, 650–655.

Biosoft (2004). *UnGraph® for Windows* (Version 5.0). Cambridge, UK: Author.

Brossart, D. F., Parker, R. I., Olson, E. A., & Mahadevan, L. (2006). The relationship between visual analysis and five statistical analyses in a simple AB single-case research design. *Behavior Modification*, *30*, 531–563.

Bulté, I., & Onghena, P. (2012). When the truth hits you between the eyes: A software tool for the visual analysis of single-case experimental data. *Methodology*, *8*, 104–114.

*Carbone, V. J., Sweeney-Kerwin, E. J., Attanasio, V., & Kasper, T. (2010). Increasing the vocal responses of children with autism and developmental disabilities using manual sign mand training and prompt delay. *Journal of Applied Behavior Analysis*, *43*, 705–709.

*Carter, S. L. (2010). A comparison of various forms of reinforcement with and without extinction as treatment for escape-maintained problem behavior. *Journal of Applied Behavior Analysis*, *43*, 543–546.

Choe, G. H. (2005). *Computational ergodic theory*. Berlin: Springer.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46.

Cohen, J. (1988). *Statistical power analysis for behavioral sciences*. Hillsdale, NJ: Erlbaum.

Crosbie, J. (1993). Interrupted time-series analysis with brief single-subject data. *Journal of Consulting and Clinical Psychology*, *61*, 966–974.

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

*Digennaro-Reed, F. D., Codding, R., Catania, C. N., & Maguire, H. (2010). Effects of video modeling on treatment integrity of behavioral interventions. *Journal of Applied Behavior Analysis*, *43*, 291–295.

*Dolezal, D. N., & Kurtz, P. F. (2010). Evaluation of combined-antecedent variables on functional analysis results and treatment of problem behavior in a school setting. *Journal of Applied Behavior Analysis*, *43*, 309–314.

*Falcomata, T. S., Roane, H. S., Feeney, B. J., & Stephenson, K. M. (2010). Assessment and treatment of elopement maintained by access to stereotypy. *Journal of Applied Behavior Analysis*, *43*, 513–517.

Glass, G. V., Willson, V. L., & Gottman, J. M. (1975/2008). *Design and analysis of time-series experiments*. Boulder, CO: Colorado Associate University Press.

Goodwin, M. S., Velicer, W. F., & Intille, S. S. (2008). Telemetric monitoring in the behavior sciences. *Behavior Research Methods*, *40*, 328–341.

*Grauvogel-MacAleese, A. N., & Wallace, M. D. (2010). Use of peer-mediated intervention in children with attention deficit hyperactivity disorder. *Journal of Applied Behavior Analysis*, *43*, 547–551.

*Groskreutz, N. C., Karsina, A., Miguel, C. F., & Groskreutz, M. P. (2010). Using complex auditory-visual samples to produce emergent relations in children with autism. *Journal of Applied Behavior Analysis*, *43*, 131–136.

Harrington, M., & Velicer, W. F. (2015). Comparing visual and statistical analysis in single-subject studies: Results for Kazdin Textbook Examples. [Paper in preparation.]

Harrop, J. W., & Velicer, W. F. (1985). A comparison of three alternative methods of time series model identification. *Multivariate Behavioral Research*, *20*, 27–44.

Harrop, J. W., & Velicer, W. F. (1990). Computer programs for interrupted time series analysis: A quantitative evaluation. *Multivariate Behavioral Research*, *25*, 219–231.

Hedges, L. V., Pustejovsky, J. E., & Shadish, W. R. (2012). A standardized mean difference effect size for single case designs. *Research Synthesis Methods*, *3*, 324–239.

Huitema, B. E. (2011). *The analysis of covariance and alternatives: Statistical methods for experiments, quasi-experiments, and single-case studies* (2nd ed.). Hoboken, NJ: Wiley.

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

Huitema, B. E., & McKean, J. W. (2000). A simple and powerful test for autocorrelation errors in OLS intervention model. *Psychological Reports*, *87*, 3–20.

Huitema, B. E., McKean, J. W., & Laraway, S. (2007). Time-series intervention analysis using ITSACORR: Fatal flaws. *Journal of Modern Applied Statistical Methods*, *6*(2), Article 4.

Jazi, M. A., Jones, G., & Lai, C. D. (2012). First-order integer valued AR processes with zero inflated poisson innovations. *Journal of Time Series Analysis*, *33*, 954–963.

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

Jones, R. R., Weinrott, M. R., & Vaught, R. S. (1978). Effects of serial dependency on the agreement between visual and statistical inference. *Journal of Applied Behavior Analysis*, *11*, 277–283.

Kazdin, A. E. (2011). *Single-case research designs: Methods for clinical and applied settings* (2nd ed.). New York: Oxford University.

Kratochwill, T. R., & Levin, J. R. (2010). Enhancing the credibility of single-case intervention research: Randomization to the rescue. *Psychological Methods*, *15*, 124–144.

*Kuhn, D. E., Chirighin, A. E., & Zelenka, K. (2010). Discriminated functional communication: A procedural extension of functional communication training. *Journal of Applied Behavior Analysis*, *43*, 249–264.

*Lee, M. S. H., Yu, C. T., Martin, T. L., & Martin, G. L. (2010). On the relation between reinforce efficacy and preference. *Journal of Applied Behavior Analysis*, *43*, 95–100.

*Leon, Y., Hausman, N. L., Kahng, S., & Becraft, J. L. (2010). Further examination of discriminated functional communication. *Journal of Applied Behavior Analysis*, *43*, 525–530.

*Lomas, J. E., Fisher, W. W., & Kelley, M. E. (2010). The effects of variable-time delivery of food items and praise on problem behavior reinforced by escape. *Journal of Applied Behavior Analysis*, *43*, 425–435.

Maggin, D. M., Swaminathan, H., Rogers, H. J., O'Keeffe, B. V., Sugai, G., & Horner, R. (2011). A generalized least squares regression approach for computing effect sizes in single-case research: Application examples. *Journal of School Psychology*, *49*, 301–321.

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

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

*Miller, J. R., Lerman, D. C., & Fritz, J. N. (2010). An experimental analysis of negative reinforcement contingencies for adults-delivered reprimands. *Journal of Applied Behavior Analysis*, *43*, 769–773.

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement: Interdisciplinary Research and Perspectives*, *2*, 201–211.

Molenaar, P. C. M. (2007). On the implications of the classical ergodic theorems: Analysis of developmental processes has to focus on intra-individual variation. *Developmental Psychobiology*, *50*, 60–69.

Molenaar, P. C. M. (2008). Consequences of the ergodic theorems for classical test theory, factor analysis, and the analysis of developmental processes. In S. M. Hofer & D. F. Alwin (Eds.), *Handbook of Cognitive Aging* (pp. 90–104). Thousand Oaks, CA: Sage.

Molenaar, P. C. M., & Campbell, C. G. (2009). The new person-specific paradigm in psychology. *Current Direction in Psychological Science*, *18*, 112–117.

Ottenbacher, K. J. (1990). Visual inspection of single-subject data: An empirical analysis. *Mental Retardation*, *28*, 283–290.

Ottenbacher, K. J. (1992). Analysis of data in idiographic research. *American Journal of Physical Medicine & Rehabilitation*, *71*, 202–208.

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*, 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*, 284–299

Parsonson, B. S., & Baer, D. M. (1978). The analysis and presentation of graphic data. In T. R. Kratochwill (Ed.), *Single-Subject Research: Strategies for Evaluating Change* (pp. 101–165). New York: Academic.

*Raiff, B. R., & Dallery, J. (2010). Internet-based contingency management to improve adherence with blood glucose testing recommendations for teens with type 1 diabetes. *Journal of Applied Behavior Analysis*, *43*, 487–491.

*Roscoe, E. M., Kindle, A. E., & Pence, S. T. (2010). Functional analysis and treatment of aggression maintained by preferred conversational topics. *Journal of Applied Behavior Analysis*, *43*, 723–727.

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*, 177–183.

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

Simonton, D. K. (1977). Cross-sectional time-series experiments: Some suggested statistical analyses. *Psychological Bulletin*, *84*, 489–502.

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

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

*Stokes, J. V., Luiselli, J. K., & Reed, D. D. (2010). A behavioral intervention for teaching tackling skills to high school football athletes. *Journal of Applied Behavior Analysis*, *43*, 509–512.

*Stokes, J. V., Luiselli, J. K., Reed, D. D., & Fleming, R. K. (2010). Behavioral coaching to improve offensive line pass-blocking skills of high school football athletes. *Journal of Applied Behavior Analysis*, *43*, 463–472.

*St. Peter Pipkin, C., Vollmer, T. R., & Sloman, K. N. (2010). Effects of treatment integrity failures during differential reinforcement of alternative behavior: A translational model. *Journal of Applied Behavior Analysis*, *43*, 47–70.

*Toussaint, K. A., & Tiger, J. H. (2010). Teaching early Braille literacy skills within a stimulus equivalence paradigm to children with degenerative visual impairments. *Journal of Applied Behavior Analysis*, *43*, 181–194.

*Travis, R., & Sturmey, P. (2010). Functional analysis and treatment of the delusional statements of a man with multiple disabilities: A four-year follow-up. *Journal of Applied Behavior Analysis*, *43*, 745–749.

Tummers, B. (2006). DataThief III [Computer software]. Retrieved from http://datathief.org

*Ulke-Kurkcuoglu, B., & Kircaali-Iftar, G. (2010). A comparison of the effects of providing activity and material choice to children with autism spectrum disorders. *Journal of Applied Behavior Analysis*, *43*, 717–721.

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

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

Van den Noortgate, W., & Onghena, P. (2003c). Multilevel meta-analysis: A comparison with traditional meta-analytical procedures. *Educational and Psychological Measurement*, *63*, 765–790.

Van den Noortgate, W., & Onghena, P. (2007). The aggregation of single-case results using hierarchical linear models. *The Behavior Analyst Today*, *8*, 196–209.

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

*Van Houten, R., Malenfant, J. E. L., Reagan, I., Sifrit, K., Compton, R., & Tenebaum, J. (2010). Increasing seat belt use in service vehicle drivers with a gearshift delay. *Journal of Applied Behavior Analysis*, *43*, 369–380.

Velicer, W. F., Babbin, S. F., & Palumbo, B. (2014). Idiographic applications: Issues of ergodicity and generalizability. In P. Molenaar, R. Lerner, & K. Newell (Eds.), *Handbook of Relational Developmental Systems Theory and Methodology* (pp. 425–441). New York: Guilford.

Velicer, W. F., & Colby, S. M. (2005). Missing data and the general transformation approach to time series analysis. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary Psychometrics. A Festschrift to Roderick P. McDonald* (pp. 509–535). Hillsdale, NJ: Erlbaum.

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

Velicer, W. F., & McDonald, R. P. (1984). Time series analysis without model identification. *Multivariate Behavioral Research*, *19*, 33–47.

Velicer, W. F., & Molenaar, P. (2013). Time Series Analysis. In J. Schinka & W. F. Velicer (Eds.), *Handbook of Psychology: Research Methods in Psychology* (2nd ed., Vol. 2, pp. 628–660). New York: Wiley.

*Waller, R. D., & Higbee, T. S. (2010). The effects of fixed-time escape on inappropriate and appropriate classroom behavior. *Journal of Applied Behavior Analysis*, *43*, 149–153.

*Wilder, D. A., Allison, J., Nicholson, K., Abellon, O. E., & Saulnier, R. (2010). Further evaluation of antecedent interventions on compliance: The effects of rationales to increase compliance among preschoolers. *Journal of Applied Behavior Analysis*, *43*, 601–613.

*Wilder, D. A., Nicholson, K., & Allison, J. (2010). An evaluation of advance notice to increase compliance among preschoolers. *Journal of Applied Behavior Analysis*, *43*, 751–755.

Williams, E. A., & Gottman, J. M. (1982). *A user's guide to the Gottman-Williams time-series analysis computer programs for social scientists* [Computer program manual]. Cambridge: Cambridge University Press.

APPENDIX

An example of syntax SAS v.9.2 procedure implemented to evaluate ITSA parameters using assumed ARIMA (1, 0, 0).

```
procarima;
identifyvar = data esacf p = (0:7) q = (0:7) crosscorr =
(treatment);
estimate p = 1 q = 0 input = (treatment) plot method =
cls;
run;
```

An example of syntax SAS v.9.2 procedure implemented to evaluate ITSA parameters using general transformation ARIMA (5, 0, 0).

```
procarima;
identifyvar = data esacf p = (0:7) q = (0:7) crosscorr =
(treatment);
estimate p = 5 q = 0 input = (treatment) plot method =
cls;
run;
```

"Estimate" indicates the autocorrelation (p) order and moving average (q) order

"Input" and "crosscorr" indicates the design variable, e.g.: baseline phase vs. treatment phase