Available online at www.sciencedirect.com

ScienceDirect

Behavior Therapy 46 (2015) 230–241

Behavior Therapy

www.elsevier.com/locate/bt

Evaluating Statistical and Clinical Significance of Intervention Effects in Single-Case Experimental Designs: An SPSS Method to Analyze Univariate Data

Marija Maric
University of Amsterdam and Research Priority Area Yield, University of Amsterdam

Else de Haan
Academic Medical Center and Academic Center for Child and Adolescent Psychiatry, Amsterdam

Sanne M. Hogendoorn
Academic Center for Child and Adolescent Psychiatry, Amsterdam

Lidewij H. Wolters
Academic Medical Center and Academic Center for Child and Adolescent Psychiatry, Amsterdam

Hilde M. Huizenga
University of Amsterdam, Cognitive Science Center Amsterdam, and Research Priority Area Yield, University of Amsterdam

Single-case experimental designs are useful methods in clinical research practice to investigate individual client progress. Their proliferation might have been hampered by methodological challenges such as the difficulty applying existing statistical procedures. In this article, we describe a data-analytic method to analyze univariate (i.e., one symptom) single-case data using the common package SPSS. This method can help the clinical researcher to investigate whether an intervention works as compared with a baseline period or another intervention type, and to determine whether symptom improvement is clinically significant. First, we describe the statistical method in a conceptual way and show how it can be implemented in SPSS. Simulation studies were performed to determine the number of observation points required per intervention phase. Second, to illustrate this method and its implications, we present a case study of an adolescent with anxiety disorders treated with cognitive-behavioral therapy techniques in an outpatient psychotherapy clinic, whose symptoms were regularly assessed before each session. We provide a description of the data analyses and results of this case study. Finally, we discuss the advantages and shortcomings of the proposed method.

Keywords: single-case experimental designs; univariate data analysis; treatment efficacy; reliable change; clinical research

Determining the efficacy of interventions for adult and youth problems has been one of the most critical areas of research in the contemporary psychology and social sciences. A generally recommended strategy (the so-called *golden standard*) for conducting this type of research is to use group comparison designs—randomized controlled trials (RCTs)—in which outcomes for a group that received a target treatment

The authors wish to thank Marthe Keijman for help with collecting data for this study, the anonymous reviewers, and the editor for their very valuable suggestions, and Joost Agelink van Rentergem for his careful reading of the manuscript.

Address correspondence to Marija Maric, Department of Developmental Psychology, University of Amsterdam, Weesperplein 4, 1018 XA, Amsterdam, the Netherlands; e-mail: m.maric@uva.nl.

are compared with outcomes for others who received either an alternative treatment or were placed on a waiting list (Task Force on Promotion and Dissemination of Psychological Procedures, 1995). Some authors (e.g., Barlow & Hersen, 1984) and the Task Force on Promotion and Dissemination of Psychological Procedures (1995) have long noted that single-case designs could also qualify as either a stand-alone experiment or as a research method complementary to RCTs. Currently, single-case experimental designs (SCEDs) are increasingly recognized as useful methods in clinical research practice to investigate individual clients' progress and to determine whether an intervention works (e.g., Barlow, Nock & Hersen, 2009).

In SCEDs, a single participant is repeatedly assessed on one or multiple indices (e.g., symptoms) during various phases, for example, during baseline, treatment, and follow-up. An important advantage of this method is that it can be used to test novel interventions prior to investigations in demanding and costly RCTs (e.g., Norell-Clarke, Nyander & Jansson-Fröjmark, 2011; Robey & Schultz, 1998). Further, in certain heterogeneous populations of clients, SCEDs may be the only way to investigate treatment outcomes, either because group analyses could lead to a loss of valuable information or because the condition is so rare that the group study would be impossible to conduct within a limited study period (Gaynor & Harris, 2008; Maric, Wiers & Prins, 2012). Finally, SCEDs offer the possibility to systematically document knowledge of researchers and clinicians, thereby preventing loss of information, and also offering the possibility to perform analyses at an aggregate level, combining data from several clients (Kazdin, 2008).

Despite these advantages of SCEDs, methodological challenges may preclude the proliferation of SCEDs in clinical intervention research. One challenge appears to be the analysis of SCED data (Barlow et al., 2009; Smith, 2012). Although visual analysis of such data is common, statistical analysis of such data may be preferred as it is less susceptible to biases (for a review, see Brossart, Parker, Olson & Mahadevan, 2006). Currently, several statistical techniques of single-subject data exist, including nonparametric (e.g., Edgington, 1992) and parametric (e.g., Beeson, & Robey, 2006) procedures. However, most of them may not be easily applied by clinical researchers, which may hamper the proliferation of SCEDs in clinical intervention research (e.g., Barlow et al., 2009; Borckardt et al., 2008; Smith, 2012). Therefore, the purpose of the current paper is to demonstrate how univariate data (i.e., one symptom) obtained from SCEDs can be analyzed in a straightforward manner. Specifically, we illustrate how to analyze SCEDs using the common software package SPSS version 20 to yield information on treatment efficacy (e.g., Is the change in anxious symptoms more pronounced during treatment than during baseline?) and reliable change (e.g., When compared with a normative sample, is there a clinically significant change in anxious symptoms following treatment?).

Below, we first provide a description of our approach to the SCEDs analyses. Second, we outline the results of simulation studies performed to determine the number of observation points required for each phase in a SCED analysis. Third, we provide a detailed description of a single case of an adolescent diagnosed with anxiety disorders treated with cognitive-behavioral therapy (CBT) techniques in an outpatient psychotherapy clinic. Finally, we illustrate how to apply the SCED analytic methods to the data from the single case.

Method

The most frequently used SCED in clinical research is the AB design (Barlow et al., 2009). It consists of two phases (i.e., a baseline and a treatment), and can be conceptualized as an interrupted time series (Campbell & Stanley, 1966). To obtain an adequate analysis of differences between baseline and treatment in such an interrupted time series, two requirements should be fulfilled. First, the overall pattern in the time series has to be modeled adequately. A common assumption is that an adequate model consists of two linear functions: one for baseline and one for the treatment phase (e.g., Center, Skiba & Casey, 1985; Jones, Vaught & Weinrott, 1977; Kelly, McNeil & Newman, 1973). Each of these linear functions is described by an intercept (the symptom score if time_in_phase [time points within each phase] is zero), and a slope (the change of symptom score in a phase). If the overall pattern in the data is not modeled adequately, for example, if it is assumed that slopes of the two phases are equal whereas they actually differ, estimators of intercepts and slopes will be biased and thus cannot be given a meaningful interpretation. Therefore, adequate modeling of the overall pattern is a first requirement for an adequate analysis of phase differences.

A second requirement is adequate modeling of potential correlations between residuals. That is, adequate modeling of that which remains after the overall pattern has been accounted for (Huitema & McKean, 1998). This correlation between residuals of the observations (e.g., anxiety symptoms) has been termed "autocorrelation," and it implies that the residuals are not independent. For example, a lag 1 correlation refers to the degree in which the

residual of anxiety ratings at time points 1 and 2 or time points 3 and 4 are correlated; as another example, a lag 3 correlation refers to the degree in which residuals of anxiety ratings at time points 1 and 4 are correlated. If residuals are correlated, the correlations are likely to decrease with increasing separation between time points, as described, for example, by the commonly used first-order autoregressive model (Levin, Ferron & Kratochwill, 2012).

In more statistical terms, if correlations between residuals decrease with the increasing lag between observations, the residual correlation matrix might be modeled adequately by an AR(1) correlation matrix. Suppose that there are three observations; the correlation matrix then will be

$$\begin{bmatrix} rho^0 & rho^1 & rho^2 \\ rho^1 & rho^0 & rho^1 \\ rho^2 & rho^1 & rho^0 \end{bmatrix},$$

where *rho* equals the AR(1) parameter, ranging between –1 and +1. If time points are separated by one (lag 1), *rho* is raised to the power of 1; if they are separated by two (lag 2), *rho* is raised to the power of 2, and so on, thus generating correlations that decrease with increasing lag. If correlations between residuals are not adequately modeled, for example, if it is incorrectly assumed that they are uncorrelated, tests on intercepts and slopes will be unreliable. Importantly, if positively correlated residuals are assumed to be uncorrelated, chances of finding significant results will be too high (Brossart et al., 2006). This may have important implications, for example, it might then be erroneously concluded that treatment is beneficial, whereas this actually is not the case. In the following, while describing the analyses in a more conceptual way, we turn to several of these issues in more detail.

ANALYSES INVESTIGATING TREATMENT EFFICACY

If the overall pattern is modeled by an intercept and slope for each phase, this does not yield a direct test of intercept and slope differences between phases. Therefore, it is convenient to re-parameterize the model in terms of an intercept and slope for the baseline phase and baseline-treatment *differences* in intercepts and slopes. The model is described by the following function (e.g., Center, Skiba & Casey, 1985; Jones, Vaught & Weinrott, 1977; Kelly, McNeil & Newman, 1973):

$$\begin{aligned} y(i) = {} & b_0 + b_1 * \text{phase}(i) + b_2 * \text{time\_in\_phase}(i) \\ & + b_3 * \text{phase}(i) * \text{time\_in\_phase}(i) + e(i) \quad (1) \end{aligned}$$

In Equation (1), $y(i)$ denotes the outcome variable score at time point i, phase(i) denotes the phase in which time point i is contained (coded as 0 for baseline and 1 for treatment), and time_in_phase denotes time points within each phase. The term $e(i)$ denotes the residual at time point i. The parameter b_0 is interpreted as the baseline intercept, b_1 as the treatment–baseline difference in intercepts, b_2 as the baseline slope, and b_3 as the treatment–baseline difference in slopes. Consequently, intercept differences between phases can be directly assessed by testing whether b_1 differs significantly from 0; analogously, slope differences can be assessed by testing b_3. Note that these parameter estimates can also be interpreted as an effect size (Cumming, 2014).

The baseline intercept b_0 and the baseline–treatment difference in intercept b_1 refer to symptom scores when time_in_phase is zero. Therefore, interpretation of b_0 and b_1 depends on the coding of time_in_phase. For example, if time_in_phase is coded such that it is zero at the start of each phase, b_1 indexes symptom score differences between the start of baseline and start of treatment. However, it may be more convenient to code time_in_phase such that it is zero at the end of each phase. A test on b_1 then indicates whether symptom scores differ between the end of baseline and the end of treatment, thereby providing a convincing test of treatment efficacy.

Note that this parameterization yields a test of change in the first phase and a test of the difference in change between the first and second phases. Reversing the coding of the phase variable (1 = exposure [EXP], 0 = EXP + cognitive therapy [CT]; see Table 1) and computing a new interaction term accordingly, yields a test of change in the second phase in addition to a test of the difference in change between phases.

Some controversy was raised by the question of whether residuals in interrupted time series designs are correlated, and thus whether it is necessary to model them (see Levin et al., 2012, for a review of this controversy). On the one hand, it has been argued that some studies have reported inflated correlations as the overall pattern was not modeled adequately (cf. Center et al., 1985; Huitema & McKean, 1998), for example, by incorrectly assuming that baseline and treatment slopes were zero. On the other hand, it has been argued that studies may have incorrectly concluded that correlations are nonsignificant, as tests of correlations are underpowered for short time series (Busk & Marascuilo, 1992). For these reasons, the best approach seems to be to both describe the general trend adequately *and* to account for remaining correlations, even if tests on these correlations indicate that they are nonsignificant.

These correlations can be quantified by the AR(1) parameter *rho* incorporated in the commonly used

Table 1
Structure of Data Array in SPSS

Participant	Time (weekly sessions)	Phase (0 = EXP, 1 = EXP + CT)	Time in phase	Phase × Time in phase [a]	ASICA anxious feelings	ASICA anxious thoughts	ASICA avoidance	CATS negative automatic thoughts	CATS positive automatic thoughts	SEQ-C self-efficacy
1	1	0	3	0	12	14	4	8	2	1
1	2	0	2	0	7	13	7	6	4	3
1	3	0	1	0	10	12	9	4	4	3
1	4	0	0	0	7	10	0	2	6	5
1	5	1	3	3	11	11	0	5	5	5
1	6	1	2	2	2	8	0	3	6	9
1	7	1	1	1	6	8	0	5	8	11
1	8	1	0	0	6	4	0	3	9	14

[a] This variable can be omitted from the data array in SPSS as it is automatically calculated. EXP = exposure; CT = cognitive therapy; ASICA = Anxiety Severity Interview for Children and Adolescents; CATS = Children's Automatic Thoughts Scale; SEQ-C = Self-Efficacy Questionnaire for Children.

first-order autoregressive model. This parameter may be underestimated for short time series (Marriott & Pope, 1954) and therefore, tests on the parameters in model Equation (1) may remain too liberal (Crosbie, 1993). Three solutions might be adopted to alleviate this potential problem. The first solution is to apply a small sample correction to the estimated AR(1) parameter (Crosbie, 1993; cf. Marriott & Pope, 1954). A second solution is to test parameters by implementing a permutation-based procedure (Edgington, 1967). The third, in short time series, is that tests may be carried out at more stringent levels, thereby reducing the likelihood of finding spurious results.

As will be illustrated in the Results section, the analysis described above, fitting Equation (1) to the data while accounting for autocorrelation, can easily be performed in the SPSS mixed-models module (SPSS Inc., 2011). However, before that, we now turn to describing the analyses that can be used to test reliable change.

ANALYSES INVESTIGATING RELIABLE CHANGE

Note that due to the specific coding of the phase and the session_in_phase variables, the intercept b_0 gives an estimator of the end point of the "baseline" phase, whereas $b_0 + b_1$ gives an estimator of the end of the "treatment" phase. Therefore, b_0 and $b_0 + b_1$ can be used to estimate reliable change. If the linear change model is correct, these estimators are likely to be more precise estimators of final outcomes than observed final outcomes, as they are less contaminated by error (see Figure 2, the difference between observations and regression lines), just as a mean is less contaminated by error than individual observations.

Jacobson and Truax (1991) define the Reliable Change Index (RCI) as follows:

$$\mathrm{RCI} = \mathrm{X2}–\mathrm{X1}/\mathrm{Sdiff}, \quad (2)$$

in which X1 and X2 represent a participant's pre- and posttreatment scores, respectively, and where Sdiff denotes the standard deviation of the difference between the two test scores (Jacobson & Truax, 1991). Instead of using the observed end points of phases, we suggest including the estimated end points as defined by b_0 and $b_0 + b_1$; therefore, the RCI is defined by:

$$\mathrm{RCI} = [(b_0 + b_1)–b_0]/\mathrm{Sdiff} = b_1/\mathrm{Sdiff} \quad (3)$$

To determine how much an individual has benefited from treatment, it would be useful to know the magnitude of the pre- to posttreatment effect (i.e., an effect size). However, there have been, to our knowledge, no proposals for effect sizes for reliable change indices, and therefore, also no guidelines exist on what constitutes a small, medium, or large effect.

In sum, we propose a procedure that combines the strengths of several approaches that have previously been described in the literature on single-subject analysis. First, within one integrated model, we model the general trend in the data adequately by means of two separate regression lines. Second, we formulate the regression model in such a manner that differences between the ends of phases can be tested directly. Third, we account for correlated residuals, thereby increasing the reliability of test results. Fourth, we provide an RCI. Finally, and most important, we show that this procedure can be carried out straightforwardly in the common

statistical package SPSS, which is often available to researchers in clinical settings.

The two main assumptions of this analysis are that the overall pattern in the data is modeled correctly and that the error correlation structure is modeled adequately. For example, if time effects are curvilinear instead of linear, the analysis will yield biased results. In addition, if the correlation structure is not modeled adequately, that is, if correlations do not follow an AR(1) structure but another pattern, test results may become less reliable. An additional assumption, which will not be discussed any further, is that if data are missing, they should be missing at random (Schafer & Graham, 2002). Another assumption is that the number of time points is sufficient to estimate the AR(1) parameter adequately. We therefore performed simulation studies to determine the number of observation points required per intervention phase.

SIMULATION STUDY

Simulation studies were performed to investigate the number of time points required to arrive at an acceptable Type 1 error rate for these analyses. That is, we simulated data with no effects present (all *b*'s zero) and added noise that was generated by an AR(1) process. From these simulated data we estimated the four parameters and tested whether they differed significantly from zero using a nominal alpha of 5%. We repeated this procedure 10,000 times, and scored for each parameter the number of significant outcomes, which should occur only in 5% of the cases. We varied the number of time points per phase (5, 10, 15, 20, or 100) and the degree of autocorrelation (*rho* of .1, or .3). The simulation study was carried out in R Statistical Package (*R* 3.0.2), parameter estimates were derived by using REML implemented in the gls function of the nlme package, which yields equivalent estimates as the proposed SPSS procedure.

In Figure 1a, it can be seen that Type 1 errors decrease with an increasing number of time points. They are reasonably accurate if the number of time points per phase is 10 or more, but Type 1 errors are about 10% if there are only 5 time points per phase. In addition, it can be seen that Type 1 error slightly increases with increasing autocorrelation. We therefore conclude that the procedure is reliable if the number of time points per phase is 10 or more, but too liberal if the number of time points is only 5. If there are only 5 time points per phase, it is better to test against a more stringent nominal alpha of 1%, which will yield a satisfactory percentage of false positives (see Figure 1b).

In the following section we present a detailed description of a single adolescent treated with CBT techniques. Thereafter, we apply the proposed SCED analytic methods to these data.

Case Example

The participant was a 17-year-old girl who met inclusion and exclusion criteria[1] for a broader study investigating the additional value of CT over and above EXP therapy for youth anxiety. The SCED included multiple pretreatment assessments (A phase = baseline) followed by four sessions of behavioral interventions (B phase = EXP) and four sessions of behavioral interventions and CT (C phase = EXP + CT). The study was conducted jointly by the University of Amsterdam, Developmental Psychology, and de Bascule in Amsterdam, Academic Centre for Child and Adolescent Psychiatry. Both youth and parents provided written informed consent allowing their data to be used for the purposes of scientific research. At pretreatment, the participant met criteria for social phobia (Clinical Severity Rating [CSR][2] = 6), generalized anxiety disorder (CSR = 6), and specific phobia for vomiting and doctors (CSR = 5), assessed via administration of the Anxiety Disorders Interview Schedule for Children/Parents (ADIS-C/P; Silverman & Albano, 1996).

In order to assess weekly levels of anxiety symptoms, the Anxiety Severity Interview for Children and Adolescents (ASICA; Hogendoorn, De Haan, et al., 2013) was administered. The ASICA is a 15-question semistructured clinical interview developed to assess anxious feelings, anxious thoughts, and avoidance on a regular basis. Previous evaluations of the ASICA showed that internal consistency (Cronbach's α) was good for the total score (.86), and was moderate to good for the three subscales: Feelings (.64), Avoidance (.80), and Thoughts (.78). The interview has moderate to good test–retest reliability (ICC's ranging from .66 to .85), and it has been shown to be sensitive to treatment change (Hogendoorn, De Haan, et al., 2013). In addition, negative and positive automatic thoughts were measured using the Children's Automatic Thoughts Scale–Negative/Positive (CATS-N/P; Hogendoorn et al., 2010). In a previous study, sound psychometric properties were observed for the

[1] Inclusion criteria: ages between 8 and 18, primary anxiety disorder (except posttraumatic stress disorder or obsessive–compulsive disorder), had not received protocolized evidence-based CBT in the past half year, no use of selective serotonin reuptake inhibitors (SSRIs) at the moment of intake and treatment. Exclusion criteria: suicidal ideation, psychosis, selective mutism, IQ below 80, problems with drugs or alcohol.

[2] The CSR scale consists of a 9-point Likert scale (0–8) and allows the clinician to evaluate the severity of each diagnosed condition. A score of 4 or above indicates the presence of a clinically significant disorder.

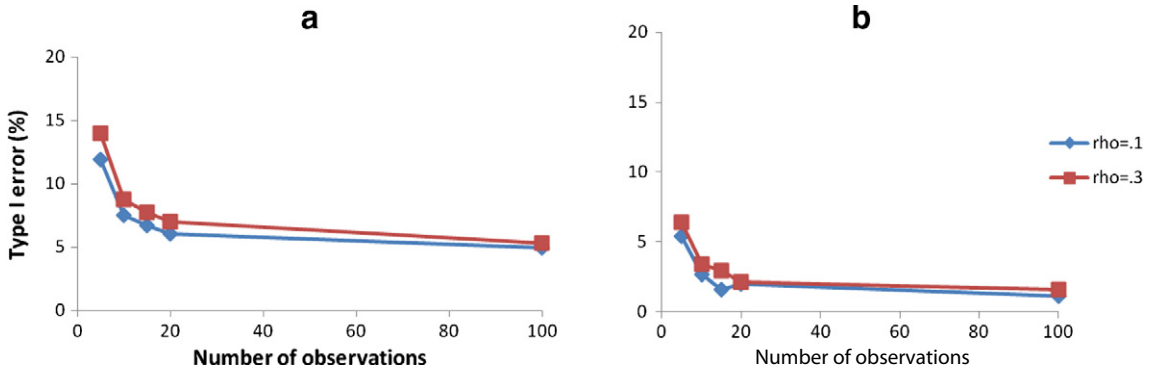


FIGURE 1 Simulation study results: False positives using a nominal alpha of 5% (1a) and 1% (1b). *rho* = autcorrelation parameter.

CATS-N/P in youth ages 8 to 18 years. Good internal reliability (Cronbach's α.83–.94) and factorial validity, and moderate test–retest reliability (Pearson's *r* .61–.77 for the total score) were reported. The questionnaire appeared to be sensitive to treatment change (Hogendoorn, Prins, et al., 2013). Self-efficacy was assessed using the Self-Efficacy Questionnaire for Children (SEQ-C; Muris, 2001), which measures social, academic and emotional self-efficacy. Internal consistency of the SEQ-C is good (Cronbach's α.85–.88).

The ASICA, CATS-N/P, and the SEQ-C were administered in full at pre-, mid-, and posttreatment. Furthermore, prior to each treatment session, the ASICA (first three items from each subscale) was administered, as well as the top three to four items (i.e., highest or lowest scores of the participant at pretreatment) from each questionnaire (CATS-N/P, SEQ-C). For this case, items 20, 26, and 41 were administered from the CATS-Negative subscale; items 3, 19, and 21 from the CATS-Positive subscale; and items 5, 9, 21, and 24 from the SEQ-C questionnaire. We utilized this client-guided, idiographic approach for two reasons. The more practical one concerns the fact that the amount of assessment that can be administered frequently on a session basis is limited, as frequent assessments may be demanding for (young) clients (Stice, Rohde, Seeley & Gau, 2010). The more conceptual reason for choosing top items concerns that we aimed to focus attention on problems that youth considered most important to be working on during treatment (Weisz et al., 2011). In this way, specific evidence can be generated on trajectories of change in those problems during treatment. A trained research assistant conducted all assessments.

At the end of each treatment session, the therapist registered the in-session activities using a checklist. To ensure adherence to therapy protocol, the first (Maric) and second (Haan) authors scored the content of each session using the session reports. The interrater reliability was good (Kappa = .81, $p < 0.001$).

The participant completed the entire eight-session treatment and all pre-, during, and posttreatment assessments. There were no missing data. To illustrate our method, the data from the two treatment phases were used: EXP and EXP + CT.

Results

Table 1 illustrates the structure of the data array. The SPSS file used for the actual data analysis should be organized in the same manner. The first column contains the participant identifier. The second column contains session number. The third column indicates whether the session consisted of EXP (Weeks 1–4, coded as 0) or whether CT was added (coded as 1). The fourth column shows the session number within each phase; note that this index runs backward (i.e., 3 at the beginning of a phase, and 0 at the end of a phase). The fifth column contains the interaction term of phase and session number within each phase. These values are automatically calculated by SPSS and can be omitted from the data array. Columns 6–11 record the symptom scores (ASICA-anxious feelings, ASICA- anxious thoughts, ASICA-avoidance, CATS-negative automatic thoughts, CATS-positive automatic thoughts, SEQ-C) registered prior to each session.

TREATMENT EFFICACY

To give an answer to the question of treatment efficacy (i.e., Is the change in anxious symptoms more pronounced during EXP + CT than during the EXP?), the analysis described in the Method section, fitting Equation (1) to the data while accounting for autocorrelation, has been performed in the SPSS mixed-models module (S.P.S.S. Inc., 2011) and is illustrated in the following video: http://youtu.be/sYGOynx-J8M.

The results for our case example can be found in Table 2 and in Figure 2. Using SEQ-C results as an example, the interpretation of each parameter is as follows: The parameter b_0 is the self-efficacy score at the end of EXP. As the EXP phase was coded 0

Table 2
Treatment Efficacy Results Using Mixed Models

a. Results Concerning Anxious Feelings (ASICA)

	Estimate	*SE*	*p*	95% Confidence Interval	
				Lower bound	Upper bound
Intercept (b_0)	7.96	2.17	.287	-141.76	157.68
Phase (b_1)	-3.28	2.99	.572	-377.35	370.80
Time_in_phase (b_2)	0.76	1.21	.666	-31.96	33.49
Time_in_phase * phase (b_3)	0.11	1.62	.963	-140.95	141.16

ASICA = Anxiety Severity Interview for Children and Adolescents.

b. Results Concerning Anxious Thoughts (ASICA)

	Estimate	*SE*	*p*	95% Confidence Interval	
				Lower bound	Upper bound
Intercept (b_0)	10.47	0.19	**.000**	9.71	11.23
Phase (b_1)	-5.32	0.27	**.001**	-6.34	-4.29
Time_in_phase (b_2)	1.24	0.12	**.005**	0.80	1.68
Time_in_phase * phase (b_3)	0.54	0.15	.054	-0.02	1.10

ASICA = Anxiety Severity Interview for Children and Adolescents.

c. Results Concerning Negative Automatic Thoughts (CATS)

	Estimate	*SE*	*p*	95% Confidence Interval	
				Lower bound	Upper bound
Intercept (b_0)	2.19	0.26	.013	1.09	3.29
Phase (b_1)	1.62	0.36	.042	0.14	3.09
Time_in_phase (b_2)	1.9	0.15	**.005**	1.29	2.51
Time_in_phase * phase (b_3)	-1.75	0.20	**.010**	-2.55	-0.96

CATS = Children's Automatic Thoughts Scale.

d. Results Concerning Positive Automatic Thoughts (CATS)

	Estimate	*SE*	*p*	95% Confidence Interval	
				Lower bound	Upper bound
Intercept (b_0)	5.7	0.21	**.004**	4.51	6.88
Phase (b_1)	3.45	0.29	.016	1.82	5.08
Time_in_phase (b_2)	-1.12	0.12	.017	-1.71	-0.53
Time_in_phase * phase (b_3)	-0.29	0.16	.231	-1.14	0.56

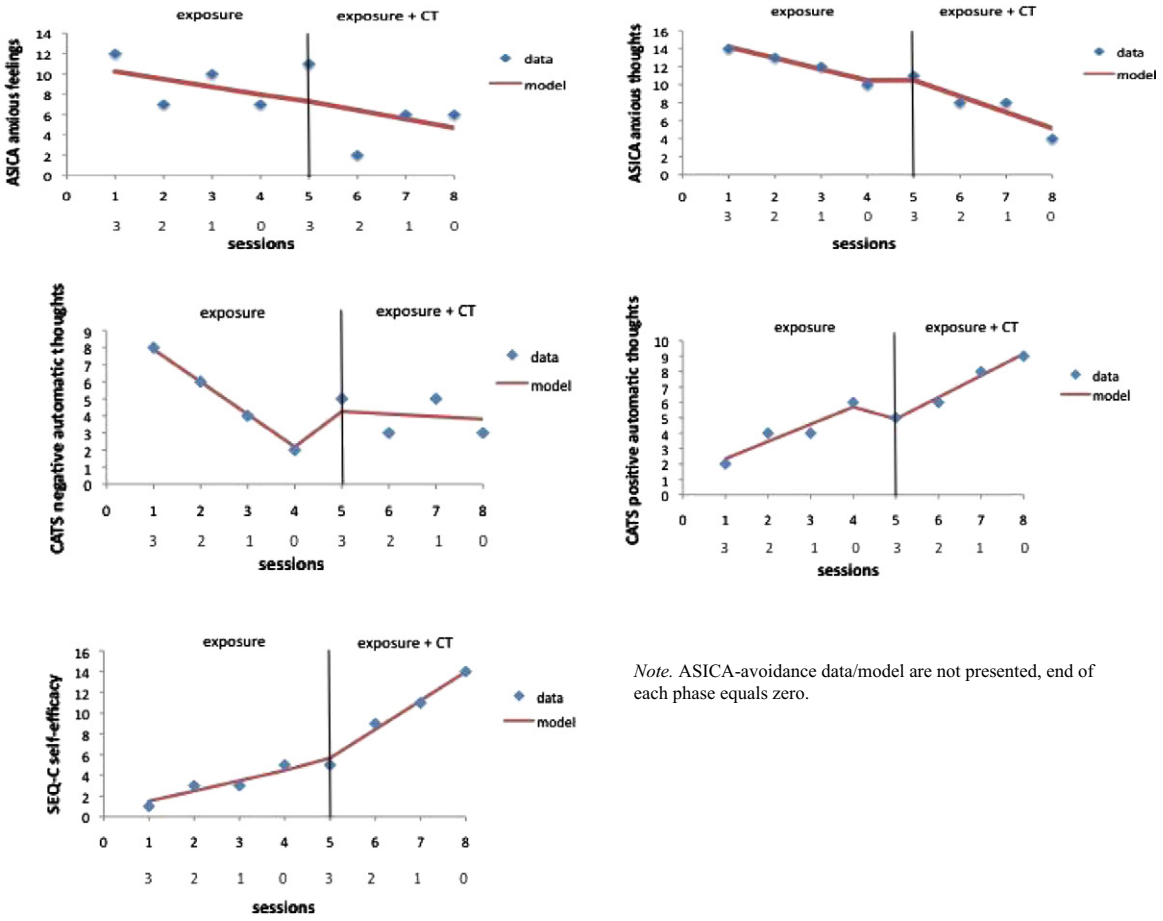
CATS = Children's Automatic Thoughts Scale.

e. Results Concerning Self-Efficacy (SEQ-C)

	Estimate	*SE*	*p*	95% Confidence Interval	
				Lower bound	Upper bound
Intercept (b_0)	4.47	0.14	**.000**	3.96	4.99
Phase (b_1)	9.51	0.19	**.000**	8.81	10.21
Time_in_phase (b_2)	-0.99	0.08	**.003**	-1.29	-0.70
Time_in_phase * phase (b_3)	-1.78	0.11	**.001**	-2.17	-1.40

Note. SEQ-C = Self-Efficacy Questionnaire for Children.
Note. *p* values in bold indicate significant estimates. Nominal alpha = .01.

and the EXP + CT phase was coded 1, b_1 is the difference in self-efficacy between end of EXP + CT and end of EXP, with positive values indicating higher levels of self-efficacy during EXP + CT. The interpretation of b_2 is the rate of change in the EXP phase. As time runs backward from the end of each phase, negative values indicate an increase in self-efficacy during the EXP phase. The interpretation of b_3 is the difference in rates of change in self-efficacy between EXP + CT and EXP. This parameter should be interpreted in combination with parameter b_2. That is, if both b_3 and b_2 estimates



Note. ASICA-avoidance data/model are not presented, end of each phase equals zero.

FIGURE 2 Visual representation of the models of the differences between the phases (exposure = sessions 1–4, and exposure + CT = sessions 5–8) on symptoms scores.

are negative (as in our self-efficacy example), this indicates that there is a higher increase in self-efficacy during the EXP + CT phase than in the EXP phase. If b_2 is negative and b_3 is positive, this indicates that there is a lower increase (in case the absolute value of b_3 is lower than that of b_2) or even a decrease (in case the absolute value of b_3 exceeds that of b_2) in self-efficacy during the EXP + CT phase. As our simulation study indicated that for approximately 5 time points or less per phase it is better to use a nominal alpha of .01 instead of .05, we use this criterion for significance in the analyses reported below.

The results of the weekly assessments indicated that after the EXP + CT, as compared with EXP, the end point of anxious thoughts (ASICA) was lower ($p < .01$), and the end point of self-efficacy was higher ($p < .01$). There were no significant phase differences in end points of anxious feelings and negative automatic thoughts and positive automatic thoughts (Table 2).

Regarding differences in slopes between the two phases, self-efficacy ($p < .01$) increased at a higher rate during EXP + CT, as compared with EXP. There were no significant phase differences in rates of changes in anxious feelings, anxious thoughts, negative automatic thoughts, and positive automatic thoughts. However, negative automatic thoughts showed a trend toward significance ($p = .01$), suggesting a higher decrease in this symptom during EXP, as compared with the EXP + CT phase.

RELIABLE CHANGE

As indicated in the Method section, we used estimated instead of observed phase end points to assess reliable change. Based on an alpha = 0.05 (two-tailed significance testing), an RCI >+1.96 or <–1.96 indicates a statistically significant reliable change (Jacobson & Truax, 1991). As shown in Table 3, two RCIs indicated a clinically significant decrease in anxious feelings and thoughts from the end of EXP to the end of EXP + CT.

Discussion and Conclusion

In the current study we presented an SPSS method to investigate treatment efficacy and reliable change

Table 3
Reliable Change Index (RCI) Results for Differences Between the Phases in Self-Report of Symptoms

	ASICA anxious feelings	ASICA anxious thoughts	CATS negative automatic thoughts	CATS positive automatic thoughts	SEQ-C self-efficacy
Reliability information [a]	.68	.71	.61	.56	.88
SD [b]	1.52	1.41	2.54	2.53	11.2
b_1 [c]	-3.28	-5.32	1.62	3.45	9.51
RCI	-2.69	-4.97	0.72	1.46	1.73

[a] For the Anxiety Severity Interview for Children and Adolescents (ASICA) scales test–retest reliability (ICC) of the first three items of each scale is used. For the Children's Automatic Thoughts Scale (CATS) negative and positive subscale test–retest reliability (Pearson's *r*) of the three items from each subscale was used. For the Self-Efficacy Questionnaire for Children (SEQ-C) Chronbach's alpha of the whole questionnaire was used (test–retest reliability information of four selected items was unavailable).
[b] Based on information from nonclinical samples.
[c] $(b_1) = ([b_0 + b_1] - [b_0])$.

in SCEDs. The results from the treatment efficacy analyses can provide information regarding difference in within-client symptom changes between two different phases in SCEDs (e.g., changes in anxiety levels between baseline and treatment phases). Additionally, results from the reliable change analyses can provide valuable information about whether this change in anxious symptoms within the client is significant in comparison to the scores of a norm group. Some additional information regarding the implications and optimal implementation of the method might be helpful for clinical researchers. In this section, this information is presented followed by a short discussion of the results based on the analyses of the case example data used in this study. In concluding paragraphs, several benefits of the SCEDs for clinical research practice are highlighted.

With regard to utilization of the current method, the method can be generalized in various ways. First, it can be generalized to more than two phases such as ABAC (baseline–treatment B–baseline–treatment C) designs. In this case, additional phase variables and their interaction with time_in_phase should be added. Second, as the method is also valid if some session data are missing at random (Schafer & Graham, 2002), the method can also be used if symptoms are probed at irregular time intervals. Third, longer-term follow-up data can be seen as an additional phase. Finally, using meta-analytic techniques, results across participants can be combined, that is, by weighting intercept or slope differences by their respective standard errors (Huizenga, Visser & Dolan, 2011; Van den Noortgate & Onghena, 2008).

Several points regarding this method merit further consideration. First, the methodology requires assumptions. The two main assumptions are that the overall pattern in the data as well as the correlation structure are modeled adequately. A third assumption is that if data are missing, they should be missing at random. These assumptions also hold for other methods proposed for SCED analysis and therefore are not unnecessarily restrictive. Second, the results of our simulation study indicate that the method is most reliable in the case of 10 or more time points per phase. Note, however, that these simulations only considered a few circumstances, that is, an autocorrelation parameter of .1 and .3, equal number of time points per phase, and no missing values. Therefore, these results should be treated as tentative rules of thumb instead of fixed guidelines. Third, and based on the results from the simulation study, if there are approximately 5 time points or less per phase, it is necessary to choose a more stringent nominal alpha to arrive at reliable results. Alternative approaches are applying a correction (Crosbie, 1993; Marriott & Pope, 1954) or resampling techniques (Edgington, 1992). One approach to resampling has been described by Borckardt et al. (2008). However, the latter method is based on very specific hypotheses such as "during treatment the symptom score in every subsequent session will be decreased by 2 (a slope of –2)." We believe that it is more likely that clinical researchers will specify more general hypotheses such as "during treatment the symptom score in every subsequent session will be decreased," as is being tested in the current methodology. Fourth, the analytic procedure for single-subject data described in the current paper provides indicators of baseline–treatment differences in univariate, but not multivariate, outcome variables. Therefore, it does not allow to test whether treatment-induced change in one outcome variable is preceded by that in another. In order to answer such questions, more sophisticated structural equation modeling-based techniques, like dynamic factor analysis (Gayles & Molenaar, 2013; Molenaar, 1985), are required. Fifth, our results (especially with regard to reliable change) seem to confirm the importance of selecting assessment tools that are specifically designed to assess clinical symptoms on a regular basis (such as ASICA) and/

or instruments that have sound psychometric properties concerning sensitivity to treatment change. In the absence of significant results, one may incorrectly conclude that no clinically significant change has happened following treatment; the actual explanation could be that the questionnaire had low levels of sensitivity to treatment change. Finally, specifically related to the RCI, reliability information of the measures should ideally be selected from norm groups that are as similar as to the client groups evaluated in the study. For example, in the case of our data, it would have been better if reliability information was available from norm groups concerning adolescents diagnosed with anxiety disorders.

Although the present methodology is user-friendly, as it uses the common package SPSS, two factors may hamper its routine application in clinical practice. First, data collection may be too time consuming given the high workload of many clinicians. Note, however, that modern data collection methods using tablets or smartphones may be very facilitative in this respect (e.g., Lambert, Harmon, Slade, Whipple & Hawkings, 2005). Second, in some settings with an academic affiliation, clinicians may not have access to SPSS due to cost considerations, whereas in others where SPSS is available, clinicians may not use it because of time constraints. SPSS may furthermore not be available within institutions without academic affiliations. Therefore, future work may focus on the development of Web applications that incorporate single-case data analytic techniques that may even further increase application of these methods.

In the current study, we illustrated the application of this method using real-life data of an anxious adolescent treated with CBT techniques. The results suggest that, for this client, CT showed an additive effect over and above EXP, which was evidenced by an enhanced change and an enhanced end point in self-efficacy. In addition, the end point of anxious thoughts was lower after EXP + CT. In comparison with the norm group of children and adolescents, anxious feelings and thoughts showed a clinically significant decrease from end of EXP to the end of the EXP + CT phase. However, it should be noted that the effects found in this study at the EXP + CT phase might also be due to prolonged exposure. Future studies should consider reversing the order of CBT components, that is, providing "CT only" to the clients followed by additional "EXP therapy" techniques or using a design less confounded by the prolonged effects of the first intervention phase, that is, the "interaction element design" (Hayes, Barlow & Nelson-Gray, 1999). In this design, different treatment components are systematically added or subtracted after the baseline phase (e.g., baseline/EXP/EXP + CT/EXP or baseline/CT/CT + EXP/CT).

In the context of CBT treatment component analyses, SCEDs may be beneficial in several respects. First, after 50 years of research into CBT for youth anxiety, we know that CBT as a whole treatment package is efficacious, but little is known about the relative efficacy of various components of CBT (Maric et al., 2012). Testing such innovative research questions in RCTs can be challenging, given the number of participants, and properly trained therapists needed for such a study (Ougrin, 2011). Small-scale research such as SCEDs may be a useful alternative in this case, either as a stand-alone method or as a first step in generating a specific hypothesis to be tested in RCTs. Second, the extent of individual benefit from specific CBT components may vary from client to client. Some may respond better to EXP or CT, and some may do better with a combination. An additional merit of SCEDs is thus the ability to discover substantially diverse treatment effects in different individuals. This information might be lost if analyses are done on a group level only (Gaynor & Harris, 2008).

More generally, the SCEDs can be used to bridge the gap between science and practice (Borckardt et al., 2008; Kazdin, 2008; Task Force on Promotion and Dissemination of Psychological Procedures, 1995). The so-called top-down approach would involve laboratory studies informing clinical practice. Laboratory researchers could evaluate treatment protocols using SCEDs, and could inform clinicians about the most efficacious techniques to be used with certain clients (as a hypothetical example, EXP may be sufficient to alleviate anxiety in children, but adolescents need an additional CT component). Similar studies conducted by clinicians could inform laboratory researchers (the so-called bottom-up approach), for example, when the clinician uses Internet-based assessment systems and tracks symptoms before, during, and after interventions with clients characterized by heterogeneity that are often excluded from RCTs (e.g., concurrent anxiety, attention deficit/hyperactivity disorder [ADHD], and conduct disorders). Researchers could then further evaluate this evidence on effectiveness of treatment in comorbid conditions, for example, they can evaluate the efficacy of different interventions for these clients. The findings might guide future RCTs, for example, they may offer suggestions for the type, ordering, and duration of different treatment techniques needed to alleviate ADHD, anxiety, and conduct disorder symptoms.

Conclusion

In this article we presented a user-friendly method to analyze univariate data from Single-Case Experimental Designs using SPSS. The method can be used

to evaluate both statistical as well as clinical significance of treatment effects. The development of such methods seems a necessary path to be taken not only to assist science-practitioners in their evaluations of treatment progress, but also to improve current SCED guidelines and reporting standards. SCEDs offer an excellent opportunity to gain information from therapy about what works and how it works for specific clients, and what might work in future similar cases.

Conflict of Interest Statement

The authors declare that there are no conflicts of interest.

References

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

Barlow, D. H., Nock, M. K., & Hersen, M. (2009). *Single case experimental designs. Strategies for studying behaviour change* (3rd ed.). Boston, MA: Allyn and Bacon.

Beeson, P. M., & Robey, R. R. (2006). Evaluating single-subject treatment research: Lessons learned from the Aphasia literature. *Neuropsychology Review*, *16*, 161–169.

Borckardt, J. J., Nash, M. R., Murphy, M. D., Moore, M., Shaw, D., & O'Neil, P. (2008). Clinical practice as natural laboratory for psychotherapy research. *American Psychologist*, *63*, 77–95. http://dx.doi.org/10.1037/0003-066X.63.2.77

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. http://dx.doi.org/10.1177/0145445503261167

Busk, P. L., & Marascuilo, R. (1992). Statistical analysis in single-case research: Issues, procedures, and recommendations, with applications to multiple behaviors. In T. R. Kratochwill & J.R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education* Hillsdale, NJ: Erlbaum.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Chicago, IL: Rand McNally.

Center, B., Skiba, R., & Casey, A. (1985). A methodology for the quantitative synthesis of intra-subject design research. *Journal of Special Education*, *19*, 387–400. Retrieved from http://sed.sagepub.com/content/19/4/387

Crosbie, J. (1993). Interrupted time-series analysis with brief single-subject data. *Journal of Consulting and Clinical Psychology*, *61*, 966–974. Retrieved from http://psycnet.apa.org/journals/ccp/61/6/966/

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7–29.

Edgington, E. (1967). Statistical inference from $N = 1$ experiments. *Journal of Psychology*, *65*, 195–199. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/00223980.1967.10544864

Edgington, E. (1992). Nonparametric tests for single-case experiments. In T. R. Kratochwill & J.R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp. 133–157). Hillsdale, NJ: Erlbaum. Retrieved from http://psycnet.apa.org/psycinfo/1992-98100-004

Gayles, J. G., & Molenaar, P. C. M. (2013). The utility of person-specific analyses for investigating developmental processes: An analytic primer on studying the individual. *International Journal of Behavioral Development*, *37*, 549–562. http://dx.doi.org/10.1177/0165025413504857

Gaynor, S. T., & Harris, A. (2008). Single-participant assessment of treatment mediators: Strategy description and examples from a behavioral activation intervention for depressed adolescents. *Behavior Modification*, *32*, 372–402.

Hayes, S. C., Barlow, D. H., & Nelson-Gray, R. O. (1999). *The scientist practitioner: Research and accountability in the age of managed care* (2nd ed.). Boston, MA: Allyn and Bacon.

Hogendoorn, S. M., de Haan, E., Wolters, L. H., Vervoort, L., Prins, P. J. M., De Bourgraaf, A., . . . Goodman, W. K. (2013). The Anxiety Severity Interview for Children and Adolescents: An individualized repeated measure of anxiety severity. *Clinical Psychology and Psychotherapy*. Published online ahead of print. http://dx.doi.org/10.1002/cpp.1863

Hogendoorn, S. M., Prins, P. J. M., Boer, F., Vervoort, L., Wolters, L. H., Moorlag, H., . . . de Haan, E. (2013). Mediators of cognitive behavioral therapy for anxiety disordered children and adolescents. *Journal of Clinical Child and Adolescent Psychology*, published online ahead of print. http://dx.doi.org/10.1080/15374416.2013.807736

Hogendoorn, S. M., Wolters, L. H., Vervoort, L., Prins, P. J. M., Boer, F., Kooij, E., & de Haan, E. (2010). Measuring negative and positive thoughts in children: An adaption of the Children's Automatic Thoughts Scale (CATS). *Cognitive Therapy and Research*, *34*, 467–478.

Huitema, B. E., & McKean, J. W. (1998). Irrelevant autocorrelation in least-squares intervention models. *Psychological Methods*, *3*, 104–116. http://dx.doi.org/10.1037//1082-989X.3.1.104

Huizenga, H. M., Visser, I., & Dolan, C. V. (2011). Hypothesis testing in random effects meta-regression. *British Journal of Mathematical and Statistical Psychology*, *64*, 1–19.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *59*, 12–19.

Jones, R., Vaught, R. S., & Weinrott, M. (1977). Time-series analysis in operant research. *Journal of Applied Behavior Analysis*, *10*, 151–166. Retrieved from http://onlinelibrary.wiley.com/doi/10.1901/jaba.1977.10-151/abstract

Kazdin, A. E. (2008). Evidence-based treatment and practice: New opportunities to bridge clinical research and practice, enhance the knowledge base, and improve patient care. *American Psychologist*, *63*, 146–159.

Kelly, F., McNeil, K., & Newman, I. (1973). Suggested inferential statistical models for research in behavior modification. *Journal of Experimental Education*, *41*, 54–63. Retrieved from http://www.jstor.org/stable/10.2307/20150891

Lambert, M. J., Harmon, C., Slade, K., Whipple, J. L., & Hawkings, E. J. (2005). Providing feedback to psychotherapists on their patients' progress: Clinical results and practice suggestions. *Journal of Clinical Psychology*, *61*, 165–174.

Levin, J. R., Ferron, J. M., & Kratochwill, T. R. (2012). Nonparametric statistical tests for single-case systematic and randomized ABAB … AB and alternating treatment intervention designs: New developments, new directions. *Journal of School Psychology*, *50*, 599–624. http://dx.doi.org/10.1016/j.jsp.2012.05.001

Maric, M., Wiers, R. W., & Prins, P. J. M. (2012). Ten ways to improve the use of statistical mediation analysis in the practice of child and adolescent treatment research. *Clinical Child and Family Psychology Review*, *15*, 177–191.

Marriott, F., & Pope, J. (1954). Bias in the estimation of autocorrelations. *Biometrika*, *41*(3), 390–402. Retrieved from http://www.jstor.org/stable/10.2307/2332719

Molenaar, P. (1985). A dynamic factor model for the analysis of multivariate time series. *Psychometrika*, *50*, 181–202. Retrieved from http://link.springer.com/article/10.1007/BF02294246

Muris, P. (2001). A brief questionnaire for measuring self-efficacy in youths. *Journal of Psychopathology and Behavioral Assessment*, *23*, 145–149.

Norell-Clarke, A., Nyander, E., & Jansson-Fröjmark, M. (2011). Sleepless in Sweden: A single subject study of effects of cognitive therapy for insomnia on three adolescents. *Behavioral and Cognitive Psychotherapy*, *39*, 367–374.

Ougrin, D. (2011). Efficacy of exposure versus cognitive therapy in anxiety disorders: Systematic review and meta-analysis. *BMC Psychiatry*, *11*, 200.

R Statistical Package (version 3.0.2). Retrieved February 10, 2014, from http://www.r-project.org/

Robey, R. R., & Schultz, M. C. (1998). A model for conducting clinical-outcome research: An adaptation of the standard protocol for use in aphasiology. *Aphasiology*, *12*, 787–810. http://dx.doi.org/10.1080/02687039808249573

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147–177.

Silverman, W. K., & Albano, A. M. (1996). *Anxiety Disorders Interview Schedule for DSM-IV, Child and Parent Versions*. San Antonio, TX: Psychological Corporation.

Smith, J. D. (2012). Single-case experimental designs: A systematic review of published research and current standards. *Psychological Methods*. http://dx.doi.org/10.1037/a0029312

SPSS Inc. (2011). SPSS base version 20 for Windows User's Guide. Chicago, IL: Author.

Stice, E., Rohde, P., Seeley, J. R., & Gau, J. M. (2010). Testing mediators of intervention effects in randomized controlled trials: An evaluation of three depression prevention programs. *Journal of Consulting and Clinical Psychology*, *78*, 273–280.

Task Force on Promotion and Dissemination of Psychological Procedures, Division of Clinical Psychology, American Psychological Association (1995). Training in and dissemination of empirically validated psychological treatments: Report and recommendations. *The Clinical Psychologist*, *48*, 3–23.

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

Weisz, J. R., Chorpita, B. F., Frye, A., Ng, M. Y., Lau, N., Bearman, S. K., . . . Hoagwood, K. E. (2011). Youth top problems: Using idiographic, consumer-guided assessment to identify treatment needs and to track change during psychotherapy. *Journal of Consulting and Clinical Psychology*, *79*, 369–380.

RECEIVED: July 25, 2013
ACCEPTED: September 8, 2014
Available online 19 September 2014