上越教育大学研究紀要　第35巻　平成28年３月
Bull. Joetsu Univ. Educ., Vol. 35, Mar. 2016

Publication bias in meta-analysis of single-case research under a selection based on the statistical significance

Taichi OKUMURA*

(Received August 24, 2015; Accepted October 22, 2015)

ABSTRACT

This study examined the publication bias in the meta-analysis of single-case research when only significant effect sizes are available. Hypothetical data were generated under the AB design with no trends, assuming that responses were normally distributed. The standardized mean difference was tested in each case, and in meta-analysis only the significant data were integrated using a multilevel model. The results showed that the population effect size and the number of data points are the critical factors of the bias in the effect size estimates. When a treatment was ineffective in the population and data had few observations, the meta-analysis of the significant cases estimated that the treatment had been effective; the integrated effect size was approximately the median of the empirical distribution of the effect size reported in the past single-case studies. Meanwhile, other factors of the intraclass correlation and the first-order autoregressive correlation and interaction of these factors had small effects on the publication bias.

KEY WORDS
publication bias, meta-analysis, AB design, statistical significance

1. Introduction

1.1 Single-case research designs in the behavioral sciences

Single-case research designs (SCDs) have been used to describe intra-individual changes and to examine treatment effects in research on clinical psychology, school psychology, and special education. This is because, in these areas, researchers often emphasize their deep understandings of each case including their psychosocial backgrounds resulting from close relationships in real lives. In the behavioral sciences as a whole, however, between-subject designs (BSDs) are more popular than SCDs, and SCDs are even often considered to be inferior to BSDs as scientific methods (Kratochwill & Levin, 2014).

The criticism of SCDs can be summarized by three points. First, SCDs often lack internal validity, because it is difficult to control for nuisance variables that are likely confounding the treatment. In general, insufficient internal validity cannot provide strong evidence for specifying a causal relationship between a treatment and a change. Second, the external validity—the generalizability of a result to other cases or situations—cannot be guaranteed in SCDs because of the lack of information about individual differences. Third, the interpretation and evaluation of data tend to be subjective in SCDs. As a tradition, SCD data are firstly graphed and then visually analyzed by researchers, who are also often practitioners. However, past studies show that even so-called "experts" of visual analysis do not agree with one another in their judgments of the effectiveness of a treatment (Kratochwill, Levin, Horner, & Swoboda, 2014).

To compensate for these shortcomings, researchers have elaborated SCDs to enhance their scientific credibility. For example, repetitions of baseline and treatment phases in the interrupted time-series designs and setting different timings of a treatment between subjects in the multiple baseline design improve the internal and external validity of SCD research. Today, task forces in the clinical psychology, clinical child psychology, and school psychology divisions of the American Psychological Association recommend SCDs in their evidence standards (Kratochwill & Levin, 2014).

Effect size statistics and significance testing procedures have also been developed and are used widely today by SCD researchers, strengthening the scientific credibility of SCD studies. Effect size statistics for SCDs can be roughly classified into two types: non-overlap indices and standardized mean differences. The non-overlap indices

* School Education

have been developed originally for SCDs. They are more interpretable and statistically robust, but generally lack power for statistical inference and cannot be converted into well-known effect size measures for BSDs (Parker, Vannest, & Davis, 2014). Well-known standardized mean differences are Cohen's *d*, Hedges' *g*, and Glass's Δ in BSDs. Busk and Serlin (1992) proposed the reporting of one of the standardized mean differences for the ABAB design and the multiple baseline design, regarding A (baseline) and B (treatment) phases as the control and experimental groups in the BSD, respectively. The standardized mean differences are sensitive to model assumptions and seem to be less accessible intuitively, but are comparable with existing effect size measures in BSDs.

1.2 Publication bias in single-case research

SCD researchers have become interested in the meta-analysis of single-case research. Information on the inter-individual variability can be obtained through integrating many cases and it contributes to the evaluation of the generalizability of a treatment effect and factors that generate variations across cases. In meta-analysis of SCD research, the effect sizes are often integrated as is done for BSDs, whereas some researchers propose the integration of raw data using multilevel statistical models (Moeyaert, Ferron, Beretvas, & Van den Noortgate, 2014; Van den Noortgate & Onghena, 2003).

In meta-analysis, publication bias occurs when the research that appears in the published literature is systematically unrepresentative of the population of completed studies (Rothstein, Sutton, & Borenstein, 2005). Researchers have shown that publication biases exist in BSDs and they have investigated the causes of the biases. For example, it has been shown that manuscripts reporting non-significant results are less likely to be published than those reporting significant results (Egger & Smith, 1998). Accordingly, integrated results tend to be positively biased when only published articles are used for meta-analysis.

Shadish, Hedges, and Pustejovsky (2014) argued that publication bias can occur also in SCDs. They illustrated the inspection of and adjustment for publication bias in SCDs using six studies of the effects of pivotal response training on children with autism. The funnel plot was asymmetric, implying publication bias. They applied existing statistical methods for publication bias in BSDs to the data. Results were inconsistent between methods. Begg and Mazumdar's rank correlation test suggested no bias, while Egger's regression test suggested the presence of bias. After applying the trim-and-fill method, the mean effect size decreased from $g = 1.01$ to $g = 0.77$, suggesting positive bias.

Shadish et al. (2014) indicated that publication bias in SCD research could be a function of the effect size. This is because researchers may stop treating a case if treatment appears not to work and they may then not write a manuscript or submit it for publication (Shadish et al., 2014). As noted above, statistical hypothesis testing has become popular among SCD researchers for evaluating the effectiveness of treatments. Then, just as in the cases of BSDs, statistical significance could also have an effect on whether SCD research is published. If such a mechanism exists, how large is the bias in a meta-analysis when effect sizes are integrated by analyzing only the published cases?

In the light of the discussion above, this study illustrates the effects of publication bias in the meta-analysis of SCD research under controlled conditions using hypothetical data. Some constraints are provided for systematic considerations. First, this study deals with only the $(AB)^1$ design that assumes no trends, which is the most basic SCD. Second, data in the two phases are generated as being normally distributed with a common variance. Third, in each case, the treatment effect is evaluated by calculating Cohen's *d* proposed in Busk and Serlin (1992), and its significance is tested using the normal sampling distribution with approximate variance (Hedges & Olkin, 1985, p. 86). Fourth, this study assumes that only statistically significant cases are published and only data for them are used in meta-analysis. Fifth, the multilevel model is used to integrate the data as proposed by Van den Noortgate and Onghena (2003). Sixth, as factors that affect publication bias, four parameters are manipulated for the data generation: the population effect size, the intraclass correlation, the first-order autocorrelation, and the number of sessions. These constraints may seem to be too restrictive and improvisational, but this is unavoidable in the present circumstances; i.e., no clue has yet been found for identifying conditions, because almost no empirical or theoretical research on publication bias has been done for SCDs (Shadish et al., 2014).

2. Methods

2.1 Modeling the population

Data were generated from the following multilevel model as the population of $(AB)^1$ studies. As the first-level model, the response in the ith session for the jth person Y_{ij} is expressed as

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + r_{ij}, \tag{1}$$

where X_{ij} is a dummy valuable that takes a value of zero for Phase A and a value of 1 for Phase B. In this model, β_{0j} is the expected response for the jth person during the baseline (Phase A), and β_{1j} is the expected treatment effect of the person (difference between Phases A and B). The residual r_{ij} is normally distributed with a mean of zero and variance of $\sigma^2\phi^{|i-i'|}$, where ϕ is the autoregressive correlation coefficient.

The second-level model allows the intercept and slope of the first-level model (β_{0j} and β_{1j}) to be normally distributed around their grand means, which is equivalent to the variability across individuals:

$$\begin{bmatrix} \beta_{0j} \\ \beta_{1j} \end{bmatrix} \sim N\left(\begin{bmatrix} \gamma_0 \\ \gamma_1 \end{bmatrix}, \begin{bmatrix} \tau_0^2 & \tau_{01} \\ \tau_{10} & \tau_1^2 \end{bmatrix} \right), \tag{2}$$

where γ_0 is the grand mean during the baseline and γ_1 is the grand mean of the treatment effect. The variances of β_{0j} and β_{1j} are respectively τ_0^2 and τ_1^2, and they are correlated with the covariance of $\tau_{01}(=\tau_{10})$.

From Equations (1) and (2), total variances in responses in Phases A and B are

$$\mathrm{Var}(Y_{ij} \mid X_{ij} = 0) = \tau_0^2 + \sigma^2 \tag{3}$$

and

$$\mathrm{Var}(Y_{ij} \mid X_{ij} = 1) = \tau_0^2 + \tau_1^2 + 2\tau_{01} + \sigma^2, \tag{4}$$

respectively. To assume the homoscedasticity of the two phases noted as the second constraint above, $\tau_1^2 + 2\tau_{01}$ must be zero in Equation (4). In addition, the level-2 variances are assumed to have common values (i.e., $\tau_0^2 = \tau_1^2$) for simplicity.

When ϕ equals zero, the model is equivalent to the model for the multisite randomized trials (Raudenbush & Liu, 2000). Spybrook, Bloom, Congdon, Hill, Martinez, and Raudenbush (2011) defined the standardized effect size δ for the multisite randomized trials as

$$\delta = \frac{\gamma_1}{\sqrt{\sigma^2}}. \tag{5}$$

In the context of the $(AB)^1$ design, this equals the grand-mean difference between Phases A and B, which is standardized by the common within-individual standard deviation. We can interpret that this is the population value of the d-type estimator proposed by Busk and Serlin (1992).

2.2 Determining the parameter values and number of sessions

To generate data in realistic situations, the parameter values and number of sessions were determined referring to reviews of SCD research.

2.2.1 Effect size (δ)

Parker, Brossart, Vannest, Long, De-Alba, Baugh, and Sullivan (2005) reviewed 77 SCD studies and reported the distributions of effect size values. According to their review, the 25th, 50th, and 75th percentiles of R^2, obtained by regressing responses on the 0/1 phase variable, were roughly .3, .5, and .7, respectively. The R^2 values can be converted to d values using a conversion formula (Cohen, 1988, Equation 2.2.6), and the d-values are 1.31, 2.00, and 3.06, respectively. To examine the effects of publication bias when the null hypothesis is true and false, data were generated setting the effect size δ at 0 and 1. The latter value corresponds roughly to the 25th percentile reported in the review.

2.2.2 Intraclass correlation (ρ)

As far as the author knows, there is no systematic review that reports the empirical distribution of the intraclass correlation in SCD research. Moeyaert et al. (2014) applied a multilevel model to the data of nine cases reported by Lambert, Cartledge, Heward, and Lo (2006) and found that the intraclass correlations were .20 and .09 for the intercept and slope, respectively. Referring to the results, the intraclass correlation ρ was set to .10 and .20 in the data generation.

Table 1 summarizes the parameter values in the multilevel model for the data generation, concerning the effect size and intraclass correlation. The grand mean in the baseline γ_0 is unrelated to δ and ρ and was thus set as 10 in this study. In manipulating the values of δ and ρ, the grand mean difference γ_1 and the variance components were set to satisfy the constraints described above within an arbitrary scale of $\tau_0^2=\tau_1^2=1$.

Table 1

Parameter values for the data generation concerning the effect size and intraclass correlation

δ	ρ	γ_0	γ_1	σ	τ_0^2	τ_1^2	τ_{01}
0	.1	10	0	3	1	1	−.5
0	.2	10	0	2	1	1	−.5
1	.1	10	3	3	1	1	−.5
1	.2	10	2	2	1	1	−.5

2.2.3 Autocorrelation (φ)

Parker et al. (2005) reported that the 25th, 50th, and 75th percentiles of the distribution were −.023, .247, and .533, respectively. Shadish and Sullivan (2011) reviewed 113 SCD studies in 21 articles and reported that the autocorrelation was .20 on average. Hence, the autocorrelation φ was set as 0.0 and 0.2 in the data generation.

2.2.4 Number of sessions per case (*n*)

According to Shadish and Sullivan (2011), the mode of the number of sessions per case was 20. The standards of What Works Clearinghouse (WWC) recommend a minimum of five sessions in a phase (Kratochwill, Hitchcock, Horner, Levin, Odom, Rindskopf, & Shadish, 2010; Shadish & Sullivan, 2011). This requires at least 10 sessions per case in the (AB)[1] design. As small and moderate sizes, 10 and 20 sessions per case were set in this study, where the two phases have an equal number of data points (i.e., equal numbers of sessions).

2.3 Data generation and analysis

Publication bias was examined for the total of $2^4 = 16$ conditions (i.e., δ [0 or 1] × ρ [0.1 or 0.2] × φ [0.0 or 0.2] × *n* [10 or 20]). For each condition, 10,000 cases were generated and the data were meta-analyzed employing the multilevel model and using all cases and then using the cases for which the effect sizes were significantly larger than zero with a significance level of $\alpha = .05$. R version 3.1.0 (R Core Team, 2014) was used for the simulation; the MASS package was used for the data generation, the effsize package to test effect sizes, and the nlme package for the meta-analysis.

3. Results

Table 2 gives parameter estimates for the 16 conditions. When using all cases, estimates accurately reproduce the parameter values for all conditions. In contrast, when using only significant cases, estimates were biased more or less, depending on the conditions. As a consistent tendency, level-2 variance of the treatment effects (τ_1^2) and the covariance between baseline outcomes and the treatment effects (τ_{01}) were estimated to be nearly zero for almost all conditions. When $\delta = 0$ and $n = 10$, the estimates of intraclass correlation ρ were positively biased. The autocorrelation φ tended to be underestimated, especially when $\delta = 0$.

Biases in the standardized effect size δ ($Bias(\hat{\delta}) = \hat{\delta}^{(significant\ cases)} - \hat{\delta}^{(all\ cases)}$) are given in the final column of Table 2. It is seen that the effect size was overestimated for all conditions when integrating significant cases only.

The bias reached a maximum of 2.011 when $\delta = 0$, $\rho = .2$, $\phi = .2$, and $n = 10$, whereas it was at a minimum of 0.387 when $\delta = 1$, $\rho = .1$, $\phi = .0$, and $n = 20$. Generally, the bias was almost consistently around 2.0 (1.940–2.011) when $\delta = 0$ and $n = 10$, regardless of the levels of ρ and ϕ. In contrast, the bias was generally smaller when $\delta = 1$ and $n = 20$, and in these cases, the degree of bias was more variable (0.387–0.515) depending on the levels of ρ and ϕ.

Table 3 gives the results of the four-factor ($\delta \times \rho \times \phi \times n$) analysis of variance with $Bias(\hat{\delta})$ as the outcome variable. In this analysis of variance, levels of the factors are fixed and the model includes the main effects and the first-order interactions. The partial-ω^2 and the *f*-statistics show that the main effects of δ and n are distinct among the factors, and the main effects of ρ and ϕ and the first-order interactions are relatively small and practically ignorable. Figure 1 shows the main effects of the four factors. In the figure, we also see that $Bias(\hat{\delta})$ increases as δ and/or n becomes small and as ρ and/or ϕ becomes large, but ρ and ϕ have small effects on the bias.

Table 2

Results of meta-analysis under each condition when all cases are used and when significant cases are used

Factors				Parameter estimates (all cases)									Parameter estimates (significant cases)									Bias in
δ	ρ	ϕ	n	γ_0	γ_1	σ	τ_0^2	τ_1^2	τ_{01}	δ	ρ	ϕ	γ_0	γ_1	σ	τ_0^2	τ_1^2	τ_{01}	δ	ρ	ϕ	δ
0	.1	.0	10	9.99	0.01	3.00	0.99	0.97	−0.48	0.00	.10	.00	7.77	4.29	2.21	1.14	0.00	0.00	1.94	.21	−.06	1.940
0	.1	.0	20	10.00	0.02	3.00	0.99	0.98	−0.49	0.01	.10	.00	8.53	3.10	2.72	0.95	0.00	0.00	1.14	.11	−.04	1.132
0	.1	.2	10	9.99	0.04	3.04	1.06	1.09	−0.51	0.01	.11	.19	7.71	4.60	2.32	1.23	0.00	0.00	1.99	.22	.07	1.973
0	.1	.2	20	9.99	0.00	3.06	1.02	1.00	−0.51	0.00	.10	.20	8.36	3.23	2.77	0.97	0.00	0.00	1.17	.11	.14	1.168
0	.2	.0	10	10.02	−0.02	2.01	1.01	1.03	−0.52	−0.01	.20	.00	8.35	3.10	1.59	0.90	0.00	0.00	1.95	.24	−.03	1.964
0	.2	.0	20	9.99	0.02	2.00	1.00	0.98	−0.48	0.01	.20	.00	8.89	2.26	1.87	0.87	0.00	0.00	1.21	.18	−.03	1.197
0	.2	.2	10	9.99	0.02	2.03	0.99	0.99	−0.46	0.01	.19	.19	8.40	3.20	1.59	1.01	0.00	0.00	2.02	.29	.06	2.011
0	.2	.2	20	10.00	−0.01	2.05	1.00	1.01	−0.51	0.00	.19	.20	8.86	2.34	1.88	0.93	0.00	0.00	1.24	.20	.14	1.247
1	.1	.0	10	10.00	3.02	2.99	0.96	0.97	−0.46	1.01	.09	.00	8.89	5.30	2.52	1.00	0.00	0.00	2.10	.14	−.04	1.093
1	.1	.0	20	10.00	3.00	3.00	1.00	0.97	−0.52	1.00	.10	.00	9.48	4.03	2.90	0.87	0.00	0.00	1.39	.08	−.02	0.387
1	.1	.2	10	9.98	3.01	3.07	1.00	0.98	−0.52	0.98	.10	.20	8.76	5.50	2.56	1.14	0.00	0.00	2.14	.16	.10	1.163
1	.1	.2	20	9.99	3.03	3.06	1.01	1.01	−0.49	0.99	.10	.20	9.39	4.22	2.94	0.93	0.00	0.00	1.44	.09	.17	0.446
1	.2	.0	10	10.00	1.98	1.99	1.01	1.01	−0.49	0.99	.20	.00	9.10	3.71	1.73	0.94	0.00	0.00	2.14	.23	−.02	1.149
1	.2	.0	20	9.99	2.03	2.00	1.01	1.01	−0.50	1.01	.20	.00	9.56	2.89	1.96	0.87	0.22	−0.02	1.48	.17	.00	0.463
1	.2	.2	10	10.00	2.01	2.04	0.98	1.00	−0.51	0.98	.19	.21	9.09	3.82	1.76	0.96	0.00	0.00	2.18	.23	.13	1.195
1	.2	.2	20	9.98	2.02	2.04	1.00	1.02	−0.50	0.99	.19	.20	9.51	2.97	1.98	0.89	0.00	0.00	1.50	.17	.18	0.515

Table 3

Results of analysis of variance with four factors and estimation bias in δ as the outcome variable

Source	*SS*	*df*	*MS*	*F*	*p*	ω^2	*f*
δ	2.4188	1	2.4188	40271.4266	0.0000	0.9996	50.1687
ρ	0.0120	1	0.0120	200.5421	0.0000	0.9258	3.5315
ϕ	0.0097	1	0.0097	160.7170	0.0001	0.9089	3.1595
n	2.2000	1	2.2000	36629.0208	0.0000	0.9996	47.8461
$\delta \times \rho$	0.0000	1	0.0000	0.7586	0.4236	0.0000	0.0000
$\delta \times \phi$	0.0002	1	0.0002	3.8720	0.1062	0.1522	0.4237
$\delta \times n$	0.0079	1	0.0079	131.1394	0.0001	0.8905	2.8520
$\rho \times \phi$	0.0000	1	0.0000	0.0094	0.9267	0.0000	0.0000
$\rho \times n$	0.0012	1	0.0012	20.1051	0.0065	0.5442	1.0927
$\phi \times n$	0.0000	1	0.0000	0.0010	0.9755	0.0000	0.0000
Residuals	0.0003	5	0.0001				
Total	4.6501	15					

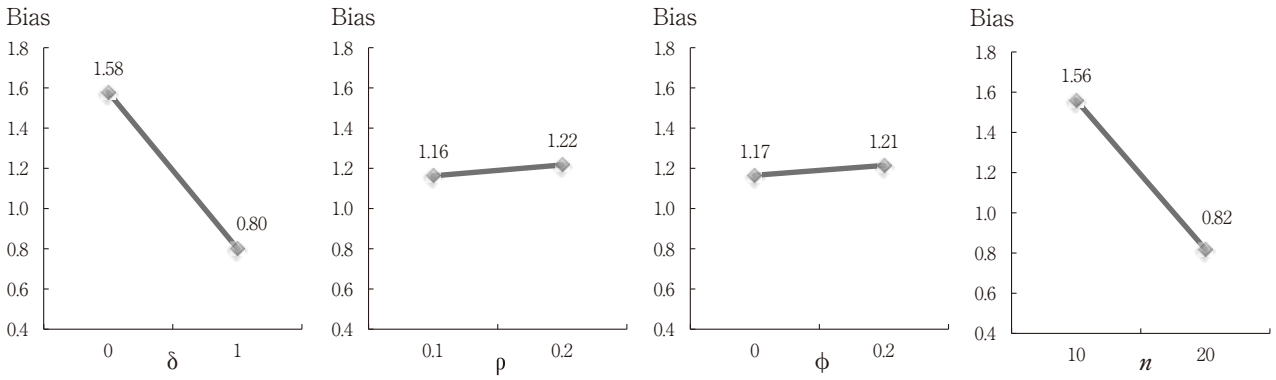

Figure 1. Plots of the mean biases in δ under each level of the four factors.

4. Discussion

The results of the simulation show that the estimate of the effect size δ becomes positively biased when the meta-analysis uses data of reported significant treatment effects. The bias was larger when the null-hypothesis was true (δ = 0) than when the null-hypothesis was false (δ = 1). In other words, if other conditions are constant and publication bias does exist, as the estimated effect size increases, the population value δ approaches zero. Apparently, evaluating an actually ineffective treatment as if it were effective is a much more serious problem than overestimating the effectiveness of an actually somewhat effective treatment. The results indicate such serious misspecifications can occur in SCD research.

When data involved fewer observations, this "false positiveness" became severe, and the standardized mean difference was about 2.0 when n = 10. As noted above, Parker et al. (2005) showed that this value of the effect size roughly equals the median of the empirical distribution of d-values. In addition, Shadish and Sullivan (2011) indicated that only about a half of SCD research fulfilled the WWC's standard of at least five sessions in each phase (i.e., n = 10 in the $(AB)^1$ design). Integrating these findings, we might even make a shrewd guess that the empirical distribution that Parker et al. (2005) reported itself was an artifact of publication bias, which reflects non-effectiveness of the treatments. Of course, this is speculation as long as the true state and mechanism of publication bias are unknown in SCD research.

Integrating only significant cases also affected the evaluation of the variability in the treatment effects. For most conditions, the variability was extremely underestimated as if there were no individual differences in the effectiveness. Needless to say, this is a big threat to the scientific credibility of SCD research and its meta-analysis. Practitioners should consider maximizing benefits of clients before anything else when a treatment is actually applied. Hence, it is indispensable for them to know the variability of the treatment effect between individuals and factors that affect the effectiveness (e.g., kinds of disorder, intelligence, and personality). This is known as the aptitude–treatment interaction. To link the science and practice, behavioral scientists have the responsibility to clarify the aptitude–treatment interaction and to provide practitioners with precise information on it. The results of the simulation indicate that meta-analysis is capable of disconnecting the link using data of significant cases, which are more probably published.

Several tasks remain for future research. First, the present study sets many constraints: data are normally distributed under the $(AB)^1$ design with no trends and only statistically significant cases are available. Some constraints will not be realistic. Hence, to examine the effects of publication bias in SCD, additional research should be done under other situations, including those of binary and count responses, linear and non-linear trends, and realistic odds of publication and accessibility of cases (Shadish et al., 2014). Second, to set more realistic situations, empirical and theoretical research on publication bias in SCDs should be done. For example, analysis of SCD data mainly relies on a visual method, and at present, statistical methods are regarded as supplements to visual methods (Kratochwill, et al., 2014). Accordingly, factors that affect the publication of research and even the decision of whether to continue the treatment will not simply be the same as those for BSDs. This implies the uniqueness of

the mechanism of publication bias in the meta-analysis of SCD research. Third, on the basis of such studies, researchers should develop appropriate methods for adjusting publication bias in the meta-analysis of SCD research. Such methods are indispensable for knowing what is really effective and which treatment should be applied to specific cases. Behavioral sciences do not accomplish their aim until they can provide precise information that is beneficial to individual practices.

Acknowledgements

This study was supported by MEXT/JSPS KAKENHI Grant Numbers 24730530 and 15K17267.

References

Busk, P. L., & Serlin, R. C. (1992). Meta-analysis for single-case research. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp. 187–212). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Egger, M., & Smith, G. D. (1998). Bias in location and selection of studies. *British Medical Journal*, *316*, 61–66.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego, CA: Academic Press.

Hedges, L. V., Pustejovsky, J. E., & Shadish, W. R. (2012). A standardized mean difference effect size for single case designs. *Research Synthesis Methods*, *3*, 224–239.

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). *Single-case designs technical documentation*. Retrieved from http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_scd.pdf

Kratochwill, T. R., & Levin, J. R. (2014). Introduction: An overview of single-case intervention research. In T. R. Kratochwill, T. R., & J. R. Levin (Eds.), *Single-case intervention research: Methodological and statistical advances* (pp. 3–23). Washington, DC: American Psychological Association.

Kratochwill, T. R., Levin, J. R., Horner, R. H., & Swoboda, C. M. (2014). Visual analysis of single-case intervention research: Conceptual and methodological issues. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research: Methodological and statistical advances* (pp. 91–125). Washington, DC: American Psychological Association.

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

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

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

Parker, R. I., Vannest, K. J., & Davis, J. L. (2014). Non-overlap analysis for single-case research. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research: Methodological and statistical advances* (pp. 127–151). Washington, DC: American Psychological Association.

R Core Team (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org/

Raudenbush, S. W., & Liu, X. F. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods, 5*(2), 199–213.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005). *Publication bias in meta-analysis: Prevention, assessment, and adjustments*. West Sussex, England: Wiley.

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

Shadish, W. R., & Sullivan, K. J. (2011). Characteristics of single-case designs used to assess intervention effects in

2008. *Behavior Research Methods, 43*(4), 971–980.

Spybrook, J., Bloom, H., Congdon, R., Hill, C., Martinez, A., & Raudenbush, S. (2011). *Optimal design for longitudinal and multilevel research* [Documentation for the Optimal Design Software Version 3.0]. Retrieved from www.wtgrantfoundation.org

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