Journal of
Applied Behavior Analysis

A SURVEY OF PUBLICATION PRACTICES OF SINGLE-CASE DESIGN RESEARCHERS WHEN TREATMENTS HAVE SMALL OR LARGE EFFECTS

WILLIAM R. SHADISH, NICOLE A. M. ZELINSKY, AND JACK L. VEVEA

UNIVERSITY OF CALIFORNIA, MERCED

AND

THOMAS R. KRATOCHWILL

UNIVERSITY OF WISCONSIN, MADISON

The published literature often underrepresents studies that do not find evidence for a treatment effect; this is often called *publication bias*. Literature reviews that fail to include such studies may overestimate the size of an effect. Only a few studies have examined publication bias in single-case design (SCD) research, but those studies suggest that publication bias may occur. This study surveyed SCD researchers about publication preferences in response to simulated SCD results that show a range of small to large effects. Results suggest that SCD researchers are more likely to submit manuscripts that show large effects for publication and are more likely to recommend acceptance of manuscripts that show large effects when they act as a reviewer. A nontrivial minority of SCD researchers (4% to 15%) would drop 1 or 2 cases from the study if the effect size is small and then submit for publication. This article ends with a discussion of implications for publication practices in SCD research.

Key words: single-case design, publication bias, effect size

A central method for identifying evidence-based practices is to review relevant research on the topic. A key problem in these reviews can occur if they exclude negative results; that is, results that find no visual or statistically significant evidence of an effect (Kratochwill, Stoiber, & Gutkin, 2000). The conundrum is that negative results tend to appear less often than positive results in publications (Ferguson & Heene, 2012; Nosek, Spies, & Motyl, 2012), a phenomenon variously called *publication bias* (Epstein, 1990, 2004; Mahoney, 1977; Peters & Ceci, 1982), *positive-outcome bias* (Emerson et al., 2010), or *prejudice against the null hypothesis* (Greenwald, 1975). This bias may result in an overestimate of the size of treatment effects found by reviews that rely mostly on published literature.

Substantial evidence supports the existence of publication bias (Rothstein, Sutton, & Borenstein, 2005). It can occur if authors choose not to submit or resubmit manuscripts for publication because results are not significant, resulting in a "file drawer problem" (Rosenthal, 1979) in which subsequent researchers cannot easily find studies with negative results. Sometimes this bias occurs if a researcher omits some negative outcomes but does report positive results. Chan, Hróbjartsson, Haahr, Gøtzsche, and Altman (2004) tracked medical trials from their initial design in a clinical trial registry to final publication; publications incompletely reported over half of the outcomes described in

This research was supported in part by Grants R305D100046, R305D100033, and R305D150041 from the Institute for Educational Sciences, U.S. Department of Education. The opinions expressed are those of the authors and do not represent views of the University of California, the Institute for Educational Sciences, or the U.S. Department of Education.

Correspondence concerning this article should be addressed to William R. Shadish, School of Social Sciences, Humanities and Arts, University of California, Merced, 5200 North Lake Rd., Merced, California 95343 (e-mail: wshadish@ucmerced.edu).

doi: 10.1002/jaba.308

an initial registry, especially for negative results. Sometimes bias occurs when reviewers reject a manuscript due to negative results. Atkinson, Furlong, and Wampold (1982) found that reviewers for a mock journal article rejected manuscripts with negative results more often than those with positive results, even when the designs of the studies were identical.

Little work has been conducted on publication bias in the single-case design (SCD) literature. This may be because the assumed mechanism for publication bias is statistical significance testing, but SCD researchers rarely use significance tests. SCD researchers mostly analyze SCDs visually (Baer, 1977; Kratochwill, Levin, Horner, & Swoboda, 2014), by examining the visual difference between baseline and treatment phases, and consider the results to be negative if they do not show a clear visual functional relation. A few empirical studies suggest that publication bias nonetheless exists in the SCD literature. Mahoney (1977) asked 75 reviewers from the *Journal of Applied Behavior Analysis* to review one of five almost identical SCD manuscripts. Manuscripts differed in whether the results were positive (large visible effects), negative (no visible effects), or a mixture of the two; the manuscripts also differed in whether the discussion was positive (claiming an effective treatment) or negative (claiming an ineffective treatment). Other parts of the manuscript were identical. Reviewers rated manuscripts higher if they reported positive results than if they reported negative or ambiguous results.

Sham and Smith (2014) looked at effect sizes of 21 published studies and 10 unpublished dissertations using SCDs; none of the publications' authors were also authors of the dissertations. Published and unpublished studies did not differ on rated study quality, but the percentage of nonoverlapping data-effect size was lower for unpublished dissertations than for publications, a finding that suggests that less effective treatments are often not published.

Two studies (Shadish, Hedges, & Pustejovsky, 2014; Shadish, Hedges, Pustejovsky, Boyajian, et al., 2014) examined the issue of publication bias in SCD studies using meta-analytic tests of publication bias. Those studies did meta-analytic reviews of the literatures about the effects of a treatment for children with autism and treatments for patients with brain trauma, respectively. They found modest evidence that publication bias may exist in the reviewed studies, and adjusting for that bias can sometimes reduce observed effect sizes by a substantial proportion.

These studies each have different limitations due to the different methods they used. As these studies illustrate, researchers study publication bias in three ways. The first is by using analogue experiments with random assignment to the conditions assumed to lead to publication bias. For instance, Mahoney (1977) distributed different versions of entire fictitious manuscripts to reviewers. Alternatively, to reduce response burden, one can manipulate and distribute only part of a manuscript, such as graphs of results. The present study does the latter, with the disadvantage that components of the manuscript that provide the context for a review were removed.

A second way to study the problem is quasiexperimentally. The research by Sham and Smith (2014) is an example of a retrospective study that did not manipulate the conditions presumed to lead to publication bias. They tested whether unpublished work has smaller effect sizes than published work, which publication bias theory would predict. An advantage of this method is that the previous research reports have the ecological validity of being complete, actual manuscripts. However, the unmanipulated conditions may then be confounded with other differences; for example, unpublished studies might have used smaller doses of treatment or reported less reliable outcomes. The nature of selection bias in quasiexperiments is that we cannot know of such confounding effects for certain.

A third way to study the problem uses meta-analytic publication bias tests (Rothstein et al., 2005). Description of all these tests is beyond the scope of this article, but detailed examples of their application to SCD research are available (e.g., Shadish, Hedges, & Pustejovsky, 2014; Shadish, Hedges, Pustejovsky, Boyajian, et al., 2014). These tests attempt to determine if the review underrepresents studies with small samples and small effects, which might be the case if publication bias were operating. However, these methods make statistical assumptions (e.g., homogeneity of effect size) that are often violated, which limit their applicability.

For SCDs, all four previous studies (Mahoney, 1977; Shadish, Hedges, & Pustejovsky, 2014; Shadish, Hedges, Pustejovsky, Boyajian, et al., 2014; Sham & Smith, 2014) suggest that a bias may exist against publishing negative SCD results and that this bias may be giving preference to large effects. In SCD research, tradition often emphasizes large and visually detectable functional relations (Kazdin, 1982, 2011; Kratochwill et al., 2013; What Works Clearinghouse, 2014). Kazdin (1982, p. 232) says, “Weak results will not be regarded as meeting the stringent criteria of visual inspection. Hence, visual inspection will serve as a filter or screening device to allow only clear and potent interventions to be interpreted as producing reliable effects.” Publication bias might result if SCD researchers (a) stop treating a case that does not yield a large functional relation, (b) do not write a manuscript or submit it for publication if results are negative, or (c) include only cases with a large functional relation. Reviewers or editors may reject manuscripts without a large functional relation. The end result of these possibilities would be literature reviews that overestimate the effects of treatment.

The present study adds to the existing literature about possible SCD publication bias by manipulating the effect size in SCD studies to determine whether this manipulation has an impact on publication-related preferences. The study uses a standardized mean difference statistic for SCDs that is in the same metric as the standardized mean difference statistic (*d*) used in between-groups studies (Hedges, Pustejovsky, & Shadish, 2012, 2013). Hedges' *g* is *d* after correction for small-sample-size bias. We use this terminology for *d* and *g*, and most of the findings in this study are in the metric of *g*.

We chose this effect-size measure for two reasons. First, Hedges et al. (2012, 2013) derived it from statistical theory, and it has a known distribution, standard error, and significance test, all of which give confidence in statistical conclusion validity. Second, it is in the same metric as the standardized mean difference statistic used in between-groups studies, so it connects the publication bias literatures in the two domains.

In summary, this article reports results of a survey that varies the size of the standardized mean difference statistic to evaluate whether such variation influences publication judgments made by SCD researchers. We looked at three forms of potential publication bias: not submitting research to a journal at all, dropping some cases and then submitting, and not recommending for publication as a reviewer. We also examine several potential moderators of responses.

METHOD

Sample and Participant Selection

We identified authors of SCD research who published at least one SCD study in one of 34 journals in 2012 (see Supporting Information for a list of the journals). This list contains journals that we knew at the time to publish SCD research in psychology and education, though the list may not be exhaustive. These 34 journals yielded 704 unique authors. Of these, we obtained e-mail addresses for 590 authors of the articles by searching the Internet or through correspondence with coauthors. When sent an e-mail invitation to participate in the study, 375 authors opened the e-mails but did not click on the survey link; of those, 295 clicked the link and started the survey, and of those, 243 authors (41.2%) completed the survey.

We compared the original 704 authors to the 243 authors who completed the survey to see if they differed on geographic region, journal type, or having a first author paper that year (Table 1). Authors who completed the survey originated from the United States more often than the complete set of authors contacted, χ^2 (3, N = 243) = 10.01, p = .018. Journal type differed significantly between completers and the original authors, χ^2 (3, N = 243) = 10.07, p = .018. Completers published less often in journals that focus on specific disorders, less often in the *Journal of Applied Behavior Analysis*, and more often in education-related journals. Finally, the percentage of completers with first-author publications did not significantly differ from the original authors, χ^2 (1, N = 243) = 3.55, p = .060.

Materials

The survey contained (a) demographic questions regarding the participant's characteristics and experiences with SCDs, (b) a vignette of a hypothetical study (see Table 2), and (c) eight simulated SCD figures (each figure contained three SCD graphs) purportedly from that hypothetical study. To reduce response burden, the eight figures were a random sample from 16 possible figures resulting from manipulations of different figure characteristics to be described shortly (see Figure 1 for an example; see Supporting Information for details on how the graphs were generated, including R code). We asked researchers to assume that the quality of the study was high and would be submitted for publication to a top journal. Each figure was accompanied by the same three publication bias questions:

1. If you obtained these results from a study you conducted, how likely would you be to submit a manuscript based on the study for publication?
2. Assuming you did submit, if you obtained these results from a study you conducted, how likely would you be to drop one or two of the three cases, and then submit a

Table 1
Demographics of the Originally Identified Authors and Authors who Completed the Survey

	Originally identified		Completed survey	
	n	Percentage	*n*	Percentage
Geographic region of authors				
U.S.	605	85.94	225	92.59
Europe	44	6.25	5	2.06
Asia and Pacific	28	3.98	6	2.47
Canada	27	3.84	7	2.88
Journal type				
JABA	199	28.27	56	23.05
School	200	28.41	88	36.21
Disorders	246	34.94	74	30.45
Mixture	59	8.38	25	10.29
Authorship order of respondents				
First author	224	31.8	91	37.4
Not first author	480	68.2	152	62.6
Total	704	100	243	100

Table 2
Instructions to Respondents

SURVEY INSTRUCTIONS

The following pages present graphs of simulated results from a hypothetical single-case design study. We are mainly interested in your opinions about the publishability of the results.

The hypothetical study used an ABAB design and it has three cases in which the treatment was presented simultaneously and independently to each case. To provide some context for the research, assume that:

- the three cases are teenage boys with developmental disabilities,
- the cases have a presenting problem of overly rapid eating,
- each boy chose a favorite food to consume at lunch,
- the treatment was focused on increasing time spent eating that item of food,
- the treatment was administered over 24 consecutive school days at lunchtime in a school setting, and
- an observer used a digital stopwatch to record the total duration in seconds of eating time for the chosen item of food, and a second observer conducted interobserver agreement which involved an acceptable measure and high agreement.

For purposes of this survey, assume that the study is otherwise very well done: for example, the rationale for the study is compelling and grounded in previously published research, treatment implementation integrity data are good, interrater reliability is high, experimental conditions are well controlled, and the selected outcome measure is judged to be credible for this problem and to have good social validity. Because single-case design researchers often report statistics describing the outcome within phase for each case, the graphs present the mean number of seconds for each phase in addition to the raw outcome data.

For methodological reasons, we have kept the scale of the vertical axes the same over all the graphs. We recognize that sometimes a researcher might have chosen to reduce the scale of the graph so as to make a possible functional relation more visibly salient. This is another reason we have presented the phase means in the graphs, as one more aid to see how phases differed from each other.

After each graph, we will ask you three questions about whether you would be likely to try to publish (or recommend publishing) the results in the graphs on that page. We realize that publication judgments are never made solely on the basis of results, which is one reason why we ask you to assume that the hypothetical study is otherwise very well done and appropriate for this area of research. In addition, we request that for present purposes you consider irrelevant as a publishability criterion the novelty and/or potential impact of the study's findings per se (i.e., whether the findings support/replicate or disconfirm previously published findings in the literature). The three questions are:

1. If you obtained these results from a study you conducted, how likely would you be to submit a manuscript based on the study for publication?
2. If you obtained these results from a study you conducted, how likely would you be to drop one or two of the three cases and then submit a manuscript containing only the remaining cases for publication?
3. If you were a journal reviewer asked to review a manuscript containing these results, how likely would you be to recommend that the study be published?

We will ask you to answer the questions for each set of graphs using a 5-point Likert-type rating scale:

Very likely	Somewhat likely	Unsure	Somewhat unlikely	Very unlikely
5	4	3	2	1

Please go to the next page and begin.

manuscript containing the remaining cases for publication?

3. Assuming the study is otherwise well done, if you were a journal reviewer asked to review a manuscript containing these results, how likely would you be to recommend that the study be published?

We had to choose how to scale the vertical axis for the three cases in each figure: either keeping it constant across all three cases within a figure or varying the scale for each case in each figure to reflect the lowest and highest observations in that case. To help with this decision, we examined common practice in published SCD studies. In a sample of over 100 SCD studies we reviewed (Shadish & Sullivan, 2011), 69% contained graphs that used the same scales for all graphs using that dependent variable. Therefore, we made the scales of the graphs the same for all graphs within each figure. For example, the scales in the graphs of each case in Figure 1 range from 55 s to 185 s, even though the observed range

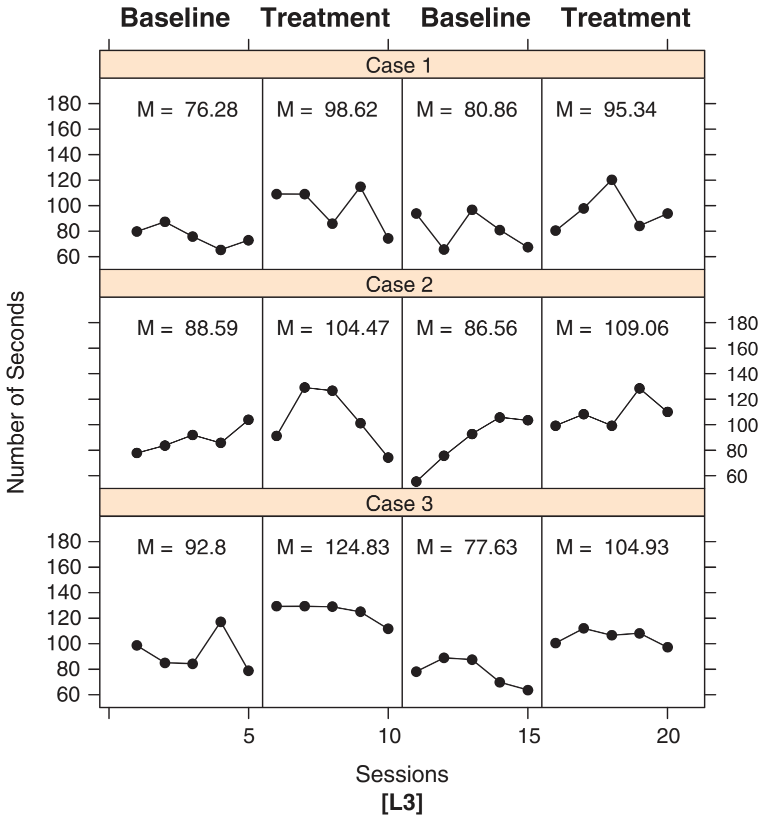

Figure 1. Figure for an effect size of 1.37, high overlap, high variability.

of observations for each case was not so wide. On the other hand, we allowed the scales of the graphs for different figures to change to reflect that each figure represented hypothetical study results that were independent from each other figure. For example, although the scales on Figure 1 ranged from 55 to 185 s, another figure ranged from 80 to 120 s to reflect the minimum and maximum observations in the three cases in that figure.

We asked about the education and career of each author, including (a) the number of years of education each author had completed, (b) how many of their coauthored publications used or discussed SCDs, (c) how many years each had used SCDs in research or practice, (d) if the author had ever taught a class or workshop on SCDs, (e) if the author had taken graduate level coursework on SCDs, (f) their primary work setting (higher education, K-12, health or mental health agency, private practice, or other), and (g) if that work setting was in higher education, if they were assistant professor, associate professor, full professor, or other.

The main stimuli in this survey were 16 figures of hypothetical SCD results, each figure displaying results from three cases. Each figure varied on all possible combinations of three independent variables: (a) effect size (small-sample bias corrected g = 0.425, 0.82, 1.375, 2.5), (b) overlap (high or low) among data points within a case, and (c) total variability (high or low) of data points within and between cases.

To identify a plausible range of effect sizes, we examined two previously published samples of 74 SCD effect sizes (Shadish, Hedges, & Pustejovsky, 2014; Shadish, Hedges, Pustejovsky, Boyajian, et al., 2014) that used the Hedges et al. (2012, 2013) g statistic adjusted for small sample bias. These effect sizes ranged from $g = -0.324$ to 11.476, with a mean of $g = 1.589$, median of $g = 0.978$, and standard deviation = 1.977; hence, the data were positively skewed. The cutoffs for the deciles of the effect sizes were $g = 0.156$, 0.305, 0.425, 0.664, 0.978, 1.376, 1.699, 2.305, and 4.520. Using quartiles to indicate small, medium, and large effect sizes, the 25th percentile is $g = 0.373$, the 50th percentile is $g = 0.978$, and the 75th percentile is $g = 1.874$. The four effect sizes we used to create graphs ($g = 0.425$, 0.82, 1.375, 2.5) approximated the 30th, 45th, 60th, and 80th percentiles. We did not create graphs for the smallest (0.156) and largest (4.520) effect sizes because initial testing on a convenience sample of a few SCD researchers suggested that those effect sizes produced no variability in the likelihood of publication. Originally, we planned to present more effect sizes to respondents, but eventually settled on four because we also had to manipulate two other independent variables (overlap and variability), so including more effect sizes would increase the number of graphs from 16 to a higher number than we believed response burden would allow.

The second independent variable was data overlap. An effective treatment should result in outcomes during the treatment phase that have little or no overlap with outcomes during the baseline phase. SCD researchers visually analyze the overlap between a baseline and treatment phase to determine treatment effectiveness (Lane & Gast, 2014), so overlap may influence publication decisions. We operationalized high overlap and low overlap with an intraclass correlation (ρ) near 0.75 and 0.10, respectively. The intraclass correlation is the ratio of between-cases variability to total variability (between cases plus within case). If we hold the size of g constant, then a low intraclass correlation allocates variability within rather than between cases, creating less overlap within cases. Later, we present overlap effect sizes to demonstrate the effectiveness of this manipulation.

The third independent variable was total data variability, operationalized by creating data with a within-case standard deviation of 20 for high variability and of 10 for low variability. For a fixed effect size g and fixed intraclass correlation ρ, larger within-case variability will increase the total variability across all three graphs. Higher total variability may make it more difficult for visual analysts to detect a difference between treatment and baseline conditions.

Fully crossing effect size, overlap, and variability yielded 16 figures (see Supporting Information for more detail and R code for producing the simulated data and graphs). We gave each participant eight randomly chosen figures to reduce response burden. All participants randomly received a figure with high or low total variability for all combinations of effect size and overlap. Thus, missing data (due to nonresponse or technical error) are randomly distributed across combinations, which facilitates subsequent analysis. The three publication-bias questions described above followed each figure with responses on 5-point Likert-type scales (1 = *highly unlikely* to 5 = *highly likely*).

Procedure

We used Qualtrics (Version 61959; Qualtrics, 2014) to e-mail each potential participant a link to the survey. We offered a $10 Amazon gift card as an incentive to complete the survey. After the initial invitation, we sent two reminder e-mails to participants who had not yet completed the survey and had not opted out of taking it. We e-mailed the first invitation in mid-July 2014, the first reminders in early August 2014, and the second reminders in September 2014. We received 126, 88, and 29 completions, respectively, for each of the three mailings. Thus, 243 of 590 invitees (41.2%) completed the survey.

RESULTS

Demographics

For years of education, many responses did not provide the expected answers of 1 for first grade through at least 20 for a doctorate, with the expectation that authors on a publication would have no less than a high school degree. For example, 104 respondents reported (a) number of years as 60 or 6, (b) responses combining numbers with text, such as 9 (undergraduate and graduate work) or 6 graduate, (c) the name of their degree instead of a number (e.g., PhD), or (d) no response. To ensure comparable responses for this variable, we assumed that undergraduate education lasted 4 years, and that the combination of primary and secondary education lasted 12 years. For respondents who said only that they had a doctorate degree, we used the median of all other respondents with a doctorate degree (22 years). We emailed 37 authors with unclear responses and received 24 responses with exact numbers. For the remaining 13 authors, we looked up educational information online (e.g., as listed in curricula vitae) and applied the above rules as necessary. After these procedures, we had no missing data for this variable.

We used online information to fill in six additional instances of missing data (three for years of SCD research, one for taking coursework in SCD, one for teaching SCDs, and one for work in higher education). We rounded two answers for years of SCD research to the nearest whole number. For the work and higher education variables, subsequent model fit indices suggested that a dichotomous variable coded for jobs inside or outside higher education yielded the best fit, so that coding was applied.

Table 3
Counts, Means, and Standard Deviations for Figure Questions

	Low overlap		High overlap	
g	High variability	Low variability	High variability	Low variability
Total responses per figure				
2.5	112	133	123	125
1.375	100	146	122	123
0.82	123	122	126	121
0.425	122	123	126	118
Mean of sample likely to submit manuscript				
2.5	4.63 (0.70)	4.50 (0.86)	3.82 (0.99)	3.94 (1.04)
1.375	4.16 (0.99)	4.07 (1.00)	3.06 (1.16)	2.82 (1.13)
0.82	3.03 (1.17)	3.23 (1.20)	2.40 (1.25)	2.14 (1.04)
0.425	2.21 (1.19)	2.22 (1.27)	1.78 (1.03)	1.84 (1.00)
Mean of sample likely to drop one or two cases				
2.5	1.53 (0.81)	1.35 (0.84)	1.75 (1.07)	1.72 (1.16)
1.375	1.44 (0.88)	1.71 (1.06)	1.84 (1.22)	1.81 (1.11)
0.82	1.87 (1.21)	1.72 (1.09)	1.69 (1.11)	1.85 (1.18)
0.425	1.75 (1.13)	1.57 (1.02)	1.68 (1.08)	1.53 (1.01)
Mean of sample likely to recommend for publication				
2.5	4.56 (0.55)	4.29 (0.89)	3.57 (1.04)	3.72 (0.98)
1.375	3.97 (0.97)	3.84 (1.03)	2.76 (1.10)	2.62 (1.07)
0.82	2.82 (1.19)	3.03 (1.10)	2.29 (1.12)	2.07 (0.98)
0.425	2.09 (1.06)	2.07 (1.16)	1.81 (0.92)	1.77 (0.93)

Note. Standard deviations presented in parentheses.

Participants had a mean of 21.76 years of education and a mean of 19.21 coauthored SCD publications; they had worked with SCDs for a mean of 13.38 years. Of the 243 participants, 63.79% had taught SCDs, 88.48% had taken graduate level classes on SCDs, and 79.42% worked in or closely with higher education.

Descriptive Results on Main Outcomes

Table 3 presents sample sizes, means, and standard deviations for all responses and conditions. Restriction of range was not a problem, given that all responses ranged from 1 to 5 except when $g = 2.5$ with low overlap and high variability, when responses ranged only from 3 to 5. Due to a technical error, Qualtrics gave seven participants more than eight randomly selected graphs (two had nine graphs, three had 11 graphs, and two had 13 graphs). The extra responses from these participants were included in the data set.

Manipulation Check for Overlap

If our manipulation of overlap was successful, it should correlate with nonoverlap of all pairs (NAP; Parker & Vannest, 2009) and tau U statistics (Parker, Vannest, Davis, & Sauber, 2011), the two best developed nonoverlap statistics. Figure 2 confirms this to be the case. Figures we manipulated to have high overlap had smaller nonoverlap statistics than the low-overlap condition. All three overlap statistics increased rapidly as *g* increased, and reached an asymptote (nonoverlap = 1.00) at approximately $g > 1.00$. Hence, nonoverlap statistics show significant ceiling and floor effects, and restriction of range, compared to *g*. They also have a nonlinear relation with *g*.

Main Analysis

We used a hierarchical linear model to account for nesting of responses to the eight graphs within participants, using R Version 3.1.1

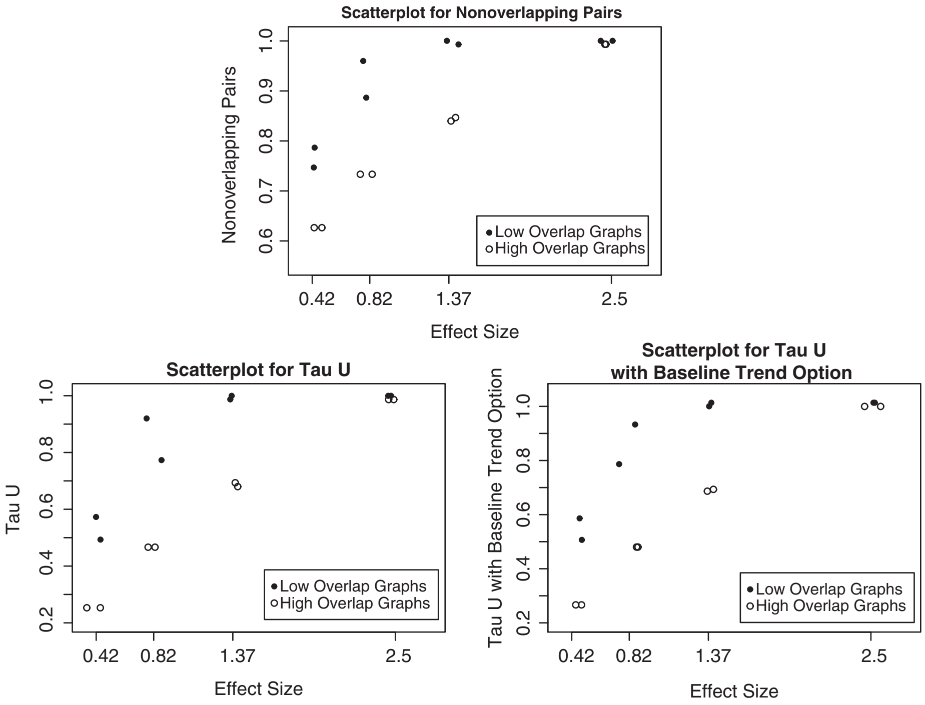

Figure 2. Nonoverlapping pairs (NAP) and tau U statistics plotted against effect size.

with the package *nlme* (Pinheiro, Bates, DebRoy, Sarkar, & R Core Team, 2014). The model predicts each of the three outcomes from (a) graph characteristics: effect size, overlap, variability, three two-way interactions between each of the three main effects, and a three way interaction of effect size, overlap (high vs. low), and variability (high vs. low); and (b) participant characteristics: years of education, number of SCD publications, years using SCDs, having taught SCDs, taken graduate courses in SCDs, and the dummy variable for job type. The Supporting Information contains details about the model, tests for collinearity (high correlation between variables that causes estimation problems), additional tables and graphs of results, and follow-up tests for significant interactions.

Results for Submitting a Manuscript for Publication

Respondents said they were significantly more likely to submit a manuscript for publication in the presence of a larger effect size (*g*) and when the data overlap was low (Figure 3); no interaction terms were significant. They were less likely to submit if they had more SCD-related publications, had more years of experience with SCDs, or took graduate coursework in SCDs. These predictors do not account for all of the variance in submitting a manuscript for publication, so additional research is needed to predict the variance in respondents' propensity to submit a manuscript.

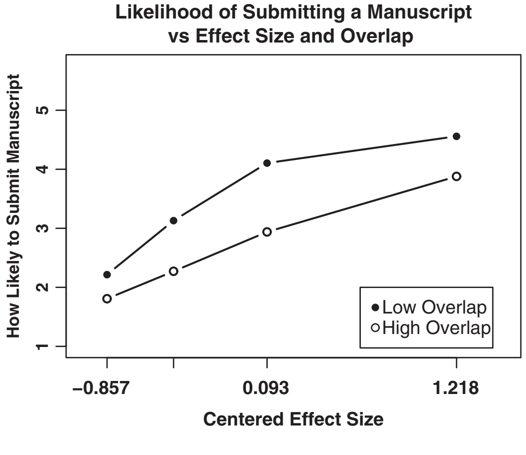

Figure 3. Mean values of likelihood of submitting a manuscript for publication on a 5-point Likert-type scale for graphs of different effect sizes and high or low overlap.

Results for Dropping One or Two Cases

Most respondents said they were unlikely or very unlikely drop a case before submitting a manuscript for publication. A small minority (4% to 15%) would drop a case that had a low effect size and high overlap, with most of this effect due to effect size and not overlap (Figure 4). Respondents were more likely to drop a case if they worked in higher education and had more years of education but had fewer years of experience with SCDs. These predictors do not account for all of the variance in decisions about dropping a case.

Results for Recommending a Manuscript for Publication

The analysis of recommending a manuscript for publication during the peer review process

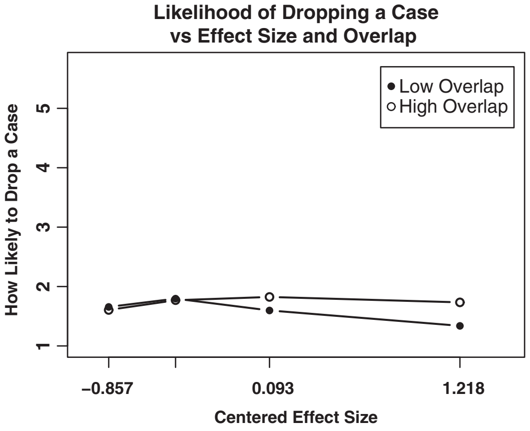

Figure 4. Mean values of likelihood of dropping one or two cases before submitting a manuscript for publication on a 5-point Likert-type scale for graphs of different effect sizes and high or low overlap.

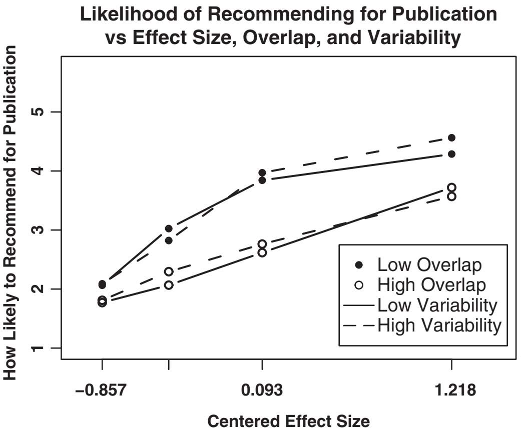

Figure 5. Mean values of likelihood of recommending a manuscript for publication on a 5-point Likert-type scale for graphs of different effect sizes, high or low overlap, and high or low overall variability.

yielded a three-way interaction among effect size, overlap, and variability (Figure 5). Researchers were more likely to recommend a manuscript for publication as effect size increased, but even more so with low rather than high overlap, and even more so when low overlap was accompanied by high variability, although the moderating effect of high variability is trivial from a practical point of view. By trivial we mean that the difference in mean response between low- and high-variability conditions never exceeded 0.27 on the 5-point Likert scale response format. This small difference almost never changed, whether a response was higher or lower than one of the five anchor points. Researchers were less likely to recommend a manuscript for publication if they had more SCD-related publications, had more years of experience with SCDs, and had taken graduate coursework in SCDs. These predictors do not account for all of the variance in researchers' propensities to recommend a manuscript for publication as a reviewer.

Participants' Comments

Participants had the option to leave comments at the end of the survey, which we categorize in Table 4. The most common comment was that researchers should not drop cases before submitting a manuscript for publication because doing so is unethical. Some comments claimed personal knowledge of dropping cases, such as "I am not in favor of dropping cases when the results do not conform to other individual cases. This is somewhat intellectually dishonest but I know it is done quite frequently," and "I believe it unethical to drop subjects from any study (any design) based on results, and I even asked to have my name removed from a paper for that very reason." A smaller number of respondents said that results needed to be positive (not null), or of practical or clinical significance, to warrant publication.

Other respondents commented on the design of the survey and stimuli. Some said they based their answers on detecting trend (even though the simulated data had no trend), in-phase instability, overlap, study quality, and perceived practical or clinical significance. Others would have preferred a different vertical axis for each graph in each figure. Still others said that their answers would depend on the full paper, that they would have liked to be able to revisit previous answers and the study vignette, and that the stimuli should have included multiple baseline designs and more data.

DISCUSSION

In general, SCD researchers said they would give preference to data that showed large effects and small overlap when they make publication decisions. Total data variability had very little impact on such decisions. This preference for large effects was even stronger for more experienced SCD researchers, who were also less likely to submit or recommend a manuscript for publication in general, perhaps suggesting that they are slightly more discriminating in their expectations about what is publishable. Experienced SCD researchers were also less

Table 4
Frequency of Comments for Each Category

Category	Frequency
When research should be published	
Researchers should not drop negative cases (because this is unethical)	45
Ideally all high-quality research should be published	10
Ideally negative results should be published	4
Research's effect should be practical or of clinical significance	15
Potentially negative results should only be published with extra motivation	6
Perceived influences of participants' responses	
Felt that the full paper would influence participants' responses	21
Perceived trend influenced participants' responses	18
Perceived within-phase stability or instability influenced participants' responses	13
Perceived overlap influenced participants' responses	8
Scale of the graph influenced participants' responses	8
Design criticisms	
The hypothetical research should have included more data or more cases	6
Should include a temporal control (e.g., a multiple baseline design)	3
Study criticisms	
Should allow participants to go back to previous pages	4
The questions about dropping cases were confusing	2

likely to drop a case before submitting a manuscript for publication, perhaps reflecting a more proficient understanding of the ethics of such an action.

We have argued that failure to publish SCD results with small effect sizes is a bias, which leads to evidence-based practice reviews that systematically overestimate the effectiveness of treatments in SCD research. However, at least some SCD researchers believe that demonstrating a visually compelling functional relation (and thus a large effect size) is not a bias but rather is good SCD research practice and should be an important consideration in publication decisions. These researchers assert that studies that do not demonstrate a visually large functional relation are uninterpretable; for example, a negative result may not mean that the treatment failed but rather that the researcher failed to implement the treatment adequately or failed to measure the outcome with enough reliability or validity. These statements may be true, although it would be better to base publication decisions on direct evidence about poor treatment implementation or poor measurement reliability than on indirect evidence of small effect sizes. Even so, a negative result may sometimes mean the treatment does not work well. SCD researchers need to better define professional standards for publishing negative effects and the process for documenting intervention ineffectiveness. Knowledge of what does not work should have just as great a place in evidence-based practice reviews as knowledge of what does work. Also, studies with negative results may differ from studies with positive results in having different kinds of cases, settings, treatment variations, or outcomes. Omitting results that are negative for this reason deprives the field of knowledge about what moderates the size of an effect.

Some SCD researchers might have (if given the option) responded to the stimuli with "not likely" to submit the minor or noneffect graphs because they would consider the study to be incomplete. They might argue that the standard should be to keep working until they get an effect, but to show all of their work and data in the publication. Similarly, reviewers might reject negative findings in the belief that more work needs to be done. This suggests further research to examine published articles to find evidence of this possibility. Examples of this evidence might include direct statements by the authors that they used this strategy, including whether or not they presented preliminary results, or indirect evidence such as the size of the effect systematically increasing over sequentially presented cases. This issue also provides fertile ground for professional discussion. One matter is if and how this "keep trying" mechanism should be implemented, for example, by presenting negative results in the primary text or in supplementary materials of publications. Another matter is how to distinguish negative results that do and do not warrant acknowledgment in publication (more on this below). The reader can no doubt think of many more such specific matters.

Various groups have developed standards for which study designs should be included in evidence-based practice reviews (Smith, 2012; Wendt & Miller, 2012), sometimes referred to as "meets design standards." These groups include the American Psychological Association Division 16 Task Force on Evidence-Based Interventions in School Psychology (Kratochwill & Stoiber, 2002), the Council for Exceptional Children (Cook et al., 2015; Horner et al., 2005), and the American Speech-Language-Hearing Association (2004). Among all such guidelines, only the What Works Clearinghouse *Single-Case Design Pilot Standards* make a distinction between design quality and evidence of an effect (Kratochwill et al., 2013; What Works Clearinghouse, 2014). This important distinction separates the judgment of whether research is well designed (i.e., the study would meet SCD design standards) from the judgment of whether the intervention shows an effect. In other guidelines, this distinction is lost if demonstration of a visually large functional relation (and hence a large effect) is a requirement for meeting design standards.

An analogy to the between-groups literature illustrates this point. Some guidelines consider a randomized experiment as meeting design standards, and all such studies are included whether or not they show a large effect. Only then would the reviewer calculate an effect (whether for the individual study or in a meta-analysis) and judge whether the intervention is effective. Assessment of whether a study demonstrates a functional relation in the SCD literature is analogous to the step of computing the effect size in the between-groups literature. SCD design standards should only allow an intervention the opportunity to demonstrate a functional relation; they should not require a functional relation any more than a randomized experiment should be required to have a large effect size. A good illustration of the importance of this distinction is a recent review of sensory-based treatments for children with disabilities (Barton, Reichow, Schnitz, Smith, & Sherlock, 2015). In this review, the authors summarized studies in which the investigators used a credible design to test the sensory-based treatment but found no effect on the outcome measures. These negative results are not only important from a scientific basis but also could potentially affect policy decisions for treatment options for children with disabilities.

Those who conduct evidence-based practice reviews would ideally locate the full literature, published or not. Unfortunately, finding file-drawer studies is notoriously difficult (Shadish, Doherty, & Montgomery, 1989). Fortunately, reviewers who compute appropriate effect sizes from digitized SCD data can use publication-bias methods developed for meta-analysis (Rothstein et al., 2005). These analyses do two things; first, they assess the likelihood that publication bias is present in the effect sizes being studied. Methods for doing this include simple statistical tests (e.g., Begg & Mazumdar, 1994; Egger, Davey Smith, Schneider, & Minder, 1997), inspecting a funnel plot, trim-and-fill analyses, weight function methods, and determining if effect sizes from published and unpublished studies differ in the direction predicted by publication bias (Rothstein et al., 2005). Second, some methods also estimate what the average effect size would be if all studies were included in the meta-analysis. Methods for doing this include Egger et al.'s (1997) statistical test, trim-and-fill analysis (Duval, 2005; Duval & Tweedie, 2000a, 2000b), and selection model methods (Hedges & Vevea, 1996; Vevea & Hedges, 1995). None of these methods are perfect, but their use will encourage SCD researchers to think more about whether the problem needs attention. Shadish, Hedges, and Pustejovsky (2014; see also Shadish, Hedges, Pustejovsky, Boyajian, et al., 2014) provide worked-through examples for computing all these analyses, including syntax. These meta-analytic publication-bias methods are just starting to be used in the SCD meta-analytic literature (e.g., Dart, Collins, Klingbeil, & McKinley, 2014), although their use with effect-size indices that they were not designed for, such as tau U in Dart et al. (2014), is of unknown validity. The reason is that most of these methods require knowledge of the sampling error of the effect-size statistic being used, where sampling error is a measure of the precision of the effect size. Sampling error is well established for the usual effect sizes such as *g*, *d*, *r*, and the odds ratio (Shadish & Haddock, 2009), but the overlap statistics like tau U do not at this time have either known distributions or well-derived sampling error. In between-groups studies, sample size can sometimes substitute adequately for sampling error because it is the main contributor to sampling error. That is not the case for SCD effect sizes, the precision of which reflects the number of cases, number of observations per phase, intraclass correlation, autocorrelation, and number of phase changes. Some SCD effect sizes have a well-developed standard error that explicitly takes all these factors into account (Hedges et al., 2012, 2013; Pustejovsky, Hedges, & Shadish, 2014); others might approximate this need reasonably well (e.g., Swaminathan, Rogers, & Horner, 2014), but most do not. Therefore, simply substituting sample size for a well-developed standard error will not suffice for tests of publication bias in SCD research.

The artificially generated graphs that we used as stimuli were intended to reflect common practices of SCD researchers, but they may have differed from actual graphs in several ways: (a) All conditions included the same number of data points per panel, (b) the displayed data had no trend, and (c) the graphs displayed condition means prominently in each condition. These features bolstered experimental control over the manipulations, but studies that used graphs with more realistic features might yield different results. For example, if trends were present, difficulties distinguishing trend from treatment effects might have made results less clear.

A second limitation is that we do not know if respondents' stated opinions about what they would publish correlate well with actual behavior. Their hypothetical answers to hypothetical questions based on hypothetical results may differ from what they would actually do. For example, pressure to have additional publications for tenure might lead to behaviors that differ from stated preferences, or respondents might have underreported whether they would delete cases because they were aware that doing so is a dubious practice.

A third limitation is that, although we obtained a reasonable return rate (41.2%), we do not know how well results would generalize to the full target population of SCD researchers. Respondents were slightly more likely than the target population to come from the U.S., to publish in school psychology, and to have been first author in the 2012 publications that we used to identify the respondent pool. Table 1 suggests these differences are small in percentage terms, but one can think of possible (though likely small) biases as a result. For example, we found that more experienced and well-published researchers were less likely to recommend publication of smaller effects. If first authors fit this description, then our slight overrepresentation of first authors might have biased results towards finding more reluctance to publish negative results than exists in the full population. Given the very large size of the effects in this study, however, it seems unlikely that bias would change the overall preference for large effects in publication decisions.

In summary, there are reasons to think that current professional wisdom about publication practices in the SCD research encourages researchers to think that a manuscript is not worth publishing without demonstrating a large effect. This issue might not be crucial if researchers continued to improve a treatment until it worked for all cases (Barlow, Nock, & Hersen, 2009). Kazdin (1982) hints at this rationale while simultaneously acknowledging that knowledge of negative results may still be important:

> Visual inspection may be too stringent a criterion that would reject interventions that produce reliable but weak effects. Such interventions should not be abandoned because they do not achieve large changes initially. These interventions may be developed further through subsequent research and eventually produce large effects that could be detected through visual inspection. Even if such variables would not eventually produce strong effects in their own right, they may be important because they can enhance or contribute to the effectiveness of other procedures. (p. 243)

Unfortunately, we cannot count the effectiveness of a mechanism by which one researcher publishes only positive results at the end of a program of research with mixed results. Researchers are constrained by time, resources, and external pressures like tenure that can encourage them to abandon such programs of research before they bear fruit. In addition, some interventions may never be effective no matter how much further they are developed. We need to know about these interventions to know what does not work. This knowledge allows other researchers to avoid unnecessary duplication of past efforts simply because they did not know an intervention had already been tried and failed. This knowledge also helps researchers to modify research programs that predecessors abandoned but that may still hold promise by making public the variations on a treatment that have and have not been tried to date.

This is not to say that all negative results should always be published. Pilot research intended to explore underlying processes, for example, may not contribute much to the literature in the absence of clear functional relations. Nor should a negative result deserve publication if serious question is present about whether it is attributable to problems of design or implementation. On the other hand, research intended to evaluate the efficacy and dissemination of established, emerging, or popular (but understudied) treatments should be published, regardless of the outcome, if it was rigorously designed and implemented. So, in the presence of negative findings, the researcher will always need to take into account the full context of the study rather than just the finding itself; but if the study is well designed and implemented, the mere presence of negative effects should not preclude publication.

Ultimately, evidence-based practice reviews need to tell us what does not work just as much as what does work (Kratochwill et al., 2000). If a treatment or method does not actually have an effect and previous research demonstrating that fact is not published, our knowledge of what does not work is limited, and our knowledge of the size of the effect for interventions that do work may be overstated. The SCD community needs to make it a high priority to address this dilemma by reconsidering the role that negative results should play in both publication and evidence-based practice reviews.

REFERENCES

American Speech-Language-Hearing Association. (2004). *Evidence-based practice in communication disorders: An introduction* [Technical Report].

Atkinson, D. R., Furlong, M. J., & Wampold, B. E. (1982). Statistical significance, reviewer evaluations, and the scientific process: Is there a (statistically) significant relationship? *Journal of Counseling Psychology, 29*, 189–194. doi: 10.1037/0022-0167.29.2.189

Baer, D. M. (1977). "Perhaps it would be better not to know everything." *Journal of Applied Behavior Analysis, 10*, 167–172. doi: 10.1901/jaba.1977.10-167

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

Barton, E. E., Reichow, B., Schnitz, A., Smith, I. C., & Sherlock, D. (2015). A systematic review of sensory-based treatments for children with disabilities. *Research in Developmental Disabilities, 37*, 64–80. doi: 10.1016/j.ridd.2014.11.006

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics, 50*, 1088–1101. doi: 10.2307/2533446

Chan, A., Hróbjartsson, A., Haahr, M. T., Gøtzsche, P. C., & Altman, D. G. (2004). Empirical evidence for selective reporting of outcomes in randomized trials: Comparison of protocols to published articles. *Journal of the American Medical Association, 291*, 2457–2465. doi: 10.1001/jama.291.20.2457

Cook, B. G., Buysse, V., Klingner, J., Landrum, T. J., McWilliam, R. A., Tankersley, M., & Test, D. W. (2015). CEC's standards for classifying the evidence base of practices in special education. *Remedial and Special Education, 36*, 220–234. doi: 10.1177/0741932514557271

Dart, E. H., Collins, T. A., Klingbeil, D. A., & McKinley, L. E. (2014). Peer management interventions: A meta-analytic review of single-case research. *School Psychology Review, 43*, 367–384. doi: 10.17105/SPR-14-0009.1

Duval, S. (2005). The trim and fill method. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment, and adjustments* (pp. 127–144). Chichester, UK: Wiley. doi: 10.1002/0470870168.ch8

Duval, S., & Tweedie, R. (2000a). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association, 95*, 89–98. doi: 10.1080/01621459.2000.10473905

Duval, S., & Tweedie, R. (2000b). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*, 455–463. doi: 10.1111/j.0006-341X.2000.00455.x

Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal, 315*, 629–634. doi: 10.1136/bmj.315.7109.629

Emerson, G. B., Warme, W. J., Wolf, F. M., Heckman, J. D., Brand, R. A., & Leopold, S. S. (2010). Testing for the presence of positive-outcome bias in peer review: A randomized controlled trial. *Archives of Internal Medicine, 170*, 1934–1939. doi: 10.1001/archinternmed.2010.406

Epstein, W. M. (1990). Confirmational response bias among social work journals. *Science, Technology and Human Values, 15*, 9–38. doi: 10.1177/016224399001500102

Epstein, W. M. (2004). Confirmational response bias and the quality of the editorial processes among American social work journals. *Research on Social Work Practice, 14*, 450–458. doi: 10.1177/1049731504265838

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science, 7*, 555–561. doi: 10.1177/1745691612459059

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82*, 1–20. doi: 10.1037/h0076157

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. doi: 10.1002/jrsm.1052

Hedges, L. V., Pustejovsky, J. E., & Shadish, W. R. (2013). A standardized mean difference effect size for multiple baseline designs across individuals. *Research Synthesis Methods, 4*, 324–341. doi: 10.1002/jrsm.1086

Hedges, L. V., & Vevea, J. L. (1996). Estimating effect size under publication bias: Small sample properties and robustness of a random effects selection model. *Journal of Educational and Behavioral Statistics, 21*, 299–332. doi: 10.3102/10769986021004299

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children*, *71*, 165–179. doi: 10.1177/001440290507100203

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

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

Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2013). Single-case intervention research design standards. *Remedial and Special Education, 34,* 26–38. doi: 10.1177/0741932512452794

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. doi: 10.1037/14376-004

Kratochwill, T. R., & Stoiber, K. C. (2002). Evidence-based interventions in school psychology: Conceptual foundations of the *Procedural and Coding Manual* of Division 16 and the Society for the Study of School Psychology Task Force. *School Psychology Quarterly, 17*, 341–389. doi: 10.1521/scpq.17.4.341.20872

Kratochwill, T. R., Stoiber, K. C., & Gutkin, T. B. (2000). Empirically supported interventions in school psychology: The role of negative results in outcome research. *Psychology in the Schools, 37*, 399–413. doi: 10.1002/1520-6807(200009)37:5<399:AID-PITS1>3.0.CO;2-Y

Lane, J. D., & Gast, D. L. (2014). Visual analysis in single case experimental design studies: Brief review and guidelines. *Neuropsychological Rehabilitation, 24*, 445–463. doi: 10.1080/09602011.2013.815636

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research, 1*, 161–175. doi: 10.1007/BF01173636

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. Perspectives on Psychological Science*, 7*, 615–631. doi: 10.1177/1745691612459058

Parker, R. I., & Vannest, K. (2009). An improved effect size for single-case research: Nonoverlap of all pairs. *Behavior Therapy, 40*, 357–367. doi: 10.1016/j.beth.2008.10.006

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. doi: 10.1016/j.beth.2010.08.006

Peters, D. P., & Ceci, S. J. (1982). Peer-review practices of psychological journals: The fate of published articles, submitted again. *Behavioral and Brain Sciences, 5*, 187–195. doi: 10.1017/S0140525X00011183

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team. (2014). *Nlme: Linear and Nonlinear Mixed Effects Models_. R package version 3.1-117* [Software]. Available from http://CRAN.R-project.org/package=nlme

Pustejovsky, J. E., Hedges, L. V., & Shadish, W. R. (2014). Design-comparable effect sizes in multiple baseline designs: A general modeling framework. *Journal of Educational and Behavioral Statistics, 39*, 368–393. doi: 10.3102/1076998614547577

Qualtrics. (2014). *Qualtrics.* Provo, UT. Available from http://www.qualtrics.com/

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86*, 638–641. doi: 10.1037/0033-2909.86.3.638

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005). *Publication bias in meta-analysis: Prevention, assessment and adjustments.* Chichester, UK: Wiley. doi: 10.1002/0470870168

Shadish, W. R., Doherty, M., & Montgomery, L. M. (1989). How many studies are in the file drawer? An estimate from the family/marital psychotherapy literature. *Clinical Psychology Review, 9*, 589–603. doi: 10.1016/0272-7358(89)90013-5

Shadish, W. R., & Haddock, C. K. (2009). Combining estimates of effect size. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 257–277). New York, NY: Sage.

Shadish, W. R., Hedges, L. V., & Pustejovsky, J. E. (2014). Analysis and meta-analysis of single-case designs with a standardized mean difference statistic: A primer and applications. *Journal of School Psychology, 52*, 123–147. doi: 10.1016/j.jsp.2013.11.005

Shadish, W. R., Hedges, L. V., Pustejovsky, J. E., Boyajian, J. G., Sullivan, K. J., Andrade, A., & Barrientos, J. L. (2014). A *d*-statistic for single-case designs that is equivalent to the usual between-groups *d*-statistic. *Neuropsychological Rehabilitation, 24*(3–4), 528–553. doi: 10.1080/09602011.2013.819021

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. doi: 10.3758/s13428-011-0111-y

Sham, E., & Smith, T. (2014). Publication bias in studies of an applied behavior-analytic intervention: An initial analysis. *Journal of Applied Behavior Analysis, 47*, 663–678. doi: 10.1002/jaba.146

Smith, J. D. (2012). Single-case experimental designs: A systematic review of published research and current standards. *Psychological Methods, 17*, 510–550. doi: 10.1037/a0029312

Swaminathan, H., Rogers, H. J., & Horner, R. H. (2014). An effect size measure and Bayesian analysis of single-case designs. *Journal of School Psychology, 52*, 213–230. doi: 10.1016/j.jsp.2013.12.002

Vevea, J. L., & Hedges, L.V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika, 60*, 419–435. doi: 10.1007/BF02294384

Wendt, O., & Miller, B. (2012). Quality appraisal of single-subject experimental designs: An overview and comparison of different appraisal tools. *Education and Treatment of Children, 35*, 235–268. doi: 10.1353/etc.2012.0010

What Works Clearinghouse. (2014). *What Works Clearinghouse procedures and standards handbook* (Version 3.0). Retrieved from http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v3_0_standards_handbook.pdf

Received July 26, 2015
Final acceptance December 3, 2015
Action Editor, John Borrero

APPENDIX

Journals marked with a D indicate journals categorized as disorder related, and journals marked with an E indicate journals categorized as education related. Journals without a letter did not contain SCD research in 2012 and so were not categorized.

American Journal of Intellectual and Developmental Disabilities
Autism (D)
Behavior Modification (D)
Behavior Research and Therapy
Behavior Therapy
Child and Family Behavior Therapy (D)
Education and Treatment of Children (E)
Exceptional Children (E)
Focus on Autism and other Developmental Disabilities (D)
Journal of Applied Behavior Analysis (D)
Journal of Autism and Developmental Disorders (D)
Journal of Behavioral Education (E)
Journal of Early Intervention (D)
Journal of Emotional and Behavioral Disorders (D)
Journal of Intellectual Disability Research
Journal of Learning Disabilities (E)
Journal of Positive Behavior Interventions (D)
Journal of School Psychology (E)
Journal of Special Education (E)
Journal of Speech, Language, and Hearing Research (D)
Journal of Sport and Exercise Psychology
Language, Speech, and Hearing Services in Schools (E)
Learning Disability Quarterly (E)
Neuropsychological Rehabilitation (D)
Psychology in the Schools (E)
Psychology of Sport and Exercise
Remedial and Special Education (E)
Research in Autism Spectrum Disorders (D)
Research in Developmental Disabilities (D)
School Psychology Quarterly (E)
School Psychology Review (E)
Topics in Early Childhood Special Education (E)
Topics in Language Disorders (D)