Review

Exceptional Children
2023, Vol. 90(1) 76-92
© The Author(s) 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/00144029231165505
journals.sagepub.com/home/ecx
Sage

Empirically Derived Single-Case-Design Effect Size Distributions of Engagement and Challenging Behavior in Early Childhood Research

Jason C. Chow[1], Jennifer R. Ledford[2], Sienna Windsor[2], and Paige Bennett[2]

Abstract

The purpose of this study is to present a set of empirically derived effect size distributions in order to provide field-based benchmarks for assessing the relative effects of interventions aimed at reducing challenging behavior or increasing engagement for young children with and without disabilities. We synthesized 192 single-case designs that represented data from 162 individuals and nine groups of individuals (e.g., classes) in 53 reports. We generated distributions of standardized mean difference and log-response ratio using 197 effect sizes for engagement and 146 effect sizes for challenging behavior. We examined benchmarks as a function of publication, disability, dependent variable (primary/secondary), and functional relation status and reported distributions separately by engagement and challenging behavior. Overall, the effect size distributions suggest considerable variability in the magnitude of change associated with school-based interventions for engagement and challenging behavior assessed in the context of single-case designs. Data suggest that expected magnitude of change for engagement and challenging behavior interventions may be difficult to predict and that study and effect size characteristics influence the variability of the distributions. Our results have direct implications for researchers relative to assessing the efficacy of interventions aimed at reducing challenging behavior or increasing engagement in young children with and without disabilities.

As of December 2022, Cohen's (1988) primer on statistical power analysis that includes benchmarks for standardized mean difference effect size interpretation has been cited nearly 240,000 times. Suggested benchmarks were magnitude values of 0.20, 0.50, and 0.80, to be interpreted respectively as small, medium, and large magnitudes for group design research. These arbitrary benchmarks provided a much-needed set of general agreements on effect size interpretation in the social sciences and, as such, have been used and applied to many areas of research. However, this broad application has led to a misinterpretation of Cohen's original intent that included the recommendation to consider context, intervention domain, and other relevant data sources whenever possible. Since these original benchmarks were provided to the field, researchers have made clear that

[1] University of Maryland at College Park
[2] Vanderbilt University

Corresponding Author:
Jason C. Chow, Counseling, Higher Education and Special Education, University of Maryland at College Park, 3115 Benjamin Building, 3942 Campus Dr, College Park, MD 20742, USA.
Email: jcchow@umd.edu

effect size interpretations should consider the context of specific areas of research (Harris, 2009; Hill et al., 2008; Kraft, 2020).

Studies of Effect Size Variation

Recently, Tanner-Smith et al. (2018) synthesized 1,100 studies from 74 meta-analyses that examined the empirical effect size distributions of universal mental health promotion and prevention programs for school-age youth. Authors reported on separate sets of benchmarks for effect size distributions based on the type of prevention program and outcome measure domain. Depending on the type of program and outcome measure domain of the prevention program, the median universal mental health program effect size magnitude varied from $d = -0.11$ to 0.72. In the area of autism intervention research, Chow, Zhao, et al. (2021) generated empirical distributions using 1,552 observed effect sizes from 144 early childhood autism interventions. Similar to Tanner-Smith et al., effect size magnitude varied as a function of outcome and measure properties, including the outcome domain and the outcome measure boundedness, proximity, and assessment approach. Across these benchmarks of group design studies, median values of the distributions ranged from $d = 0.12$ to 0.55, suggesting that the middle effect size values in the early childhood autism intervention literature ranged substantially.

Researchers in education (Bakker et al., 2019; Bloom et al., 2008; Taylor et al., 2018) and criminal justice (Lipsey & Cullen, 2007) have conceptualized similar work related to differences in effect sizes. Bloom et al. (2008) reported that effect sizes of educational interventions varied by the grade level, socioeconomic status, and race-ethnicity. Taylor et al. (2018) reported the treatment effect of science education interventions also varied from −0.044 to 0.149 as a function of study design, population, interventionist, science content area, assessment type, and grade level. Cheung and Slavin (2016) reported on the influence of assessment type on effect size magnitude of educational interventions by synthesizing the results from 645 studies reported in 12 reviews from the Best Evidence Encyclopedia. The authors also reported effect sizes by type of publication and sample size, finding that effect sizes were almost twice as large for those published studies with small sample sizes and researcher-developed measures, compared with the unpublished studies that used large sample sizes with standardized assessments.

To generate an appropriate interpretation of experimental results, effect sizes should be interpreted in conjunction with active consideration of the specific context in which the study was situated as well as other technical or methodological characteristics of the research (Bakker et al., 2019; Chow, 2020; Hill et al., 2008). Together, these studies provide evidence that appropriate interpretation of treatment effect sizes requires a nuanced approach.

The Need for Effect Size Benchmarks in Single-Case-Design Research

Recently in special education, researchers have adopted effect size metrics in an effort to standardize the interpretation of single-case-design (SCD) research, continuing the proliferation of effect sizes in quantitative research methods. However, effect sizes applicable for differences between groups cannot be directly applied to SCD data, and little methodological research has been conducted to provide guidance for interpretation of effect size magnitudes. The consequences of the lack of guidance can lead to arbitrary interpretations that lack empirical expected values. We argue that empirical benchmarks for effect size magnitude are needed in order to provide the field with appropriate guidelines to interpret study effects as well as predict anticipated effects. Given the seemingly increasing ubiquity of effect sizes in quantitative research, planful guidance for the field is necessary to allow for important, relative comparisons that support continuous knowledge generation in what works for improving outcomes for children. This may be especially important in special education, school psychology, and clinical psychology research given the prevalence of SCD studies in these fields.

Because SCD research relies on different assumptions and methods than group design research (e.g., randomized controlled trials [RCT]), effect size estimation across the two types of research is theoretically and practically divergent (Maggin et al., 2019). That is, when SCDs are used, condition ordering (e.g., time-lagged implementation, rapid iterative alternation; Ledford & Gast, 2018; What Works Clearinghouse [WWC], 2020) is used to control for and identify threats to internal validity, and changes in behavior over time are evaluated repeatedly for a single case (usually a single participant). SCD is also conceptually different because they are based on within-case comparisons and effectiveness is based directly on replication of effects (i.e., functional relations). Conversely, effects of a group RCT are based on comparing the mean performance of participants between a group that does and does not receive the intervention, and there are typically some participants assigned to the control group who demonstrate comparable gains to those in the intervention group (the groups have overlapping distributions). Visual analysis, which determines whether intervention changes are functionally related to intervention implementation, is the traditional method of analysis for SCD. Effect sizes are recommended to complement visual analysis (Barton et al., 2017; WWC, 2020) and represent the magnitude of behavior change between conditions rather than whether a functional relation exists. As expected, effect size estimates for single-case research are somewhat dissimilar in magnitude from those identified in similar group design studies (e.g., Barton et al., 2017).

Two commonly used effect sizes appropriate for SCD are the log response ratio (LRR) and the standardized mean difference (SMD). LRR is an effect size for single-case research studies that is closely related to response ratios used to evaluate group research (Pustejovsky, 2018). The within-case SMD is closely related to the most commonly used effect sizes in group research, Cohen's *d* and Hedges's *g*. Whereas LRR describes relative differences between conditions (i.e., ratios), SMD describes mean differences between conditions relative to the standard deviation. SCDs are comparing data between phases, and group designs are comparing data between the means of two groups. As such, both metrics describe within-participant differences, which differs from common-group-design (between-group) effect size metrics. Interpretive guidelines for these metrics (e.g., benchmarks for small, medium, and large effects) have not been published. We also based our use of LRR and SMD on ease of interpretation for LRR (researchers can calculate an index of percentage change) and recommend use of between-case SMD by the Institute of Education Sciences. As mentioned already, SMD is mathematically similar to group-design, between-group-comparison effect sizes.

In this study, we generate and synthesize effect sizes to produce empirically derived effect size distributions for SCD experiments of school-based interventions. We intend for these data to help researchers interpret effect sizes, the quantitative reflection of the magnitude of a phenomenon used to address a question or problem (Kelley & Preacher, 2012), in SCD research. We selected engagement and challenging behavior for synthesis because they are two commonly measured variables in SCD research, and they are measured in ways that allow for use of common effect size metrics (e.g., SMD, LRR; Pustejovsky, 2018; Shadish et al., 2014).

Purpose

The purpose of this study was to generate empirically driven effect size distributions in SCD research to provide a data-based method for assessing the relative magnitude of intervention effects. This will provide researchers with benchmarks that allow comparisons between an observed finding from a single study to what is documented in the broader literature. Specifically, we synthesize experimental data from interventions aimed at reducing challenging behavior or improving engagement in preschool-age children. This synthesis of effect size distributions of SCD effect sizes will (a) support realistic, data-based expectations for intervention effectiveness and (b) provide an appropriate mechanism for researchers to assess the promise or success of individual studies in demonstrating meaningful behavior change.

This study prospectively explores sources of variation in SCD effect sizes. Specifically, we examine effect size distributions by broad outcome (i.e., engagement, challenging behavior), and we also descriptively examine differences based on theoretically and empirically determined moderators: (a) effect size sample (e.g., whether the effect size represents data from a group of children compared with an individual child), (b) child-level characteristics (i.e., disability status), publication status (i.e., if the study was published in a peer-reviewed journal compared with being an unpublished dissertation), (c) whether or not the measure was a primary or secondary outcome (i.e., whether the variable was used to make experimental decisions or whether it was a corollary behavior), and (d) functional relation determination.

Method

Summary of Coding Team

Members of the study team included two faculty, four doctoral students, and five master's students, all in special education departments. Both faculty members and all doctoral students completed title and abstract screening. Both faculty members and three doctoral students completed full-text screening and article coding. The five master's students were trained to reliably extract effect size data using PlotDigitizer (Huwaldt & Steinhorst, 2015) for data extraction for several ongoing SCD research reviews; their point-by-point reliability was high (98%), and we did not provide additional training for data extraction specific to this study. We double coded all sources at the title-and-abstract-screening phase. Then, we resolved discrepancies by identifying all sources where coders did not agree and reviewing as a team during weekly meetings. For the first round of full-text screening, we made a resource-based decision to not double screen the full texts of the 1,248 sources at this phase. We then double coded the full texts of all included sources.

Inclusion Criteria and Literature Search

Included sources met the following criteria: (a) full-text publication in English; (b) inclusion of young children served in early childhood (before kindergarten) settings with or without disabilities, with data collected in the school-based setting (preschool, daycare, prekindergarten); (c) observational data collected in the context of a SCD with at least three potential demonstrations of effect and three data points in each condition, and data were presented via line graph to allow for data extraction; (d) inclusion of measurement of observational data on engagement or on-task behaviors, disengagement or off-task behaviors, or challenging behaviors (see our supplementary materials for detailed descriptions of our definitions); and (e) inclusion of a nonpharmacological intervention in comparison to a baseline or control condition. We excluded nonconcurrent time-lagged designs, graphs with nonadjacent conditions (e.g., A-B-C-A-B), and studies using measurement systems that were not on a ratio scale (e.g., rating scales).

We searched PsycINFO, PubMed, and ProQuest Dissertations and Theses using parallel search strategies for each of the behavior constructs of interest. Search 1 included terms for problem behavior: ((ab("problem behavio*" OR "behavio* problems" OR aggress* OR disruptive OR challenging)) AND (class* OR school*) AND ab(treatment* OR intervention* OR prevent* OR training*)). Search 2 included terms for engagement: ((ab(engage* OR "on-task" OR "on task")) AND (class* OR school*) AND ab(treatment* OR intervention* OR prevent* OR training*). Both searchers included the same set of design terms: ab("single-subject" OR "single subject" OR "single-case" OR "single case" OR "multiple baseline" OR "multiple-baseline" OR "multiple probe" OR "multiple-probe" OR "changing criterion" OR "withdrawal" OR "ABAB" OR "A-B-A-B" OR "reversal" OR "alternating treatment*")). We deduplicated records from the .csv files we retrieved prior to article screening. It is important to note that this study was a part of a larger project, and the original search included participants from ages 0 to 21. We divided eligible sources into grade-based categories (preschool, elementary school, middle school, high school, mixed). The sources included in this review included only children attending early childhood (before kindergarten) school programs.

Article Screening

We screened articles at the title and abstract levels using the online citation screening tool Abstrackr followed by full-text screening of article PDFs. We requested PDFs that were unavailable at either of the libraries accessible by the first two authors via interlibrary loan. We did not screen sources that were unavailable via interlibrary loan. At the full-text level, all authors conducted preliminary full-text screening of all articles from the original search. These included studies of all eligible age groups (0 to 21). This process involved ensuring each study had eligible participants, designs, and variables. We systematically screened and sorted full-text articles into groups based on age. Though we report the larger search for transparency, the present study includes data from sources that included children attending early childhood education programs. The codebook used for abstract and full-text screening is available in the online supplemental materials.

Article Coding

Four authors coded all sources using REDCap electronic data-capture tools hosted at Vanderbilt University (Harris et al., 2009). We conducted several rounds of practice coding to calibrate codes, operational definitions, and the coding form structure. In order to ensure agreement, we double coded all studies independently. Then, our team met weekly to discuss any disagreements, come to group consensus, and resolve the disagreement in our final coding spreadsheet. The codebook is available in the online supplemental materials.

We coded data at the source, participant, and outcome levels. For each included source, we coded year of publication (for published sources) or completion (for unpublished sources), publication status, journal name (if applicable), design type (e.g., multiple-baseline across participants, A-B-A-B, alternating treatments design), and nature of comparison (e.g., demonstration, comparison, combination).

At the individual-participant level, we coded the identifier (e.g., pseudonym used by the author), gender, age, race, ethnicity, and primary and secondary disabilities for each participant. If it was unclear if the participant had a disability and the coders could not reasonably infer a disability, that participant was coded as not having a disability. We coded for Individuals With Disabilities Education Act–designated disability categories and coded the disability only if the participant was specified as having the disability or was receiving special education services. We did not code screening data or “at risk” status as having a disability. If data were presented for groups (e.g., classwide data), we coded the identifier, percentage of individuals who were male, mean age, percentage for each race and ethnicity category, and whether children with disabilities were included in the group.

At the outcome level, we coded data for each individual design within the study. For example, if study authors conducted two A-B-A-B designs (one for engagement and one for challenging behavior) for each of two participants, we coded four designs. For each design included in the analysis, we coded the identifier, figure number, data series identifier, dependent variable type, design type, whether there was a sufficient number of potential demonstrations and data points, whether baseline conditions allowed for potential demonstration of experimental control, and if so, whether a functional relation existed. In this study, we used visual analysis to determine the presence or absence of a functional relation by evaluating within- and between-condition data patterns using design-specific guidelines proposed by Ledford et al. (2022). Specifically, functional relations were identified when consistent, immediate, and therapeutic changes in data patterns occurred concurrently with changes between baseline and intervention conditions. Baseline data not allowing for experimental control were those for which, even if treatment data were optimal, we coded as "no functional relation could be identified." Examples include data at floor or ceiling levels (e.g., problem behavior occurring 0% of the time during baseline sessions) or trending to floor or ceiling levels (e.g., problem behavior occurring 0% of the time during the final two baseline sessions). We also coded whether the dependent variable represented the primary outcome or a secondary outcome of the study (see rules in coding document).

Interobserver Agreement (IOA)

We double coded all sources at the title-and-abstract-screening phase. Then, we resolved discrepancies. We then conducted full-text screening; coders were assigned batches to conduct independent full-text review. We conducted several rounds of practice coding to calibrate codes, operational definitions, and the coding form structure. In order to ensure agreement, we double coded all studies independently. Our team met weekly to discuss any disagreements, come to group consensus, and resolve the disagreement in our final coding spreadsheet. Each of the four authors served as the primary coder for approximately 20 studies and secondary coder (IOA) for approximately 20 studies.

Data Extraction

We extracted data directly from study sources using graphical presentations provided by study authors using PlotDigitizer (Huwaldt & Steinhorst, 2015). We digitized data from baseline and intervention conditions only (i.e., did not include maintenance, follow-up, or generalization data). We made corrections to data that were below 0 and above 100 (when data were reported as a percentage of time, opportunities, or intervals), given that these data values are not possible (e.g., data extracted as −0.01 were identified automatically in Microsoft Excel and transformed to 0, and data extracted as 100.1 were transformed to 100). Plot digitized data are available via online supplementary materials.

Data Analysis

We selected SMD and LRR effect sizes given relative robustness to procedural variabilities (Pustejovsky, 2019). We note that these effect sizes are calculated using different strategies and thus that the values between SMD and LRR are not directly comparable. We used the SingleCaseES package (Pustejovsky & Swan, 2018) in the R statistical environment (R Core Team, 2018) to calculate effect sizes. For the purposes of analysis, we identified two effect sizes with the most desirable characteristics (Pustejovsky, 2019): LRR and SMD. LRR has two variations: LRRi (increasing) for data expected to increase during intervention conditions (e.g., engagement) and LRRd (decreasing) for data expected to decrease during intervention conditions (e.g., off-task, challenging behaviors). For LRRi, larger values indicate more therapeutic changes (e.g., higher levels of engagement in intervention conditions), whereas for LRRd, smaller values indicate more therapeutic changes, with negative values interpreted as therapeutic decreases (e.g., lower levels of challenging behavior in intervention conditions). LRR cannot be calculated if the mean of any condition is zero, and SMD cannot be calculated if there is no variability in baseline conditions, resulting in the removal of nine comparisons. Statistical code and data are available via online supplementary materials.

After we calculated effect sizes, we removed outliers (effect size calculations that were more than three standard deviations from the median; seven LRR effect sizes and 22 SMD effect sizes) and then generated effect size distributions and empirical benchmarks of these distributions for each primary outcome (engagement, challenging behavior). Specifically, we created effect size distributions for each outcome domain across all interventions using the median and interquartile ranges of corresponding effect size metrics

(LRR and SMD) to establish empirical benchmarks. Because the purpose of our study was to summarize the observed empirical effect size distributions in the research literature, we did not apply inferential statistical analyses to these data. Rather, we focused on generating effect size distributions that represent the observed and comparable magnitudes in SCD interventions for young children. In order to provide a detailed description of the empirical distributions for interventions with different outcomes, we further summarized the effect size distributions for interventions that varied based on broad outcome (e.g., engagement, challenging behavior), effect size sample (e.g., whether the effect size represents data from a group of children compared with an individual child), child-level characteristics (e.g., gender, disability status), publication status (i.e., if the study was published in a peer-reviewed journal compared with being an unpublished dissertation), and whether or not the measure was a primary or secondary outcome.

Results

Initial electronic database searches yielded 4,522 records and a final corpus of 1,248 unique reports after deduplication. After our systematic search process (see Figure 1), we identified 53 sources, 34 of which were published in peer-reviewed journals; the remainder were dissertations or theses. These sources included 192 SCDs. We generated effect size distributions using 197 effect sizes for engagement and 146 effect sizes for challenging behavior. Generally, behavior was measured with partial interval recording or momentary time sampling. Intervention and assessment most often occurred with endogenous implementers (n = 124 for

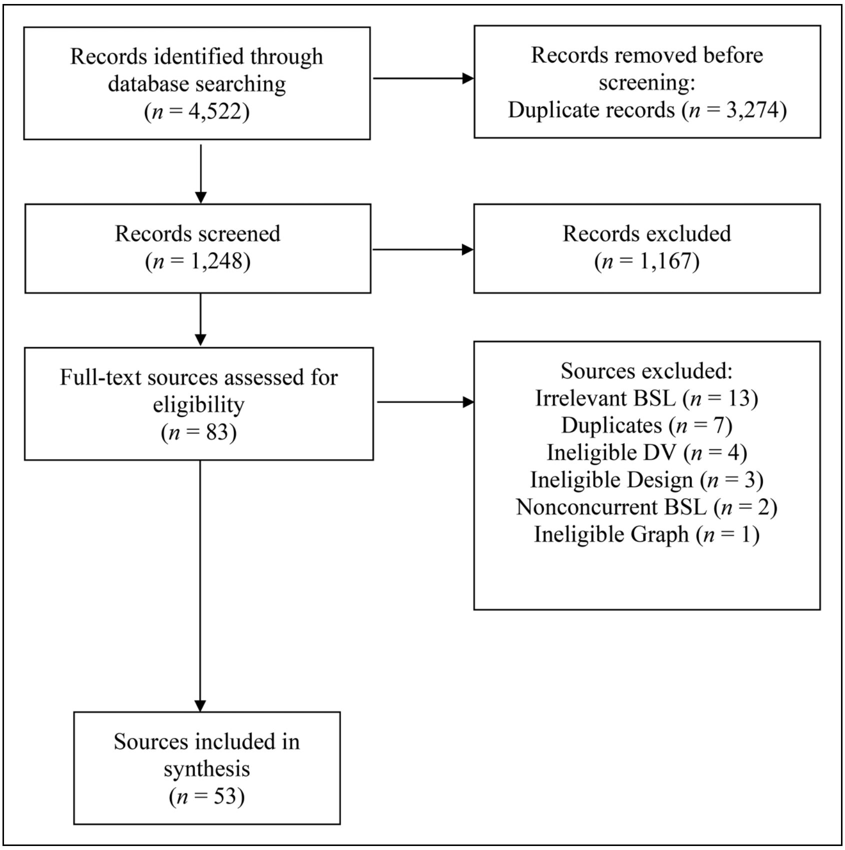
Records identified through database searching (n = 4,522)
Records removed before screening: Duplicate records (n = 3,274)
Records screened (n = 1,248)
Records excluded (n = 1,167)
Full-text sources assessed for eligibility (n = 83)
Sources excluded: Irrelevant BSL (n = 13) Duplicates (n = 7) Ineligible DV (n = 4) Ineligible Design (n = 3) Nonconcurrent BSL (n = 2) Ineligible Graph (n = 1)
Sources included in synthesis (n = 53)

Figure 1. Visualization of study idenfication process.

intervention, n = 136 for assessment) and in individual (n = 88 for intervention, n = 63 for assessment) or whole-group (n = 61 for intervention, n = 64 for assessment) activities. For designs, we included a total of 192, with 57 A-B-A-B designs, 49 multiple-baseline or multiple-probe designs, and 86 alternating-treatments designs (including ongoing control conditions for baseline vs. intervention comparisons). See Appendix C in the online supplemental materials for tables with additional description information.

Most sources included participants with behavior measured individually (e.g., one participant's data in an A-B-A-B design, n = 163), and nine sources measured group behavior (e.g., classwide data in an A-B-A-B design). When individual participants were included (n = 162), most participants were male (n = 120) and did not have a disability (n = 119). The most common primary disabilities reported were autism (n = 46) and developmental delay (n = 25). Race and ethnicity were rarely reported; when reported (n = 65), race was often reported as White (n = 30) or Black (n = 28). In terms of ethnicity, studies reported ethnicity on only 30 participants; of these participants, 16 were reported as Hispanic and 14 were reported as not Hispanic; see Table 1 for detailed participant information. For studies that measured group behavior, participant characteristics were described using various methods, making data difficult to synthesize. As such, we provide summary descriptive information of these studies in Appendix C in the online supplemental materials.

Effect Size Distributions

We generated effect size distributions for each construct (challenging behavior and engagement) by first presenting the overall distributions of our selected effect size metrics (SMD and LRR). As shown in the scatterplots of SMD and LRR values (Figure 2; for ease of interpretation, LRRd values are transformed such that positive values are interpreted as therapeutic for challenging behavior outcomes), the SMD distributions were wider for both challenging behavior and engagement (e.g., values on the vertical axis are larger than those on the horizontal axis). SMD and LRR values were more highly correlated for engagement (r = .80) than for challenging behavior (r = .34). We describe the distribution data for challenging behavior and engagement that can also be found represented in Tables 2 and 3.

Challenging Behavior. The overall distribution for SMD effect sizes for challenging behavior had a median value of 1.16 with a first quartile

Table 1. Participant Information.

Gender	Primary disability	Secondary disability	Race	Ethnicity	Age
Male = 120	None = 49	None = 119	White = 30	Hispanic = 16	<3 = 2
Female = 35	ASD = 46	ASD = 3	Black = 28	NH = 14	3 = 30
NR = 7	EBD = 6	EBD = 1	Asian = 3	NR = 132	4 = 81
	OHI = 4	OHI = 2	AI = 1		5 = 26
	DD = 25	DD = 2	Multiple = 2		6 = 12
	ID = 7	ID = 6	Other = 1		7 = 1
	SLI = 11	SLI = 13	NR = 97		NR = 10
	OI = 1	SLD = 1			
	VI = 3	OI = 1			
	SLD = 1	Other = 8			
	Other = 3	Multiple = 6			
	Multiple = 6				

Note. NR = not reported; ASD = autism spectrum disorder; EBD = emotional and behavioral disorders; OHI = other health impairment; DD = developmental delay; ID = intellectual disability; SLI = speech and language impairment; SLD = specific learning disability; OI = orthopedic impairment; AI = American Indian; ME = Middle Eastern; NH = not Hispanic.

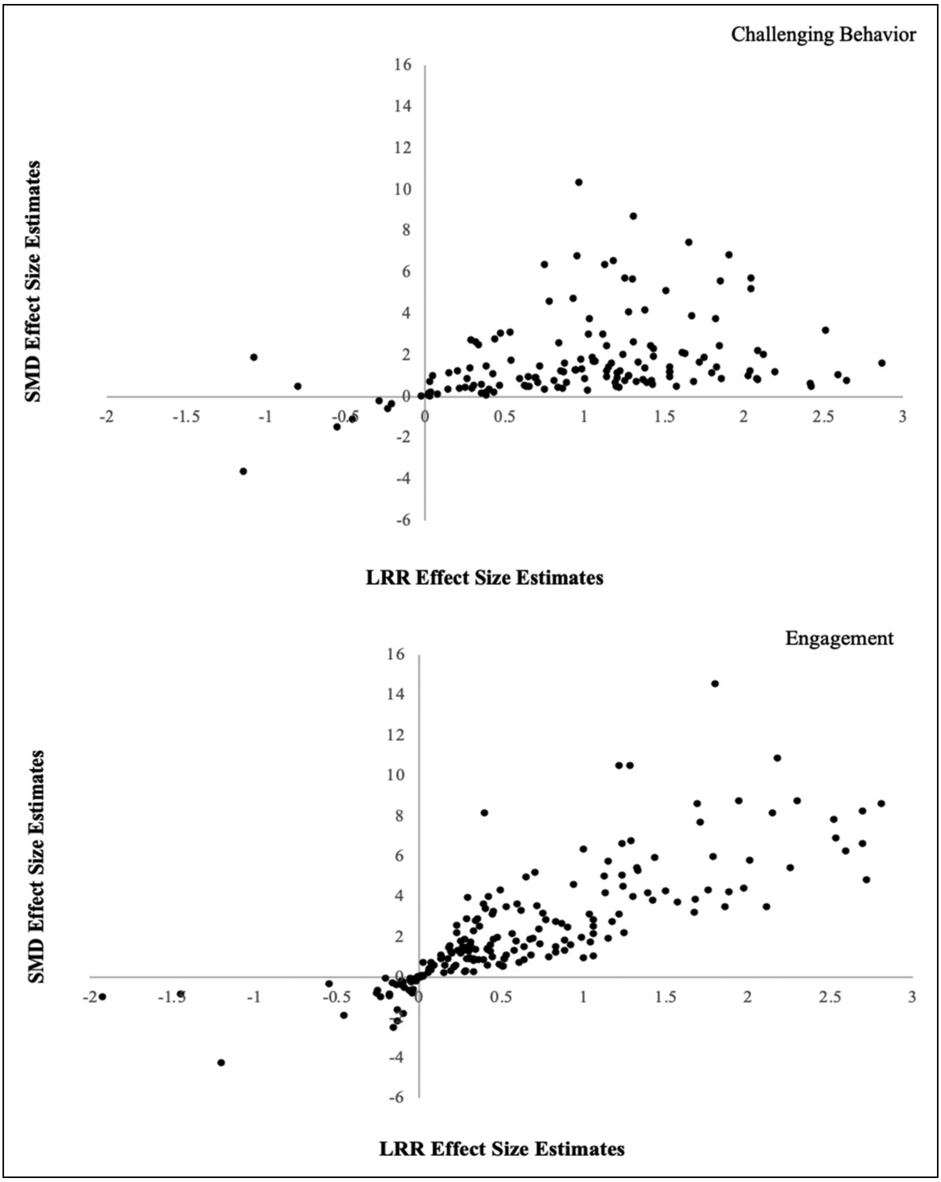

Figure 2. Scatter plot of effect size estimate metrics for challenging behavior and engagement.

effect size of 0.60 and a third quartile value of 2.22. The overall distribution for LRRd effect sizes for challenging behavior had a median value of –1.04 with a first quartile effect size of –0.47 and a third quartile value of 1.42. Only 4% (SMD) and 6% (LRR) of effect sizes were counter therapeutic. See Table 2 for overall and subcategory effect size distributions.

Publication status. The distribution for SMD effect sizes generated from published studies had a median value of 1.21 with a first quartile effect size of 0.79 and a third quartile value of 2.11. The distribution for SMD effect sizes generated from unpublished studies had a median value of 0.96 with a first quartile effect size of 0.43 and a third quartile value of 2.47. The distribution for LRRd effect sizes generated from published studies had a median value of –1.25 with a first quartile effect size of –0.94 and a third quartile value of –1.67. The distribution for LRRd effect

Table 2. Effect Size Distributions for Challenging Behavior.

Comparison		Min.	1st	*Mdn*	3rd	Max.
SMD	Overall (*n* = 147)	–3.61	0.60	1.16	2.22	10.34
	Published (*n* = 93)	–1.07	0.79	1.21	2.11	10.34
	Unpublished (*n* = 54)	–3.61	0.43	0.96	2.47	6.36
	Primary (*n* = 105)	–3.61	0.84	1.24	2.59	8.74
	Secondary (*n* = 42)	–1.08	0.36	0.66	1.41	10.34
	FR (*n* = 81)	0.45	1.19	1.77	3.15	10.34
	No FR (*n* = 66)	–1.08	0.33	0.66	1.03	5.70
	Disability[a] (*n* = 53)	–3.61	0.51	0.93	2.58	10.34
	No disability[a] (*n* = 79)	–1.08	0.74	1.11	1.85	6.87
LRRd	Overall (*n* = 147)	1.14	–0.47	–1.04	–1.42	–2.87
	Published (*n* = 93)	0.46	–0.94	–1.25	–1.67	–2.65
	Unpublished (*n* = 54)	1.14	–0.21	–0.54	–1.03	–2.87
	Primary (*n* = 105)	1.14	–0.63	–1.14	–1.45	–2.65
	Secondary (*n* = 42)	0.80	–0.38	–0.90	–1.29	–2.87
	FR (*n* = 82)	1.08	–0.95	–1.22	–1.57	–2.51
	No FR (*n* = 65)	0.80	–0.14	–0.53	–1.07	–1.86
	Disability[a] (*n* = 80)	1.14	–0.38	–0.95	–1.29	–2.87
	No disability[a] (*n* = 53)	0.46	–0.69	–1.31	–1.80	–2.65

Note. SMD = standardized mean difference; LRR = log-response ratio; FR = functional relation. For LRRd, the most positive values (e.g., "min.") indicate a countertherapeutic change (e.g., an increase in challenging behavior during intervention) and the smallest values (e.g., "max") indicate the largest therapeutic changes.

[a] Sum is not 147 because group data are excluded.

Table 3. Effect Size Distributions for Engagement.

Comparison		Min.	1st	*Mdn*	3rd	Max.
SMD	Overall (*n* = 190)	–4.23	0.46	1.52	3.52	14.57
	Published (*n* = 83)	–4.23	0.91	1.82	3.82	14.57
	Unpublished (*n* = 107)	–2.49	–0.03	0.97	3.18	10.88
	Primary (*n* = 129)	–4.23	0.04	1.56	3.29	14.57
	Secondary (*n* = 61)	–0.76	0.87	1.44	3.96	10.88
	FR (*n* = 106)	–4.23	1.35	2.79	4.37	14.57
	No FR (*n* = 84)	–2.48	–0.21	0.29	1.44	8.75
	Disability (*n* = 144)	–4.23	0.29	1.35	3.54	14.57
	No disability (*n* = 25)	0.02	1.14	1.92	3.96	8.75
	Classwide (*n* = 21)	0.02	1.14	1.92	3.96	8.75
LRR	Overall (*n* = 204)	–1.92	0.15	0.45	1.15	2.96
	Published (*n* = 113)	–1.92	0.23	0.61	1.18	2.70
	Unpublished (*n* = 91)	–0.54	–0.02	0.33	1.06	2.96
	Primary (*n* = 137)	–1.92	0.02	0.44	1.09	2.81
	Secondary (*n* = 67)	–0.05	0.22	0.49	1.22	2.96
	FR (*n* = 112)	–1.20	0.34	0.69	1.33	2.81
	No FR (*n* = 92)	–1.92	–0.05	0.11	0.72	2.96
	Disability (*n* = 155)	–1.92	0.14	0.49	1.13	2.96
	No disability (*n* = 28)	0.00	0.30	0.77	1.38	2.30
	Classwide (*n* = 21)	–0.09	0.05	0.26	0.43	2.70

Note. SMD = standardized mean difference; LRR = log-response ratio; FR = functional relation.

sizes generated from unpublished studies had a median value of –0.54 with a first quartile effect size of –0.21 and a third quartile value of –1.03. These data suggest that published studies, on average, were associated with larger effect sizes than unpublished studies (i.e., larger decreases in challenging behavior; for LRRd, this is indicated via smaller values).

Primary or secondary measure. The distribution for SMD effect sizes generated from measures that were identified as primary study measures had a median value of 1.24 with a first quartile effect size of 0.84 and a third quartile value of 2.59. The distribution for SMD effect sizes generated from measures that were identified as secondary study measures had a median value of 0.66 with a first quartile effect size of 0.36 and a third quartile value of 1.41. The distribution for LRRd effect sizes generated from measures that were identified as primary study measures had a median value of –1.14 with a first quartile effect size of –0.63 and a third quartile value of –1.45. The distribution for LRRd effect sizes generated from measures that were identified as secondary study measures had a median value of –0.90 with a first quartile effect size of –0.38 and a third quartile value of –1.29. Both SMD and LRRd effect size distributions for primary measures had larger average effects (i.e., larger decreases in challenging behavior; for LRRd, this is indicated via smaller values) and more variability than those for secondary measures.

Functional relation determination. For designs that were determined by coders to represent a functional relation, the distribution of SMD effect sizes had a median value of 1.77. The distribution of effect sizes for designs determined to have no functional relation had a median value of 0.66. The distribution of LRRd effect sizes generated from studies with a functional relation had a median value of – 1.22, and the median value of the distribution from designs with no functional relation was – 0.53. Thus, for both SMD and LRRd, effect sizes were considerably larger on average (i.e., larger decreases in challenging behavior; for LRRd, this is indicated via smaller values) and were more variable when they were associated with a design for which a functional relation had been identified.

Disability status. The distribution for SMD effect sizes for children with disabilities had a median value of 0.93 with a first quartile effect size of 0.51 and a third quartile value of 2.58. The distribution for SMD effect sizes for children without disabilities had a median value of 1.11 with a first quartile effect size of 0.74 and a third quartile value of 1.85. The distribution for LRRd effect sizes for children with disabilities had a median value of –0.95 with a first quartile effect size of –0.38 and a third quartile value of –1.29. The distribution for LRRd effect sizes for children without disabilities had a median value of –1.31 with a first quartile effect size of –0.69 and a third quartile value of –1.80. Median values were larger, and the variability was greater for effect sizes associated with children without disabilities (i.e., larger decreases in challenging behavior are indicated via smaller values for LRRd).

Engagement. The overall distribution for SMD effect sizes for engagement had a median value of 1.52 with a first quartile effect size of 0.46 and a third quartile value of 3.52. The overall distribution for LRRi effect sizes for engagement had a median value of 0.45 with a first quartile effect size of 0.15 and a third quartile value of 1.15. For LRRi, 17.9% of effect sizes were less than zero, and the same was true for 17.3% of SMD effect sizes. See Table 3 for overall and subcategory effect size distributions.

Publication status. The distribution for SMD effect sizes generated from published studies had a median value of 1.82 with a first quartile effect size of 0.91 and a third quartile value of 3.82. The distribution for SMD effect sizes generated from unpublished studies had a median value of 0.97 with a first quartile effect size of –0.03 and a third quartile value of 3.18. The distribution for LRRi effect sizes generated from published studies had a median value of 0.61 with a first quartile effect size of 0.23 and a third quartile of 1.18. The distribution for LRRi effect sizes generated from unpublished studies had a median value of 0.33 with a first quartile effect size of –0.02 and a third quartile value of 1.06. Estimates from both LRRi and SMD effect sizes suggest that unpublished data sources yield smaller and less variable effect sizes than published data sources.

Primary or secondary measure. The distribution for SMD effect sizes generated from measures that were identified as primary study measures had a median value of 1.56 with a first quartile effect size of 0.04 and a third quartile value of 3.29. The distribution for SMD effect sizes generated from measures that were identified as secondary study measures had a median value of 1.44 with a first quartile effect size of 0.87 and a third quartile value of 3.96. The distribution for LRRi effect sizes generated from measures that were identified as primary study measures had a median value of 0.44 with a first quartile effect size of 0.02 and a third quartile value of 1.09. The distribution for LRRi effect sizes generated from measures that were identified as secondary study measures had a median value of 0.49 with a first quartile effect size of 0.22 and a third quartile value of 1.22. The median values for primary and secondary measures were similar for both SMD (primary = 1.56, secondary = 1.44) and LRRi (primary = 0.44, secondary = 0.49).

Functional relation determination. For designs that were determined by coders to represent a functional relation, the distribution of SMD effect sizes had a median value of 2.79. The distribution of SMD effect sizes for designs determined to have no functional relation had a median value of 0.29. The distribution of LRRi effect sizes generated from studies with a functional relation had a median value of 0.69, whereas the median value of the distribution from designs with no functional relation was 0.11. Thus, for both SMD and LRRi, effect sizes were considerably larger when they were associated with a design for which a functional relation had been identified.

Disability status. We then generated distributions that were based on effect sizes from samples of students with and without an identified disability. The distribution for SMD effect sizes for children with disabilities had a median value of 1.35 with a first quartile effect size of 0.29 and a third quartile value of 3.54. The distribution for SMD effect sizes for children without disabilities had a median value of 1.92 with a first quartile effect size of 1.14 and a third quartile value of 3.96. The distribution for LRRi effect sizes for children with disabilities had a median value of 0.49 with a first quartile effect size of 0.14 and a third quartile value of 1.13. The distribution for LRRi effect sizes for children without disabilities had a median value of 0.77 with a first quartile effect size of 0.30 and a third quartile value of 1.38. Effect sizes for participants without versus with disabilities were larger for both SMD and LRRi.

Discussion

Overall, the calculated effect size distributions suggest considerable variability in the magnitude of change associated with school-based interventions for engagement and challenging behavior assessed in the context of SCDs, with the majority of designs resulting in positive effects, and more positive effects for challenging behavior (4%–6% countertherapeutic values) compared with engagement (17%–18% countertherapeutic effects). These data suggest that expected magnitude of change for engagement and challenging behavior interventions may be difficult to predict but that a large minority of interventions for engagement are not effective. Although outside the scope of this review, additional research that identified components associated with smaller or larger effects (e.g., specific intervention components) is warranted. In comparison with SMD benchmarks for group-design studies and estimates of average effect sizes from those studies (e.g., Sandbank et al., 2021), the values identified in this study were generally large. That is,

effect sizes from these single-case studies were consistently larger than those identified by reviews of group-design research, including those focused on challenging behavior and *engagement (e.g., Luo et al., 2020). The reasons for this are likely multifaceted, but two likely explanations are the use of researcher-derived direct observation measures (rather than standardized measures) and the common emphasis on context-bound behavior associated with SCD (cf., Ledford & Windsor, 2021).*

Consistent with what has been colloquially called the file-drawer effect and with previous research (Ekholm & Chow, 2018; Gage et al., 2017), average effects sizes for engagement were nearly double for published versus unpublished sources for both LRRi and SMD. For challenging behavior, SMD effect sizes were larger for published studies, but only LRRi effect sizes had a similar pattern in which published sources were associated with an average effect size of approximately double that from unpublished sources. Regardless, this outcome provides empirical support for previous suggestions that attempts to aggregate effect sizes should include sources from gray literature, such as dissertations and theses (Chow et al., 2021; Cumming et al., 2022; Pigott & Polanin, 2020).

More equivocal results were noted for variables identified as primary versus secondary variables. Our a priori hypothesis was that effect sizes would be larger for primary variables, given these are the behaviors on which experimental decisions are made and, based on previous research, that showed that functional relations were more commonly identified for primary than for secondary variables (Ledford & Windsor, 2021). These equivocal results could be explained because effect sizes may be less sensitive to internal validity concerns than functional relation determinations based on visual analysis. However, average effect sizes were smallest for comparisons categorized as not demonstrating a functional relation compared with those with no functional relation. Effect sizes were associated with A-B comparisons, and functional relations were demonstrated on a design level (e.g., one A-B comparison for alternating treatments designs, two A-B comparisons for A-B-A-B designs, and at least three A-B comparisons for multiple-baseline designs), so we would not expect exact convergence, but general alignment of the two metrics indicates a relation between the two methods of assessing single-case data. Thus, we support continuing calls for effect sizes to be used alongside systematic visual analysis (Maggin et al., 2019).

In relation to participant characteristics, both LRR and SMD average effect sizes were smaller for children with disabilities in comparison to children without disabilities. This may suggest that between-condition changes for children without disabilities are likely to be larger than changes for children with disabilities. The reasons for this are beyond the scope of this review but could be explained by any number of factors, including a difference in history of supports between groups and the type and intensity of interventions used for children with and without disabilities. Though tangential to the primary purpose of this review, our findings align with other reviews that document lack of reporting of race (60% not reported) and ethnicity (over 80% not reported) in special education research (Chow et al., 2022; Robertson et al., 2017).

We generated effect size distributions that represented the observed magnitudes of outcomes in the field of early childhood research, relying on quartiles to compare effect size values to available research rather than to assign categories to these values (e.g., small, medium, large). That is, researchers might interpret a medium effect as one that represents a value near the median and a large effect as one that represents a value greater than 75% of studies. However, there are difficulties with this interpretation, given the possibility that many studies did not reflect positive changes between conditions or that many studies resulted in meaningful effects. That is, an effect size larger than 75% of included studies does not necessarily translate into a *practically* large change in behavior.

Future Directions

This review provides several directions for future work, primarily for researcher audiences. We synthesized the effect sizes for studies including young children that aimed at improving engagement or reducing challenging behavior. We limited our scope to two effect size metrics as well as measures using direct observational systems. There are many ways in which future studies can expand on this work to include different populations and types of interventions and outcomes. For example, given the variability within the broad category of challenging behavior (e.g., frequency of aggressive behavior, eloping, talking out of turn), a more nuanced analysis of effect size variation by type of challenging behavior could provide more specific insight into differential response to intervention. Future studies can pursue a deeper dive into variation in effects as a function of measurement systems used as well as measurement properties, such as type of assessment, informant, and the boundedness and distality of the measures used (Sandbank et al., 2021). Expanding to older age groups would allow for researchers to include achievement outcomes as a dimension of outcomes for field-based benchmarks. Future studies can also synthesize additional types of metrics, such as indices of overlap.

Given our findings around publication status, future research should examine study quality of the published and unpublished literature to provide additional evidence of the threats of the publication process biasing single-case research syntheses (Chow & Ekholm, 2018). Although our review is primarily catered toward research audiences, our findings suggest variability in effect sizes across several factors, suggesting that practitioners should be aware that there are methodological factors that can substantively influence the amount of behavior change that researchers document in their studies. In addition, practitioners can expect that the effects of the interventions they implement may vary as a function of the type of behavior they are trying to change.

Limitations

There are several effect size metrics available for SCDs; we did not calculate some effect size metrics that are commonly used (Vannest & Sallese, 2021). We selected metrics that have demonstrated desirable properties for synthesizing a group of studies with highly variable procedural characteristics (e.g., session length, measurement system; Pustejovsky, 2019). This manuscript reports results for preschool-age children rather than school-age (K–12) populations; because engagement and challenging behavior in preschool and K–12 contexts may be considerably different, more work is needed in this area for older participants. We also report effect sizes for interventions conducted in school settings and do not include data on interventions outside of these contexts (e.g., clinic, home, community), which limits generalizability. In addition, it was difficult to synthesize information about participants in nine sources that assessed group-level data, as they reported different data (e.g., percentage of students of a given race but without specifying the number of students). We are unable to determine whether differences in operationalization of variables were related to effect size distribution and did not assess whether effect sizes were associated with optimal levels of behavior change (e.g., socially valid changes in behavior).

Though we double coded all sources and resolved discrepancies as a group during weekly meetings, we were unable to calculate IOA for our coding given the complexity of the data entry structure. We also acknowledge that we did not factor in study quality. This is a limitation given that there may be important study quality factors that may be associated with effect size magnitude (e.g., treatment fidelity). Additional research could create distributions as a function of different study quality variables. In this study, we do impose a level of study quality for inclusion where studies had to include at least three potential demonstrations of effect and include at least three data points per condition, common standards for SCD. Related to this issue is the possible conflation between study quality and publication status. We included dissertations in this review to ensure that we included a representative sample of research, but it is possible that dissertations, on average, presented lower study quality than published studies. Our findings should be interpreted accordingly.

Given that the purpose of this article was to synthesize the observed magnitude of effects of SCDs, we did not use meta-analysis or apply inferential statistics. This includes not applying statistical procedures to account for nesting within the data set (i.e., effect size dependency) nor understanding variation and precision based on standard errors. Analogous to subgroup analysis in meta-analysis, we generated distributions as a function of effect size and study characteristics instead of moderator analyses (e.g., metaregression) that are often conducted in meta-analysis. Future research could apply meta-analytic synthesis techniques to answer research questions using moderators, which would also allow for models to statistically control for other study- and effect-size-level factors and account for effect size dependency.

Conclusions

Despite limitations, this review provides critical information to single-case researchers and research synthesists that may allow them to evaluate single-study and aggregate data in comparison to values empirically derived from existing research. Our data reveal meaningful variability across effect size metrics and outcomes, and in particular, SMD effect sizes reached substantially larger values than LRR. Additional research is needed to examine relations between different effect sizes and to establish average effect sizes for school-age children.

References

Bakker, A., Cai, J., English, L., Kaiser, G., Mesa, V., & Van Dooren, W. (2019). Beyond small, medium, or large: Points of consideration when interpreting effect sizes. *Educational Studies in Mathematics*, *102*(1), 1–8. https://doi.org/10.1007/s10649-019-09908-4

Barton, E. E., Pustejovsky, J. E., Maggin, D. M., & Reichow, B. (2017). Technology-aided instruction and intervention for students with ASD: A meta-analysis using novel methods of estimating effect sizes for single-case research. *Remedial and Special Education*, *38*(6), 371–386. https://doi.org/10.1177/0741932517729508

Bloom, H. S., Hill, C. J., Black, A. R., & Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness*, *1*(4), 289–328. https://doi.org/10.1080/19345740802400072

Candbank, M., Chow, J., Bottema-Beutel, K., & Woynaroski, T. (2021). Evaluating evidence-based practice in light of the boundedness and proximity of outcomes: Capturing the scope of change. *Autism Research*, *14*(8), 1536–1542. https://doi.org/10.1002/aur.2527

Cheung, A. C., & Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educational Researcher*, *45*(5), 283–292. https://doi.org/10.3102/0013189X16656615

Chow, J. C. (2020). Commentary: Classroom motivation and learning disabilities. Consensus points and recommendations. *Learning Disabilities: A Multidisciplinary Journal*, *25*(2), 54–60. https://js.sagamorepub.com/ldmj/article/view/10734

Chow, J. C., & Ekholm, E. (2018). Do published studies yield larger effect sizes than unpublished studies in education and special education? A meta-review. *Educational Psychology Review*, *30*(3), 727–744. https://doi.org/10.1007/s10648-018-9437-7

Chow, J. C., Morse, A., Zhao, H., Kingsbery, C., Fisk, R., & Soni, I. (2022). A systematic review of the characteristics of students with emotional disturbance in special education research. *Remedial and Special Education*. https://doi.org/10.1177/07419325221125890

Chow, J. C., Sjogren, A. L., & Zhao, H. (2021). Reporting and reproducibility of meta-analyses in speech-language hearing research. *Journal of Speech, Language, and Hearing Research*, *64*(7), 2786–2793. https://doi.org/10.1044/2021_JSLHR-21-00047

Chow, J., Zhao, H., Sandbank, M., Bottema-Beutel, K., & Woynaroski, T. (2021). Empirically-derived effect size distributions of interventions for young children on the autism spectrum. *Journal of Clinical Child & Adolescent Psychology, 52*(2), 1–13. https://doi.org/10.1080/15374416.2021.2007485

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

Cumming, M. M., Bettini, E., & Chow, J. C. (2022). High-quality systematic literature reviews in special education: Promoting coherence, contextualization, generativity, and transparency. *Exceptional Children*. https://doi.org/10.1177/00144029221146576

Ekholm, E., & Chow, J. C. (2018). Addressing publication bias in educational psychology. *Translational Issues in Psychological Science*, *4*(4), 425–439. https://doi.org/10.1037/tps0000181

Gage, N. A., Cook, B. G., & Reichow, B. (2017). Publication bias in special education meta-analyses. *Exceptional Children*, *83*(4), 428–445. https://doi.org/10.1177/0014402917691016

Harris, D. N. (2009). Toward policy-relevant benchmarks for interpreting effect sizes: Combining effects with costs. *Educational Evaluation and Policy Analysis*, *31*(1), 3–29. https://doi.org/10.3102/0162373708327524

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of biomedical informatics*, *42*(2), 377–381.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*(3), 172–177. https://doi.org/10.1111/j.1750-8606.2008.00061.x

Huwaldt, J. A., & Steinhorst, S. (2015). *Plot Digitizer* (Version 2.6. 8) [Computer software]. https://sourceforge.net/projects/plotdigitizer

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods*, *17*(2), 137–152. https://doi.org/10.1037/a0028086

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, *49*(4), 241–253. https://doi.org/10.3102/0013189X20912798

Ledford, J. R., & Gast, D. L. (2018). *Single case research methodology: Applications in special education and behavioral sciences*. Routledge.

Ledford, J. R., Lambert, J. M., Pustejovsky, J. E., Zimmerman, K. N., Hollins, N., & Barton, E. E. (2022). Single-case-design research in special education: Next-generation guidelines and considerations. *Exceptional Children*. Advance online publication. https://doi.org/10.1177/00144029221137656

Ledford, J. R., & Windsor, S. A. (2021). Systematic review of interventions designed to teach imitation to young children with disabilities. *Topics in Early Childhood Special Education, 42*(2), 202–214. https://doi.org/10.1177/02711214211007190

Lipsey, M. W., & Cullen, F. T. (2007). The effectiveness of correctional rehabilitation: A review of systematic. *The Annual Review of Law and Social Science*, *3*, 297–320.

Luo, L., Reichow, B., Snyder, P., Harrington, J., & Polignano, J. (2020). Systematic review and meta-analysis of classroom-wide social–emotional interventions for preschool children. *Topics in Early Childhood Special Education*. https://doi.org/10.1177/0271121420935579

Maggin, D. M., Cook, B. G., & Cook, L. (2019). Making sense of single-case design effect sizes. *Learning Disabilities Research & Practice*, *34*(3), 124–132. https://doi.org/10.1111/ldrp.12204

Pigott, T. D., & Polanin, J. R. (2020). Methodological guidance paper: High-quality meta-analysis in a systematic review. *Review of Educational Research*, *90*(1), 24–46. https://doi.org/10.3102/0034654319877153

Pustejovsky, J. E. (2018). Using response ratios for meta-analyzing single-case designs with behavioral outcomes. *Journal of School Psychology*, *68*, 99–112. https://doi.org/10.1016/j.jsp.2018.02.003

Pustejovsky, J. E. (2019). Procedural sensitivities of effect sizes for single-case designs with directly observed behavioral outcome measures. *Psychological Methods*, *24*(2), 217–235. https://doi.org/10.1037/met0000179

Pustejovsky, J. E., & Swan, D. M. (2018). *SingleCaseES: A calculator for single-case effect sizes. R Package (Version 0.4, 3) [Computer software].*

R Core Team. (2018). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. http://www.R-project.org

Robertson, R. E., Sobeck, E. E., Wynkoop, K., & Schwartz, R. (2017). Participant diversity in special education research: Parent-implemented behavior interventions for children with autism. *Remedial and Special Education*, *38*(5), 259–271.

Shadish, W. R., Hedges, L. V., & Pustejovsky, J. E. (2014). Analysis and meta-analysis of single-case designs with a standardized mean difference statistic: A primer and applications. *Journal of School Psychology*, *52*(2), 123–147. https://doi.org/10.1016/j.jsp.2013.11.005

Tanner-Smith, E. E., Durlak, J. A., & Marx, R. A. (2018). Empirically based mean effect size distributions for universal prevention programs targeting school-aged youth: A review of meta-analyses. *Prevention Science*, *19*(8), 1091–1101. https://doi.org/10.1007/s11121-018-0942-1

Taylor, J. A., Kowalski, S. M., Polanin, J. R., Askinas, K., Stuhlsatz, M. A., Wilson, C. D., & Wilson, S. J. (2018). Investigating science education effect sizes: Implications for power analyses and programmatic decisions. *AERA Open*, *4*(3). https://doi.org/10.1177/2332858418791991

Vannest, K. J., & Sallese, M. R. (2021). Benchmarking effect sizes in single-case experimental designs. *Evidence-Based Communication Assessment and Intervention*, 1–24. https://doi.org/10.1080/17489539.2021.1886412

What Works Clearinghouse. (2020). *What Works Clearinghouse standards handbook, Version 4.1*. U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

Authors' Note

We would like to thank Kate Chazin and Kelsey Dunn for their support in the initial stages of this project. This research was supported, in part, by the Office of Special Education Programs (H325H190004; H325D190005).

ORCID iDs

Jason C. Chow https://orcid.org/0000-0002-2878-7410

Jennifer R. Ledford https://orcid.org/0000-0002-2392-7103

Open-Science Badges

For publishing their materials and data, Chow et al. received badges for open materials and open data. The public content may be retrieved from https://osf.io/6q7g2/?view_only=bc8cb4bff10a494d98c57bcd158889e3.

Supplemental Material

The supplemental material is available with the online version of the article.

Manuscript received February 2022; accepted February 2023.