

Bridging SCED and Group Designs: Demonstrating the Ratio-Metric BC-IRR Effect Size in a Meta-Analysis

Wen Luo

wluo@tamu.edu

Texas A&M University <https://orcid.org/0000-0002-0706-204X>

Chendong Li

James Madison University

Eunkyeng Baek

Texas A&M University

Method Article

Keywords: between-case incidence rate ratio, between-case standardized mean difference, design comparable, meta-analysis, single-case experimental design

Posted Date: August 18th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10704580/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Bridging SCED and Group Designs: Demonstrating the Ratio-Metric BC-IRR Effect Size in a Meta-Analysis

Abstract

This study introduces the between-case incidence rate ratio (BC-IRR), a ratio-metric effect size for single-case experimental designs (SCEDs), and demonstrates its use in a reanalysis of Barton et al.'s (2017) meta-analysis of technology-aided instruction and intervention (TAII) for students with autism spectrum disorder (ASD). We applied generalized estimating equation (GEE) models to estimate BC-IRR for SCED outcomes and computed mean response ratios for group designs. Across 36 eligible outcomes (29 from five group-design studies and 8 from eight SCED studies), separate random-effects models showed that group-design studies yielded a pooled mean response ratio of 1.23 (95% CI [1.00, 1.52]), indicating approximately 23% higher positive outcomes under TAI relative to comparison conditions, whereas SCEDs produced a pooled BC-IRR of 4.52 (95% CI [1.71, 11.92]), indicating more than a fourfold increase over baseline. We discuss BC-IRR's advantages over BC-SMD, namely, its invariance to outcome variance, insensitivity to observation session length, and direct percent-change interpretation, alongside practical limitations, including the requirement of nonzero baselines. These findings illustrate how BC-IRR can support design-comparable, interpretable synthesis of SCED and group-design evidence while underscoring the importance of triangulating ratio-metric and SMD-type effect sizes when evaluating intervention impact.

Keywords: between-case incidence rate ratio, between-case standardized mean difference, design comparable, meta-analysis, single-case experimental design

Introduction

Single-case experimental designs (SCED) offer a rigorous method for evaluating intervention effects at the individual level. Unlike typical group-based designs (i.e., between-subject designs) that assign participants into different treatment conditions, SCEDs systematically manipulate experimental conditions and measure outcomes repeatedly over time for one or more participants (Kratochwill et al., 2026). The proliferation of SCED studies across diverse fields, such as behavioral, educational, psychological, and clinical research, has created a substantial empirical literature. However, synthesizing findings from SCED studies with those from traditional group-design studies remains methodologically challenging.

A primary methodological obstacle that has hindered the incorporation of empirical findings from SCED studies into the research synthesis is the comparability of effect sizes from SCEDs with those from group-based designs. The treatment effects found in SCEDs represent a within-subject effect, that is, the effect of a treatment on an individual. On the other hand, the treatment effects found in group designs represent a between-subject effect, that is, the effect of a treatment on the average outcome of a group of individuals. These two types of effects are not directly comparable, because the variation in the within-subject effects does not reflect the variation across individuals whereas the variation of the between-subject effect is attributable to both within-subject and between-subject variations.

A second challenge is the measurement heterogeneity. Group designs typically employ score-based measures, such as standardized achievement tests or Likert-scale ratings. SCEDs, by contrast, typically utilize direct observation of specific behaviors, which are often defined based on individual cases. For example, in their meta-analysis of the effectiveness of technology-aided instruction and intervention for students with Autism Spectrum Disorder (ASD), Barton and

colleagues (2017) documented that group-design studies tended to use standardized measures, such as the Benten test of facial recognition (Benton et al., 1983). On the other hand, SCED studies included measures of target behaviors such as requesting, task engagement, and communication behavior in general. The vastly different natures of the outcomes complicate direct comparison between the two design traditions.

Strengths and Limitations of Between-case Standardized Mean Difference

The between-case standardized mean difference (BC-SMD) is a viable solution to the aforementioned challenges with several notable strengths (Hedges et al., 2013; Pustejovsky et al., 2014). BC-SMD can be used in replicated AB designs, multiple-baseline across participants, settings, or behaviors, clustered multiple baseline designs, and multivariate multiple baseline across participants (Chen et al., 2023). By using variation from both within and between participants, BC-SMD is on the same scale as effect sizes (e.g. Cohen's d) from group-design studies. It can also explicitly model key features of SCED data, including the nested data structure, trend, and autocorrelation, which improves the accuracy of effect size estimation. In simulation work, BC-SMD has shown robust performance under mild to moderate violations of normality with more than four cases and a series length of at least 16 (Chen et al., 2024).

On the other hand, BC-SMD has several limitations. Like Cohen's d , BC-SMD expresses the treatment effect in the unit of a standard deviation, which makes effects seemingly comparable regardless of the original metric. However, this standardization introduces its own interpretative and methodological constraints.

First, standardization does not create comparability when the underlying variables are of different nature. For example, consider two studies from Barton et al. (2017): one measures an

effect using scores on the Benten test of facial recognition, while the other measures the frequency of requesting behavior in a period of time. Simply converting these effects from their original units into standard deviation units does not make them more comparable. A one-SD change in facial recognition scores is qualitatively different from a one-SD change in the frequency of requesting behavior. Averaging an effect size of 2 SD from a cognitive test with an effect size of 3 SD from a behavior frequency count is as nonsensical as averaging 2 SD of length with 3 SD of weight (Alfarone et al., 2026). Converting both to standard deviation units obscures rather than resolves the fundamental incommensurability of these constructs.

Second, like Cohen's d , BC-SMD confounds intervention effect with the outcome variance, especially when the variances are not comparable across studies, which is often a norm than an exception (Greenland et al., 1986; Kim & Ferree Jr, 1981; King, 1986, Baguley, 2009). Buchner and Mayr (2004) illustrated this problem compellingly in the auditory negative priming literature: unstandardized reaction times revealed larger negative priming effects for older than younger adults (86 ms versus 53 ms), yet because older adults exhibited more variable responses, the standardized effect (Cohen's d) was smaller for older participants than younger ones ($d = 0.58$ versus $d = 0.83$). Had researchers relied exclusively on SMD, they would have concluded erroneously that negative priming was weaker in older adults, when the opposite was true.

Third, BC-SMD exhibits sensitivity to the observation session length (Pustejovsky, 2019). If one study records behavior frequency over a 10-minute session and another a 60-minute session, the resulting Cohen's d may differ even if the underlying treatment effect is identical (Wilson, 2021). This dependency on the observation session length violates the principle that effect size should be invariant to arbitrary measurement choices.

Fourth, BC-SMD is not readily understood by practitioners in schools and clinics. According to May's (2004) guidelines for meaningful statistical analysis, effect sizes should be: (a) understandable, devoid of jargon and assuming little about statistical knowledge, (b) interpretable, with a familiar or easily explained metric, and (c) directly usable, with little further manipulation. In our opinion, BC-SMD fails to meet these expectations. Telling a layperson that a treatment improves on-task behavior by "2 standard deviations" provides little practical insight. Furthermore, practitioners often set goals in terms of percentage improvement, such as increasing on-task behavior by 25% or decreasing disruptive behavior by 50%. A Cohen's d cannot be directly applied to these goals. While researchers have explored converting Cohen's d into non-overlap indices (which are more aligned with single-case research) to improve interpretability (Parker & Hagan-Burke, 2007), this conversion is valid only when data are approximately normally distributed. Single-case experimental design (SCED) data, particularly behavioral frequency counts, frequently violate these normality assumptions.

Finally, although BC-SMD is robust under mild violations of the normality, Chen et al. (2024) also showed that its relative bias increases as non-normality intensifies. Thus, it is unclear whether BC-SMD performs adequately under the kinds of pronounced non-normality that are common in behavioral outcomes. Furthermore, Sun & Cheung (2020) showed that the adverse impact of nonnormal data in primary studies persists in meta-analysis, even when the average sample size and the number of studies are large. While the specific impact on SCED meta-analyses remains to be fully evaluated, the risks associated with violating the normality assumption underlying BC-SMD should not be overlooked when conducting meta-analyses.

Between-Case Incidence Rate Ratio as an Alternative Effect Size

In response to these limitations, recent methodological work has introduced the between-case incidence rate ratio (BC-IRR) as an alternative effect size metric. Developed within a generalized linear mixed model (GLMM) framework for count data (Luo et al., 2025), BC-IRR has shown reliable performance in simulations of multiple-baseline across participants designs. Extensions based on generalized estimating equations (GEE) further simplify estimation while retaining satisfactory statistical performance (Li et al., 2025).

BC-IRR shares several key advantages with BC-SMD. Most notably, BC-IRR is design-comparable, allowing for direct comparison with the incidence rate ratios obtained from group designs (Luo et al., 2025). Additionally, it allows for adjustment of baseline trends, a critical feature when baseline data show systematic temporal patterns. Beyond these shared capabilities, BC-IRR offers several distinct strengths. First, it is not confounded with outcome variance, which makes it robust to heterogeneous outcome variability across studies. Second, it is insensitive to the observation session length, eliminating bias introduced by arbitrary measurement decisions. Third, BC-IRR offers intuitive interpretability. Defined as the ratio of incidence rate in the intervention phase to that in the baseline phase, it directly quantifies the treatment effect as a percentage increase or decrease. Crucially, BC-IRR does not require data to be normally distributed and can accommodate overdispersion (i.e., the excess variability) common in behavioral count data, without compromising statistical validity.

Purpose

Although BC-IRR is a promising alternative metric to BC-SMD, it remains relatively new to the field. For applied researchers to fully understand and implement this method in their own work, it is important to demonstrate its utility through real-world meta-analytic applications.

Using a reanalysis of the meta-analysis conducted by Barton et al. (2017), the present study demonstrates the computation of BC-IRR in primary SCED studies for both count and proportion outcomes, and the application of BC-IRR in meta-analyses. This demonstration includes a comparison with ratio-based treatment effects from group-design studies to highlight the cross-design comparability of the approach. To support the application, we used a newly developed R package that automates the computation process.

The remainder of this paper is organized as follows. We first introduce the GEE models used to estimate BC-IRR, followed by a description of the data used for the demonstration. We then outline the procedures for computing effect sizes in the primary studies: BC-IRR for SCED and ratio-based effect size for group-design studies, as well as the model used for the meta-analysis. Next, we present and interpret meta-analytic results. Finally, we discuss the findings in relation to the original Barton et al. meta-analysis and emphasize the additional insights provided by this new metric.

BC-IRR based on the Generalized Estimating Equation Models

General Introduction of GEE

The GEE is a general method for analyzing correlated continuous and noncontinuous outcomes (Liang & Zeger, 1986) are typically interpreted as population-average (marginal) models (Zorn, 2001). The regression coefficients represent average effects over the population rather than subject-specific effects. A GEE model is defined by three components: a link function, a linear combination of the predictors, and a working correlation matrix. The link function is chosen according to the outcome type. For count data, the log link is typically used. The linear combination of the predictors allows researchers to model various treatment effects (e.g., immediate treatment effect, trend effect) and adjust for baseline trends in SCEDs.

The working correlation matrix captures within-case dependence across repeated measures. Common options include first-order autoregressive specification where correlation declines with temporal lag; exchangeable, where all within-case correlations are equal; and unstructured, where each within-case correlation is freely estimated. Although the choice of working correlation does not affect the large-sample consistency of the regression estimates, misspecification can bias the model-based standard errors. To guard against this, it is recommended to use robust (sandwich) standard errors, as developed by White (1980) and extended to GEEs by Liang and Zeger (1986).

BC-IRR Based on Three Specific GEE Models

Based on for an across-participants multiple-baseline design, the first and most basic GEE model for count outcome is shown in Equation 1. For count data, it is typical to assume a Poisson or negative binomial distribution. This basic specification (Model 1) includes an immediate treatment effect but no baseline trend or treatment effect on the trend.

Model 1 (immediate treatment effect only)

Let Y_{ij} denotes the count outcome for case i and measurement j ; λ_{ij} be its expected value. $Phase_{ij}$ is a binary variable coded 0 for the baseline phase and 1 for treatment phase. The model assumes

$$Y_{ij} \sim \text{Poisson}(\lambda_{ij}) \tag{1}$$

$$\log(\lambda_{ij}) = \gamma_{00} + \gamma_{10}Phase_{ij}$$

where γ_{00} represents the baseline level and γ_{10} the immediate treatment effect on the log-count scale. Exponentiating γ_{10} (i.e., $e^{\gamma_{10}}$) yields BC-IRR, interpreted as the proportional change in the expected count between phases.

The basic model can be extended to adjust for a baseline trend. Model 2 (Equation 2) assumes an immediate treatment effect while allowing for a linear baseline trend.

Model 2 (adjusting for baseline trend):

$$\log(\lambda_{ij}) = \gamma_{00} + \gamma_{10}Phase_{ij} + \gamma_{20}Time_{ij} \quad (2)$$

In Model 2, $Time_{ij}$ is a time variable indexing measurement occasions or sessions. Center $Time_{ij}$ at a chosen value sets the reference time point and changes the interpretation of the intercept. The coefficient γ_{20} represents the baseline trend. Because this model still includes only an immediate treatment effect, BC-IRR is again derived as the exponentiation of the same coefficient γ_{10} (i.e., $e^{\gamma_{10}}$).

Model 2 can be further extended to allow the treatment to change the outcome trend over time. In Model 3, we estimate both an immediate treatment effect and a treatment effect on trend, providing a full specification for level and slope changes between baseline and treatment phases.

Model 3 (immediate and trend treatment effects):

$$\log(\lambda_{ij}) = \gamma_{00} + \gamma_{10}Phase_{ij} + \gamma_{20}Time_{ij} + \gamma_{30}Time'_{ij}Phase_{ij}. \quad (3)$$

Here, $Time'_{ij}Phase_{ij}$ captures the difference in trend between the treatment and baseline phase. It is common to center $Time'_{ij}$ such that zero corresponds to the start of the treatment phase, so that γ_{10} represents the immediate effect at the first treatment session compared with the counterfactual outcome that would have been observed had the participant remained in the control condition.

Based on Equation 3, the BC-IRR for the full model is given by $e^{(\gamma_{10} + \gamma_{30}(B-A))}$, where γ_{10} and γ_{30} represent the marginal estimates of the immediate level change and the change in trend between phases. To define the trend effect, A denotes the time at which treatment is first

introduced and B denotes the time at which the outcome is evaluated, so $B - A$ is the time elapsed under treatment used to compute the effect size. When $B - A = 0$, the effect is evaluated immediately after treatment onset, corresponding to the first treatment session. By choosing larger values of $B - A$, researchers can target cumulative effects. For example, setting $B - A = 3$ estimates the treatment effect after four consecutive treatment sessions. In multiple baseline designs where participants receive different numbers of treatment sessions, it is recommended to set $B - A$ to the smallest total number of treatment sessions observed across participants to avoid extrapolating beyond the observed treatment period (Luo et al., 2025).

Standard Error Estimation for BC-IRR

In Models 1 and 2, BC-IRR is defined by a single regression coefficient for the immediate treatment effect. In these cases, the standard error of BC-IRR can be obtained directly from the model-based robust standard error of that coefficient, combined with the usual transformation for an exponentiated parameter (e.g., using the variance of $\hat{\gamma}_{10}$ to derive the variance of $\exp(\hat{\gamma}_{10})$).

Model 3, by contrast, defines BC-IRR as a function of both the immediate effect and the treatment-by-time interaction (i.e., a combination of γ_{10} and γ_{30}). In this case, we use the delta method, which applies a first-order Taylor expansion to the BC-IRR function and uses the full variance–covariance matrix of the GEE estimates to approximate the variance of BC-IRR (Oehlert, 1992; Wasserman, 2004).

Although robust (sandwich) standard errors provide asymptotically valid inference, they can perform poorly with small sample sizes (McNeish & Stapleton, 2016; McNeish & Haring, 2017; Thompson et al., 2021). To improve both parameter estimates and their standard errors in

GEE with small sample size, several small-sample bias and variance corrections have been proposed (e.g., Thompson et al., 2021; Fay & Graubard, 2001; Kauermann & Carroll, 2001; Mancl & DeRouen, 2001; MBN, Morel et al., 2003). Recent simulation studies showed that the KC (Kauermann & Carroll, 2001) and MD method (Mancl & DeRouen, 200) produced relatively unbiased standard error estimates (Li et al, 2025).

Overdispersion

Overdispersion is referred to when the observed variance exceeds what is theoretically expected under a standard distribution (e.g., Poisson or binomial). It occurs typically in longitudinal or clustered data, causing standard models to produce incorrect estimates. Pustejovsky et al. (2019)'s review indicated that overdispersion is not uncommon for SCED count data, as repeated measurements for each case are usually correlated. Previous studies have demonstrated that the failure to account for overdispersion can lead to poor model fit and underestimated standard errors of regression coefficients (Hilbe, 2011, 2014).

GEE can handle overdispersion by relaxing the requirement for a fully specified likelihood and instead utilizing a quasi-likelihood framework. In a quasi-likelihood function, it only requires specifying the first two moments of the data (i.e., mean and variance) rather than the full distribution of the data. This allows the introduction of a scale or dispersion parameter that accounts for excessive variance for overdispersion (Liang & Zeger, 1986). In addition, the use of the robust standard errors (sandwich estimator) ensures that the marginal coefficient estimates are consistent, and inference remains valid even if the variance function is misspecified due to overdispersion (Hardin & Hilbe, 2012). Thus, GEEs accommodate overdispersion through variance correction, ensuring valid statistical inference, despite potential misspecification.

Extension to Proportion Outcomes

Models 1 to 3 are originally specified for count outcomes with a Poisson distribution. In many SCED studies, however, outcomes that are inherently counts (e.g., number of correct responses out of trials) are reported as proportions or percentages. Although such proportion outcome are often modeled with a logit link, yielding odds ratios effects, odds ratios are difficult to understand and often misinterpreted as risk ratios. Accordingly, methodologists recommend regression models that directly estimate risk ratios, such as log-binomial or Poisson regression with a log link and robust standard errors (e.g., Huang 2022, Knol et al., 2012).

To align proportion outcomes with the BC-IRR framework, we adopt Poisson-based GEE models for proportion outcomes in SCED, so that the exponentiated coefficients and hence the corresponding BC-IRR can be interpreted as risk ratios. Recent simulation studies show that when using robust standard error with KC small sample correction, BC-IRR can be accurately estimated for proportion outcomes, and interpreted as a risk ratio. When such studies are included in a meta-analysis using BC-IRR, the proportions are first converted back to counts to make them compatible with the Poisson-based models. This is done by multiplying the proportion by the corresponding number of trials or intervals per session and rounding to the nearest integer to recover the underlying to recover the underlying count data.

Data

The empirical data for the present demonstration were drawn from Barton et al. (2017), a systematic review and meta-analysis of technology-aided instruction and intervention (TAII) for students with autism spectrum disorder (ASD). Barton et al. (2017) defined TAII broadly as interventions in which technology served as the primary method for delivering instruction, including computer-assisted instruction, speech-generating devices, mobile devices, and virtual reality-based programs. The original review was designed both to evaluate the effectiveness of

TAII for improving communication, academic, engagement, social, emotion-recognition, and adaptive outcomes for students with ASD, and to demonstrate how BC-SMD for single-case experimental designs could be used alongside conventional group-design effect sizes. Eligible studies evaluated TAII relative to a control condition or non-TAII instruction, included a majority of participants with ASD, and used either a group design or a single-case design. Group designs included randomized or quasi-experimental comparisons across participants, whereas single-case studies were required to provide at least three opportunities to demonstrate behavior change at three different points in time.

Across the full review, Barton et al. identified 12 group-design studies and 10 single-case studies for their meta-analysis. The original analysis used BC-SMD for the single-case studies and standardized mean differences for the group-design studies. The present study reanalyzed a subset of primary studies to illustrate how the BC-IRR can be used as a ratio-metric effect size for synthesizing SCED findings and comparing them with group-design effects expressed on a compatible ratio scale. The current analysis included only outcomes that met four conditions. First, the outcome had to be nonnegative and ratio-interpretable (e.g., a count, rate, proportion, percentage correct, or correct-response score). Second, the published report or available data had to provide sufficient information to estimate a ratio-scale effect size and its sampling variance. For SCED studies, this requires session-level outcome data; and for group-design studies, group means and standard deviations. Third, the direction of the outcomes had to be consistent. In this demonstration, we used positively valenced outcomes in which larger values represented improvement because these were dominant in the included studies. Finally, baseline observations in SCEDs could not be all zeros because BC-IRR is undefined in that situation. Studies or outcomes that did not meet these criteria were excluded from the analysis.

Seven group-design studies were excluded from the main analysis. DeThorne et al. (2014), Moore and Calvert (2000), Silver and Oakes (2001), and Whalen et al. (2010) were excluded because the published reports did not provide sufficient directly extractable group-level posttest means and standard deviations, or equivalent information, to compute a log ratio and its sampling variance. Faja et al. (2007) was excluded because the available outcome was a difference score that could take negative values, making the log ratio undefined or substantively inappropriate. Mineo et al. (2009) was excluded because the study focused on a media-format engagement comparison rather than a treatment-versus-control TAI effect, and the available information was not sufficient for the present ratio-metric synthesis. Smith et al. (2014) was excluded from the main analysis because the outcomes were bounded role-play and self-confidence rating-scale scores with a nonzero scale minimum; these outcomes do not provide a clean count, rate, or proportion ratio and were therefore not retained in the main analysis. Two SCED studies were excluded. Taylor et al. (2004) and Lorah et al., (2014) were excluded because the extracted baseline observations were all zeros. Table 1 shows the characteristics of the included studies in the present meta-analysis.

Table 1
Descriptive Statistics of Included Outcomes by Condition

Study by design		Outcome (Type)		Control			Treatment		
				Mean	SD	N	Mean	SD	N
<i>Group-based Design</i>									
Beaumont & Sofronoff (2008)	Body Posture Recognition (Continuous)	21.32	2.82	23	21.81	2.97	26		
Beaumont & Sofronoff (2008)	Dylan is being Teased (Continuous)	2.64	1.56	23	5.08	2.23	26		
Beaumont & Sofronoff (2008)	Facial Expression Recognition (Continuous)	19.73	2.80	23	19.92	2.67	26		
Beaumont & Sofronoff (2008)	James and the Maths Test (Continuous)	2.00	1.11	23	3.81	1.58	26		
Golan & Baron-Cohen (2006)	CAM Face Task (Continuous)	34.80	8.20	22	37.50	7.80	19		
Golan & Baron-Cohen (2006)	CAM Face Task (Continuous)	29.30	9.50	13	36.20	8.90	13		
Golan & Baron-Cohen (2006)	CAM No. Concepts Recognized (Continuous)	11.30	5.40	22	13.60	4.80	19		
Golan & Baron-Cohen (2006)	CAM No. Concepts Recognized (Continuous)	8.50	6.30	13	13.50	5.20	13		
Golan & Baron-Cohen (2006)	CAM Voice Task (Continuous)	36.60	7.90	22	38.90	6.20	19		
Golan & Baron-Cohen (2006)	CAM Voice Task (Continuous)	31.80	10.90	13	38.90	7.60	13		

Study by design	Outcome (Type)	Control			Treatment		
		Mean	SD	N	Mean	SD	N
Golan & Baron-Cohen (2006)	Reading the Mind in Films (Continuous)	12.80	3.40	22	11.80	3.80	19
Golan & Baron-Cohen (2006)	Reading the Mind in Films (Continuous)	10.50	3.20	13	11.90	3.70	13
Golan & Baron-Cohen (2006)	Reading the Mind in the Eyes (Continuous)	23.00	7.30	22	23.80	4.70	19
Golan & Baron-Cohen (2006)	Reading the Mind in the Eyes (Continuous)	19.20	6.80	13	23.80	4.20	13
Golan & Baron-Cohen (2006)	Reading the Mind in the Voice (Continuous)	17.40	3.50	22	16.70	3.90	19
Golan & Baron-Cohen (2006)	Reading the Mind in the Voice (Continuous)	14.70	4.60	13	16.20	3.50	13
Golan et al. (2010)	Emotional Vocabulary (Continuous)	9.11	3.45	18	12.50	3.09	20
Golan et al. (2010)	Situation-Expression Matching, Level 1 (Continuous)	8.94	2.34	18	13.00	2.45	20
Golan et al. (2010)	Situation-Expression Matching, Level 2 (Continuous)	9.22	2.69	18	13.45	2.35	20
Golan et al. (2010)	Situation-Expression Matching, Level 3 (Continuous)	9.61	2.83	18	13.30	2.27	20

Study by design	Outcome (Type)	Control			Treatment		
		Mean	SD	N	Mean	SD	N
Hopkins et al. (2011)	Emotion Recognition, Total (Continuous)	5.18	2.44	14	6.57	2.28	11
Hopkins et al. (2011)	Emotion Recognition, Total (Continuous)	6.08	2.33	11	9.54	2.34	13
Hopkins et al. (2011)	Facial Recognition, Benton Long Form (Continuous)	29.36	4.24	14	32.64	6.81	11
Hopkins et al. (2011)	Facial Recognition, Benton Long Form (Continuous)	31.23	5.79	11	40.64	6.67	13
Hopkins et al. (2011)	Facial Recognition, Benton Short Form (Continuous)	12.64	4.20	14	14.64	4.96	11
Hopkins et al. (2011)	Facial Recognition, Benton Short Form (Continuous)	14.54	3.26	11	19.45	4.27	13
Kasari et al. (2014)	Total Different Word Roots (Count)	25.62	18.21	30	33.11	17.51	31
Kasari et al. (2014)	Total Number of Comments (Count)	8.10	9.14	30	14.09	9.12	31

Study by design	Outcome (Type)	Control			Treatment		
		Mean	SD	N	Mean	SD	N
Kasari et al. (2014)	Total Spontaneous Communicative Utterances (Count)	40.34	21.43	30	61.90	25.85	31
<i>SCED</i>							
Boesch et al. (2013)	Social-communicative (Count)	0.59	1.12	29	10.87	5.82	75
Choi et al. (2010)	Requesting Sequences (Percentage)	1.22	1.72	59	2.95	1.30	113
Cihak et al. (2010)	Self-monitoring (Percentage)	8.71	3.76	76	25.37	7.18	71
Flores et al. (2012)	Communication (Count)	8.16	4.03	45	21.14	12.55	28
King et al. (2014)	Requesting Skills (Count)	0.08	0.28	13	1.86	3.18	125
McKissick et al. (2013)	Map-reading (Count)	0.42	1.02	24	5.85	3.50	20
Pennington et al. (2012)	Story Writing (Count)	0.21	0.54	19	1.79	1.23	38
Richter et al. (2011)	Correct Responses (Count)	7.92	2.14	13	9.68	2.96	34

Note. *N* denotes the number of participants in group designs and the total number of person-time observations in SCED.

Analysis

A two-stage random-effects meta-analysis was conducted. In the first stage, log-ratio effect sizes and sampling variances were estimated separately for each eligible outcome within each primary study. In the second stage, the estimated log ratio-metric effect size effects were synthesized using random-effects meta-analysis models. Following the original meta-analysis, separate meta-analyses were conducted for SCEDs and group-designs. Publication biases were assessed using Egger's regression (Egger et al., 1997). Below we describe the methods in detail.

SCED Studies

For the SCED studies, log BC-IRR were estimated based on a phase-only model (i.e., Model 1 in Equation 1). This specification was selected for two reasons. First, visual inspection of the included SCED series did not suggest baseline trends. Second, this specification was chosen to remain consistent with the original Barton et al. (2017) synthesis. The analysis was conducted using the R package *bcirr* (Li et al., 2026).

Group-design Studies

For the group-design studies, the response ratio (also known as the ratio of means; Hedges et al., 1999) was calculated as a ratio between treatment group (or using the posttest group means) and control group. For each outcome, the log response ratio was calculated as

$$\log(RR) = \log(\bar{Y}_T) - \log(\bar{Y}_C), \quad (4)$$

where $\bar{Y}_T$ is the mean for the treatment group and $\bar{Y}_C$ is the mean for the control or comparison group.

The sampling variance of the log response ratio was estimated using the delta method:

$$Var[\log(RR)] = \frac{SD_T^2}{n_T \bar{Y}_T^2} + \frac{SD_C^2}{n_C \bar{Y}_C^2}, \quad (5)$$

where SD_T and SD_C are the treatment and control group standard deviations, and n_T and n_C are the corresponding group sample sizes. When model-estimated means or confidence intervals

were reported instead of raw means and standard deviations, the log ratio and its sampling variance were derived from the available model-based estimates. The resulting group-design effects were placed on the same log-ratio scale as BC-IRRs in SCEDs.

Meta-analysis

The meta-analytic model was specified as

$$\hat{\theta}_{ek} = \mu + u_k + \varepsilon_{ek}, \quad (6)$$

where $\hat{\theta}_{ek}$ is the observed log-ratio effect size for effect e in study k , μ is the average log-ratio effect, u_k is the study-level random effect, and ε_{ek} is the sampling error with known sampling variance. The study-level random effects were assumed to follow $u_k \sim N(0, \tau^2)$. Models were estimated using restricted maximum likelihood (REML). To account for the non-independence of multiple effect sizes extracted from the same study and the small number of study clusters, cluster-robust standard errors with CR3 small-sample correction were used (Tipton & Pustejovsky, 2015).

Separate meta-analytic models were fitted for group-design effects and SCED effects. For each model, the pooled log-ratio estimate was exponentiated to obtain the pooled ratio-scale effect size. The exponentiated estimate was interpreted as the multiplicative change in the expected outcome under treatment relative to the comparison condition. Confidence intervals were also exponentiated to present results on the ratio scale.

To evaluate whether effects differed between group-design and SCED studies, we fitted a design-difference meta-regression. Because preliminary results indicated that heterogeneity differed substantially across design types, the design-difference model allowed the degree of between-study heterogeneity to vary by design. The model was specified as

$$\hat{\theta}_{ek} = \beta_0 + \beta_1 D_{ek} + (1 - D_{ek})u_{Gk} + D_{ek}u_{Sk} + \varepsilon_{ek}, \quad (7)$$

where $D_{ek} = 0$ for group-design effects and $D_{ek} = 1$ for SCED effects. In this model, β_0 represents the average log-ratio effect for group-design studies, and β_1 represents the log-ratio difference between SCED and group-design effects. Therefore, $\exp(\beta_1)$ represents the ratio of the pooled SCED BC-IRR to the pooled group-design IRR.

The random effects were allowed to have design-specific variances. Specifically,

$$u_{Gk} \sim N(0, \tau_G^2), u_{Sk} \sim N(0, \tau_S^2), \text{Cov}(u_{Gk}, u_{Sk}) = 0,$$

so that separate heterogeneity variances were estimated for group-design and SCED effects using a diagonal random-effects structure. Models were estimated using REML. To account for the non-independence of multiple effect sizes extracted from the same study and the small number of study clusters, cluster-robust standard errors with CR3 small-sample correction were used.

In the R implementation, these models were estimated using *rma.mv()* from the *metafor* package (Viechtbauer, 2010), and CR3 robust tests and confidence intervals were obtained using the *clubSandwich* package (Pustejovsky, 2026).

Egger's Regression

Following Barton et al. (2017), we evaluated possible outcome reporting bias using Egger's regression. Because the available Egger analysis was based on group-design effects, it should be interpreted cautiously when applied to single-case research. Let $\hat{\theta}_{ek}$ denote the observed log-ratio effect size for effect e in study k , and let SE_{ek} denote its sampling standard error. The standardized effect size was calculated as $Z_{ek} = \frac{\hat{\theta}_{ek}}{SE_{ek}}$, and precision was calculated as

$P_{ek} = \frac{1}{SE_{ek}^2}$. Egger's regression was then specified as

$$Z_{ek} = \alpha_0 + \alpha_1 P_{ek} + r_{ek}. \quad (8)$$

In this model, Z_{ek} is the standardized log-ratio effect size, P_{ek} is the precision of the effect-size estimate, and r_{ek} is the residual error. The intercept α_0 is the Egger asymmetry

parameter. A statistically significant intercept provides evidence that the standardized effect sizes are asymmetrically distributed with respect to their precision, which may indicate small-study effects or outcome reporting bias. The slope α_1 represents the association between standardized effect size and precision. In the present analysis, Equation (8) was estimated using ordinary least squares for the group-design effects. The data and R code for all the analyses is available at https://osf.io/t246z/overview?view_only=669c56b8764f4086b1c7adaaf7161689 .

Results

Descriptive Statistics

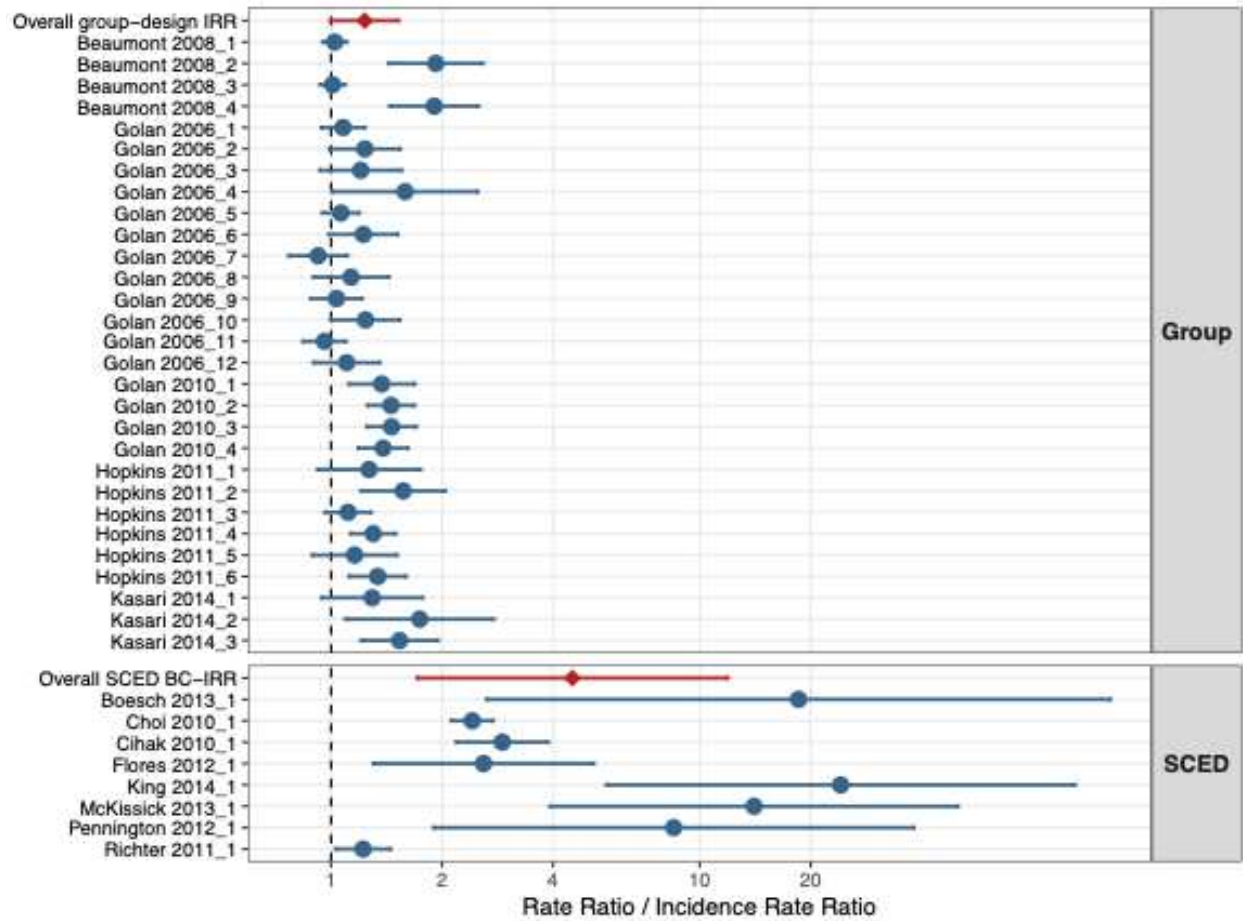
The final analytic dataset included 36 ratio-metric effect sizes from 12 studies: 29 effects from five group-design studies and 8 effects from eight SCED studies. Table 1 shows the descriptive statistics of each outcome by conditions.

Meta-analyses

The individual study effect sizes and the pooled estimates for each design type are displayed in the forest plot shown in Figure 1.

Figure 1.

Forest plot for group-based design and SCEDs



For group-design studies, the pooled log-ratio estimate was $\hat{\mu} = 0.21$, $SE = 0.08$, $t(3.93) = 2.80$, $p = .050$. After exponentiation, the pooled group-design IRR was 1.23, 95% CI [1.00, 1.52]. This indicates that, on average, outcomes in the treatment groups were approximately 23% higher than outcomes in the comparison groups. The lower bound of the confidence interval was very close to 1.00, suggesting that the group-design estimate was statistically borderline. The heterogeneity test was statistically significant, $Q(28) = 112.91$, $p < .001$, indicating variability in effects beyond sampling error. The estimated between-study variance was $\hat{\tau}^2 = 0.020$ on the log-ratio scale.

For the SCED studies, the pooled log-ratio estimate was $\hat{\mu} = 1.51$, $SE = 0.40$, $t(5.24) = 3.74$, $p = .008$. After exponentiation, the pooled BC-IRR was 4.52, 95% CI [1.71, 11.92]. This

indicates that, across the included SCED studies, the expected treatment-phase outcome was approximately 4.52 times the expected baseline-phase outcome. The confidence interval was wider than the group-design interval, reflecting the greater between-study variability. The heterogeneity test was statistically significant, $Q(7) = 74.93, p < .001$, and the estimated between-study variance was $\hat{\tau}^2 = 0.877$ on the log-ratio scale.

The design-difference meta-regression further indicated that effects differed significantly between group-design and SCED studies. The estimated design difference was $\hat{\beta}_1 = 1.30$, $SE = 0.41$, $t(6.87) = 3.16$, $p = .016$. After exponentiation, the SCED pooled effect was approximately 3.66 times the pooled group-design effect, 95% CI [1.38, 9.68].

Classic Egger's regression was conducted separately for group-design and SCED effects. For the group-design effects, the Egger intercept was statistically significant, $\hat{\alpha}_0 = 3.03$, $SE = 0.68$, $t(27) = 4.44$, $p < .001$, 95% CI [1.63, 4.43], suggesting evidence of funnel plot asymmetry or small-study effects. The precision coefficient was not statistically significant, $\hat{\alpha}_1 = -0.09$, $SE = 0.06$, $t(27) = -1.53$, $p = .137$. For the SCED effects, the Egger intercept was not statistically significant, $\hat{\alpha}_0 = 2.50$, $SE = 1.47$, $t(6) = 1.70$, $p = .139$, 95% CI [-1.09, 6.08], indicating a symmetry funnel plot. The precision coefficient was positive but not statistically significant, $\hat{\alpha}_1 = 0.48$, $SE = 0.20$, $t(6) = 2.34$, $p = .058$. However, the SCED Egger result should be interpreted cautiously because it was based on only eight effects.

Discussions

This study introduced BC-IRR, a ratio-scale effect size metric for single-case experimental designs, and demonstrated its application through a reanalysis of technology-aided instruction and intervention (TAII) for students with autism spectrum disorder. Unlike standardized mean difference metrics, BC-IRR provides a more interpretable measure of

treatment effects expressed as ratios (or equivalently, as percent change) that is comparable across SCED and group-based designs. Through this reanalysis of the Barton et al. (2017) meta-analysis, we illustrated how BC-IRR can bridge methodological traditions and offer new insights into intervention effectiveness.

Summary of Findings

We found that TAI implemented in group designs lead to approximately 23% increase in positive outcomes compared to the control condition ($IRR = 1.23$, 95% CI [1.00, 1.52]). When implemented in multiple baseline across participants designs, TAI produced a substantially larger effect, with positive outcomes increasing by over fourfold after the intervention (BC- $IRR=4.52$, 95% CI [1.71, 11.92]). Because BC-IRR is design comparable, we were able to conduct a meta-regression combining these two design types and formally test the difference which yielded statistical significance between the two designs.

Comparing these findings with the original Barton et al. (2017) results, which reported a Cohen's d of 0.66 for group designs and a BC-SMD of 1.97 for SCEDs, the interpretability advantage of BC-IRR becomes immediately apparent. Stating that treatment outcomes are 23% higher or more than 4 times baseline levels conveys meaningful information to practitioners, policymakers, and researchers without requiring specialized knowledge of standardized effect size metrics.

Understanding the Design Differences: Beyond Publication Bias

The substantial difference between SCED and group-design effect estimates has traditionally led researchers to synthesize results from the two designs separately. However, it is worth examining more carefully why such differences arise. Barton et al. (2017) noted that "this trend is consistent with theoretical expectations that effect sizes from single-case research may

generally be larger than those from group designs, even when assessed on comparable scales". They attributed this pattern primarily to publication bias, noting that "visual analysis of single-case research can reliably detect large effects better than small or moderate effects. Therefore, within the single-case research community, there is strong publication preference for studies with larger effects. This inadvertently limits the pool of studies for which to draw effect sizes and might artificially inflate average effects".

While publication bias represents a legitimate and important explanation, our analysis reveals that it does not tell the full story. The differential effect sizes across design types reflect more than selective reporting. They also reflect fundamental differences in how BC-SMD and BC-IRR quantify treatment impact. Next, we discuss two distinct statistical mechanisms through which design differences arise: the role of variance and the role of baseline level.

The Role of Variance in BC-SMD vs. BC-IRR

For BC-SMD, the apparent discrepancy between group designs and SCEDs may partially stem from differences in outcome variance across study types. As shown in the descriptive statistics of the current study (Table 1), outcome variance in group designs tends to be substantially larger than in SCEDs. Because standardized mean differences divide the mean difference by the pooled standard deviation, greater variability reduces the estimated effect size even when the underlying mean difference remains constant. This phenomenon suggests that the relatively smaller BC-SMD in group designs may partly reflect an artifact of the larger standard deviations observed in between-subjects comparisons rather than genuinely weaker treatment effects. This represents an important limitation of BC-SMD that can lead to misleading inferences when comparing effect magnitudes across design types.

By contrast, BC-IRR is not confounded with outcome variance because it does not standardize by standard deviations. The expected value of BC-IRR depends on the ratio of means between conditions, making it robust to heterogeneous outcome variability across studies. This fundamental property ensures that variance differences across study types do not artificially inflate or deflate the estimated treatment effect.

Baseline Levels and the Larger SCED Effect

Given that BC-IRR removes the confounding influence of variance, what explains the substantially larger effect observed in SCED studies compared to group designs? Our analysis points to a critical difference in baseline or control-condition means. Examination of the descriptive statistics reveals that the baseline mean in SCED studies was relatively low compared to control group means in group-design studies (as shown in Table 1). Because BC-IRR is computed as a ratio, a smaller denominator (baseline mean) naturally produces a larger ratio, even when the absolute treatment effect might be comparable.

This pattern suggests that the larger BC-IRR in SCEDs may reflect genuine differences in the starting levels of participants or in the nature of the control conditions employed across design traditions. In SCED research, baseline phases typically involve minimal or no intervention, establishing a lower functional level from which treatment effects are measured. In group designs, by contrast, comparison conditions often involve some form of standard care or alternative intervention, establishing a higher baseline against which treatment effects are evaluated.

Furthermore, SCED studies tended to focus on direct observation of specific behaviors, such as requesting, task engagement, and communication, which may have exhibited particularly low baseline rates in the absence of targeted intervention. Group-design studies, meanwhile,

more frequently employed standardized assessment measures (e.g., the Benton test of facial recognition) that may capture higher baseline functioning. These measurement differences compound the baseline-level differences, contributing to the divergent effect estimates.

Implications for Cross-Design Synthesis

The differential baseline levels revealed by BC-IRR have important implications for how we interpret and synthesize evidence across research designs. Rather than viewing the larger SCED effects solely as an artifact of publication bias, BC-IRR suggests that group-design and SCED studies may be evaluating interventions under meaningfully different conditions. SCEDs appear to test whether interventions can substantially improve functioning from a minimal baseline, while group designs more often assess whether interventions outperform existing alternatives or standard practices.

This insight underscores the value of ratio-scale effect sizes in cross-design meta-analysis. Where BC-SMD suggested a puzzling inconsistency between design types, BC-IRR provides a more interpretable and mechanistically grounded explanation: SCED studies show larger multiplicative effects because they start from lower baseline levels, not necessarily because treatments produce larger absolute improvement, or because outcome variance varies across designs, or because of publication bias.

Limitations and Future Directions

Our analysis shares some limitations with the original Barton et al. (2017) review, including the relatively small number of eligible studies and the heterogeneity of outcome domains. Additionally, because BC-IRR requires non-negative outcomes with non-zero baseline values, we excluded some studies that were included in the original BC-SMD analysis. Future

research should explore how BC-IRR performs across a broader range of outcome types and investigate methods for handling zero-inflated baseline data.

Furthermore, we applied Egger's regression to assess publication bias in both group-design and SCED studies. However, this method was developed specifically for group-based designs, and its performance for SCED has not been systematically examined. Egger's regression assumes that publication bias operates through statistical significance testing. Specifically, smaller studies with null or modest effects are less likely to be published. In SCED research, by contrast, publication decisions are more likely driven by visual determinations of experimental control and functional relationships rather than p -values. Consequently, conventional publication bias tools developed for group designs are not ideally suited for application to single-case research, and alternative methods tailored to SCED are not yet available. Future methodological studies are needed to evaluate the validity and performance of Egger's regression for SCED studies. Until such validation work is completed, results from conventional bias detection tools should be interpreted with caution when applied to single-case research.

Conclusion

This demonstration highlights BC-IRR as an alternative effect size metric to BC-SMD that provides design comparable and interpretable quantifications of treatment effect in SCEDs. These findings suggest a broader principle: no single effect size metric is likely to provide a complete characterization of intervention effectiveness. While BC-IRR does not have the inherent limitations of BC-SMD, such as confounding with outcome variance and sensitivity to observation session length, it alone does not provide a complete picture either. BC-IRR shows relative change. When absolute change is of interest, researchers should still use mean

difference-type effect sizes. However, caution is warranted when interpreting SMD-type effect sizes when outcome variances are not comparable across studies.

Reference

- Alfarone, I., & Gondan, M. (2026). Change of d-irection: Current limitations and future directions in psychological meta-analysis. *Frontiers in Psychology, 17*, Article 1717798. <https://doi.org/10.3389/fpsyg.2026.1717798>
- Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology, 100*(3), 603–617. <https://doi.org/10.1348/000712608X377117>
- 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>
- *Beaumont, R., & Sofronoff, K. (2008). A multi-component social skills intervention for children with Asperger syndrome: the Junior Detective Training Program. *Journal of child psychology and psychiatry, and allied disciplines, 49*(7), 743–753. <https://doi.org/10.1111/j.1469-7610.2008.01920.x>
- Benton, A. L., Sivan, A. B., Hamsher, K. D. S., Varney, N. R., & Spreen, O. (1983). Facial recognition: Stimulus and multiple choice pictures. In A. L. Benton, A. B. Sivan, K. D. S. Hamsher, N. R. Varney, & O. Spreen (Eds.), *Contributions to neuropsychological assessment* (pp. 30–40). Oxford University Press.
- *Boesch, M. C., Wendt, O., Subramanian, A., & Hsu, N. (2013). Comparative efficacy of the Picture Exchange Communication System (PECS) versus a speech-generating device: Effects on social-communicative skills and speech development. *Augmentative and Alternative Communication, 29*, 197–209.
- Buchner, A., & Mayr, S. (2004). Auditory negative priming in younger and older adults. *The Quarterly Journal of Experimental Psychology Section A, 57*(5), 769–787. <https://doi.org/10.1080/02724980343000431>
- Chen, L. T., Chen, Y. K., Yang, T. R., Chiang, Y. S., Hsieh, C. Y., Cheng, C., Ding, Q. W., Wu, P. J., & Peng, C. J. (2024). Examining the normality assumption of a design-comparable effect size in single-case designs. *Behavior Research Methods, 56*(1), 379–405. <https://doi.org/10.3758/s13428-022-02035-8>
- Chen, M., Pustejovsky, J. E., Klingbeil, D. A., & Van Norman, E. R. (2023). Between-case standardized mean differences: Flexible methods for single-case designs. *Journal of School Psychology, 98*, 16–38. <https://doi.org/10.1016/j.jsp.2023.02.002>
- *Choi, H., O'Reilly, M., Sigafoos, J., & Lancioni, G. (2010). Teaching requesting and rejecting sequences to four children with developmental disabilities using augmentative and alternative communication. *Research in Developmental Disabilities, 31*, 560–567.

*Cihak, D. F., Wright, R., & Ayres, K. M. (2010). Use of self-modeling static-picture prompts via a handheld computer to facilitate self-monitoring in the general education classroom. *Education and Training in Developmental Disabilities, 45*, 136–149.

DeThorne, L., Betancourt, M. A., Karahalios, K., Halle, J., & Bogue, E. (2015). Visualizing syllables: Real-time computerized feedback within a speech-language intervention. *Journal of Autism and Developmental Disorders, 45*, 3756–3763.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ, 315*(7109), 629–634. <https://doi.org/10.1136/bmj.315.7109.629>

Faja, S., Aylward, E., Bernier, R., & Dawson, G. (2007). Becoming a face expert: A computerized face-training program for high-functioning individuals with autism spectrum disorders. *Developmental Neuropsychology, 33*, 1–24.

Fay, M. P., & Graubard, B. I. (2001). Small-sample adjustments for Wald-type tests using sandwich estimators. *Biometrics, 57*, 1198–1206. <https://doi.org/10.1111/j.0006-341X.2001.01198.x>

*Flores, M., Musgrove, K., Renner, S., Hinton, V., Strozier, S., Franklin, S., & Hill, D. (2012). A comparison of communication using the Apple iPad and a picture-based system. *Augmentative and Alternative Communication, 28*, 74–84.

*Golan, O., Ashwin, E., Granader, Y., McClintock, S., Day, K., Leggett, V., & Baron-Cohen, S. (2010). Enhancing emotion recognition in children with autism spectrum conditions: An intervention using animated vehicles with real emotional faces. *Journal of Autism and Developmental Disorders, 40*, 269–279.

*Golan, O., & Baron-Cohen, S. (2006). Systemizing empathy: Teaching adults with Asperger syndrome or high-functioning autism to recognize complex emotions using interactive multimedia. *Development and Psychopathology, 18*, 591–617.

Greenland, S., Schlesselman, J. J., & Criqui, M. H. (1986). The fallacy of employing standardized regression coefficients and correlations as measures of effect. *American Journal of Epidemiology, 123*(2), 203–208. <https://doi.org/10.1093/oxfordjournals.aje.a114229>

Hardin, J. W., & Hilbe, J. M. (2012). *Generalized estimating equations* (2nd ed.). Chapman & Hall/CRC.

Hedges, L. V., Gurevitch, J., & Curtis, P. S. (1999). The meta-analysis of response ratios in experimental ecology. *Ecology, 80*, 1150–1156. [https://doi.org/10.1890/0012-9658\(1999\)080\[1150:TMAORR\]2.0.CO;2](https://doi.org/10.1890/0012-9658(1999)080[1150:TMAORR]2.0.CO;2)

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*(4), 324–341. <https://doi.org/10.1002/jrsm.1086>

Hilbe, J. M. (2011). *Negative binomial regression* (2nd ed.). Cambridge University Press.

Hilbe, J. M. (2014). *Modeling count data*. Cambridge University Press.

*Hopkins, I. M., Gower, M. W., Perez, T. A., Smith, D. S., Amthor, F. R., Wimsatt, F. C., & Biasini, F. J. (2011). Avatar assistant: Improving social skills in students with an ASD through a computer-based intervention. *Journal of Autism and Developmental Disorders*, 41, 1543–1555.

Huang, F. L. (2022). Alternatives to logistic regression models in experimental studies. *The Journal of Experimental Education*, 90(1), 213–228.
<https://doi.org/10.1080/00220973.2019.1699769>

*Kasari, C., Kaiser, A., Goods, K., Nietfeld, J., Mathy, P., Landa, R., Murphy, S., & Almirall, D. (2014). Communication interventions for minimally verbal children with autism: A sequential multiple assignment randomized trial. *Journal of the American Academy of Child & Adolescent Psychiatry*, 53, 635–646.

Kauermann, G., & Carroll, R. J. (2001). A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association*, 96(456), 1387–1396.
<https://doi.org/10.1198/016214501753382309>

Kim, J.-O., & Ferree, G. D., Jr. (1981). Standardization in causal analysis. *Sociological Methods & Research*, 10(2), 187–210. <https://doi.org/10.1177/004912418101000203>

King, G. (1986). How not to lie with statistics: Avoiding common mistakes in quantitative political science. *American Journal of Political Science*, 30(3), 666–687.
<https://doi.org/10.2307/2111095>

*King, M. L., Takeguchi, K., Barry, S. E., Rehfeldt, R. A., Boyer, V. E., & Mathews, T. L. (2014). Evaluation of the iPad in the acquisition of requesting skills for children with Autism Spectrum Disorder. *Research in Autism Spectrum Disorders*, 8, 1107–1120.

Knol, M. J., Le Cessie, S., Algra, A., Vandenbroucke, J. P., & Groenwold, R. H. (2012). Overestimation of risk ratios by odds ratios in trials and cohort studies: Alternatives to logistic regression. *Canadian Medical Association Journal*, 184(8), 895–899.
<https://doi.org/10.1503/cmaj.101715>

Kratochwill, T. R., Ferron, J., & Levin, J. R. (2026). Reporting guidelines for single-case experimental design research: Recommendations for data analysis. *Journal of School Psychology*, 115, Article 101528. <https://doi.org/10.1016/j.jsp.2025.101528>

Li, C., Luo, W., Li, H., & Baek, E. (2026). *bcirr* (Version 0.1.0) [Computer software]. GitHub. <https://github.com/cliattx/bcirr>

Li, H., Li, C., Luo, W., & Baek, E. (2025, October 17). *Between-case incidence rate ratio for single case experimental designs with count outcomes: A Monte Carlo simulation with conditional and marginal models* [Preprint]. https://doi.org/10.31234/osf.io/rx96g_v1

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1), 13–22. <https://doi.org/10.1093/biomet/73.1.13>

Lorah, E. R., Parnell, A., & Speight, D. R. (2014). Acquisition of sentence frame discrimination using the iPad™ as a speech generating device in young children with developmental disabilities. *Research in Autism Spectrum Disorders*, 8, 1734–1740.

Luo, W., Li, H., Baek, E., & Li, C. (2025, October 24). *Between-case incidence rate ratio: A design comparable effect size for count outcomes in single case experimental designs* [Preprint]. https://doi.org/10.31219/osf.io/9gdkn_v2

Mancl, L. A., & DeRouen, T. A. (2001). A covariance estimator for GEE with improved small-sample properties. *Biometrics*, 57(1), 126–134. <https://doi.org/10.1111/j.0006-341X.2001.00126.x>

May, H. (2004). Making statistics more meaningful for policy research and program evaluation. *American Journal of Evaluation*, 25(4), 525–540. <https://doi.org/10.1177/109821400402500408>

*McKissick, B. R., Spooner, F., Wood, C. L., & Diegelmann, K. M. (2013). Effects of computer-assisted explicit instruction on map-reading skills for students with autism. *Research in Autism Spectrum Disorders*, 7, 1653–1662.

McNeish, D., & Stapleton, L. M. (2016). Modeling clustered data with very few clusters. *Multivariate Behavioral Research*, 51(4), 495–518. <https://doi.org/10.1080/00273171.2016.1167008>

McNeish, D. M., & Haring, J. R. (2017). Clustered data with small sample sizes: Comparing the performance of model-based and design-based approaches. *Communications in Statistics—Simulation and Computation*, 46(2), 855–869. <https://doi.org/10.1080/03610918.2014.983648>

Mineo, B. A., Ziegler, W., Gill, S., & Salkin, D. (2009). Engagement with electronic screen media among students with Autism Spectrum Disorders. *Journal of Autism and Developmental Disorders*, 39, 172–187.

Moore, M., & Calvert, S. (2000). Brief report: Vocabulary acquisition for children with autism: Teacher or computer instruction. *Journal of Autism and Developmental Disorders*, 30, 359–362.

Morel, J. G., Bokossa, M. C., & Neerchal, N. K. (2003). Small sample correction for the variance of GEE estimators. *Biometrical Journal*, 45(4), 395–409. <https://doi.org/10.1002/bimj.200390021>

Oehlert, G. W. (1992). A note on the delta method. *The American Statistician*, 46(1), 27–29. <https://doi.org/10.1080/00031305.1992.10475842>

Parker, R. I., & Hagan-Burke, S. (2007). Useful effect size interpretations for single case research. *Behavior Therapy*, 38(1), 95–105. <https://doi.org/10.1016/j.beth.2006.05.002>

*Pennington, R. C., Stenhoff, D. M., Gibson, J., & Ballou, K. (2012). Using simultaneous prompting to teach computer-based story writing to a student with autism. *Education and Treatment of Children*, 35, 389–406.

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. (2026). *clubSandwich: Cluster-robust (sandwich) variance estimators with small-sample corrections* (R package version 0.7.0.9000) [Computer software]. <http://jepusto.github.io/clubSandwich/>

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(5), 368–393. <https://doi.org/10.3102/1076998614547577>

*Richter, S., & Test, D. (2011). Effects of multimedia social stories on knowledge of adult outcomes and opportunities among transition-aged youth with significant disabilities. *Education and Training in Autism and Developmental Disabilities*, 46, 410–424.

Silver, M., & Oakes, P. (2001). Evaluation of a new computer intervention to teach people with autism or Asperger syndrome to recognize and predict emotions in others. *Autism*, 5, 299–316.

Smith, M. J., Ginger, E. J., Wright, K., Wright, M. A., Taylor, J. L., Humm, L. B., Olsen, D. E., Bell, M. D., & Fleming, M. F. (2014). Virtual reality job interview training in adults with Autism Spectrum Disorder. *Journal of Autism and Developmental Disorders*, 44, 2450–2463.

Sun, R. W., & Cheung, S. F. (2020). The influence of nonnormality from primary studies on the standardized mean difference in meta-analysis. *Behavior Research Methods*, 52(4), 1552–1567. <https://doi.org/10.3758/s13428-019-01334-x>

Taylor, B. A., Hughes, C. E., Richard, E., Hoch, H., & Coello, A. R. (2004). Teaching teenagers with autism to seek assistance when lost. *Journal of Applied Behavior Analysis*, 37, 79–82.

Thompson, J. A., Hemming, K., Forbes, A., Fielding, K., & Hayes, R. (2021). Comparison of small-sample standard-error corrections for generalised estimating equations in stepped wedge cluster randomised trials with a binary outcome: A simulation study. *Statistical Methods in Medical Research*, 30(2), 425–439. <https://doi.org/10.1177/0962280220958735>

Tipton, E., & Pustejovsky, J. E. (2015). Small-sample adjustments for tests of moderators and model fit using robust variance estimation in meta-regression. *Journal of Educational and Behavioral Statistics*, 40(6), 604–634. <https://doi.org/10.3102/1076998615606099>

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3), 1–48. <https://doi.org/10.18637/jss.v036.i03>

Wasserman, L. (2004). *All of statistics: A concise course in statistical inference*. Springer.

Whalen, C., Moss, D., Ilan, A. B., Vaupel, M., Fielding, P., Macdonald, K., Cernich, S., & Symon, J. (2010). Efficacy of TeachTown: Basics computer-assisted intervention for the intensive comprehensive autism program in Los Angeles Unified School District. *Autism*, 14, 179–197.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4), 817–838. <https://doi.org/10.2307/1912934>

Wilson, D. B. (2022). The relative incident rate ratio effect size for count-based impact evaluations: When an odds ratio is not an odds ratio. *Journal of Quantitative Criminology*, 38(2), 323–341. <https://doi.org/10.1007/s10940-021-09494-w>

Zorn, C. (2001). Generalized estimating equation models for correlated data: A review with applications. *American Journal of Political Science*, 45(2), 470–490. <https://doi.org/10.2307/2669353>