Literature Review

Hammill Institute on Disabilities

Remedial and Special Education
1–13

DOI: 10.1177/07419325231190808
rase.sagepub.com

Sage

What Predicts Within-Participant Replication of Relative Efficiency in Single-Case Comparisons? A Logistic Regression Analysis

Lanqi Wang, PhD[1,2], Chengan Yuan, PhD, BCBA-D[3], Shahad Alsharif, PhD, BCBA-D[4,5], Qing Archer Zhang, PhD[3], and Yang Du, PhD[1,6]

Abstract

Single-case comparative studies could help identify efficient instructional procedures for individuals with disabilities. However, previous literature reported inconsistent efficiency results if multiple comparisons were conducted, indicating that within-participant replication was uncommon. In this review, we examined single-case comparative studies with multiple within-participant comparisons and identified the arrangements that may be related to within-participant replication. We performed a multilevel mixed-effects logistic regression analysis to assess the association between different arrangements and consistent efficiency results between the comparisons. We found that some arrangements, such as random assignment of targets and a combination of random assignment and equating procedure, appear more predictive of within-participant replication.

Keywords

single-case research design, comparative designs, within-participant replication, instruction

Individualized intervention approaches have long been recognized and incorporated in special education (Chow & Hampton, 2022). Due to heterogeneity among learners with disabilities, the effects of a single intervention often vary across different learners (e.g., Chow & Hampton, 2022; Kasari et al., 2021). Special educators need to identify interventions that best serve *each* student. To account for individual differences, researchers have investigated the use of single-case experimental designs to help with intervention selection (e.g., McComas & Burns, 2009). These designs allow researchers and special educators to examine an individual learner's responses under specific intervention conditions (e.g., VanDerHeyden & Burns, 2009). As such, when a learner is not responding to an existing intervention, other alternatives can be assessed using single-case experimental analysis for this learner before making a selection. A series of studies have investigated the utility of single-case experimental designs in identifying effective academic interventions for individual learners. For example, Jones et al. (2009) conducted analyses using the multitreatment design by implementing brief and extended intervention phases in sequence to identify the intervention and intervention package that produced the largest performance gains for their individual participants. The effects of the selected intervention were then verified for each participant. Thus, single-case experimental designs could be used to assess student responses under specific conditions and help select and develop effective individualized interventions.

In addition to identifying effective interventions for individual learners, single-case experimental designs have also been used to compare interventions to reveal those that promote faster skill acquisition (e.g., intervention efficiency; Ledford & Gast, 2018). As interventions that lead to quicker skill acquisition require fewer teaching sessions for a learner to master a skill, selecting more efficient interventions could allow more skills to be taught (Kodak & Halbur,

[1]The University of Iowa, Iowa City, USA
[2]Nanjing Normal University of Special Education, Nanjing, China
[3]Arizona State University, Tempe, USA
[4]Dar Al-Hekma University, Jeddah, Saudi Arabia
[5]Curtin University, Perth, Australia
[6]University of Science and Technology of China, Hefei, China

Corresponding Author:
Chengan Yuan, Division of Educational Leadership and Innovation, Mary Lou Fulton Teachers College, Arizona State University, P.O. Box 871811, Tempe, AZ 85287, USA.
Email: chengan.yuan@asu.edu

Associate Editor: Erin Barton

2021). When multiple effective interventions exist, it may be necessary for researchers and special educators to determine the most efficient intervention alternative for each learner.

To compare different interventions that target skill acquisition, single-case comparisons that include rapid iterative alternation of comparison conditions (e.g., ABABBAAB) have often been used (e.g., Holcombe et al., 1994), such as adapted alternating treatments design (Sindelar et al., 1985), parallel treatments design (Gast & Wolery, 1988), and repeated acquisition design (Kirby et al., 2021). Although some arrange multiple comparisons (e.g., parallel treatments design), each comparison typically includes rapid alternations so that the differences in effects among interventions could be revealed quickly within a comparison (Wolery et al., 2018). With much of the intervention programs for individuals with disabilities focusing on skill acquisition, comparative designs seem uniquely positioned to provide practical recommendations for special educators to select efficient procedures that can produce faster skill acquisition. With a face validity that is particularly appealing to researchers and consumers (Johnston, 1988), the number of single-case comparative studies has been increasing steadily in the last decade (e.g., Cariveau et al., 2021).

Although single-case comparative studies often aim to provide instructional recommendations for learners beyond their participants, the conclusions are likely still limited to those who share a similar combination of characteristics, target behaviors, settings, and resources as the participants in the studies (Johnston, 1988; Shabani & Lam, 2013). Thus, similar to the use of brief experimental analysis, some researchers have suggested conducting one comparison as an assessment to inform intervention selection for individual learners (e.g., Carroll et al., 2018; Kodak & Halbur, 2021; McGhan & Lerman, 2013; Yuan & Zhu, 2020). A special educator can first compare the effects of different interventions as an assessment and consider using the more efficient procedure revealed in this comparison for the same learner in the future. If this is the case, this assessment would have immediate practical implications as it has the potential to pinpoint the more efficient procedures for individual learners. This recommendation relies on the consistency in within-participant replication of the intervention effects—that is, the ability to achieve the same effects of an independent variable (e.g., intervention) on the dependent variable (e.g., behaviors) across comparisons (Ledford & Gast, 2018). Despite the potential utility of a single-case comparison for instructional recommendations, a recent review by Ledford et al. (2021) found that within-participant replications were, in fact, not common between comparisons. As limited reviews examined within-participant replication, our first purpose was to replicate the findings by Ledford et al. (2021) and examine within-participant replicability in the context of single-case comparisons. Results could inform the feasibility of using single-case comparisons to predict efficient interventions for individual learners.

Because within-participant replication was found uncommon in Ledford et al. (2021), we further examined the variables that could be associated with within-participant replication. Specifically, failure in within-participant replication is likely related to the generalizability of the findings across comparisons (i.e., external validity). As replication attempts introduce new contexts (e.g., new targets, settings, etc.) and populations beyond the specific arrangement of the initial experiment (Fabrigar et al., 2020; Kazdin, 2011), the characteristics of these context(s) and population(s) between the comparisons could moderate the intervention effects. In the case of single-case comparisons, even though the participant remains the same between replication attempts, exposure to interventions in the initial comparison could affect participant performance in the subsequent comparisons. For example, a participant may learn at a much faster rate regardless of the interventions after they have already experienced one procedure (e.g., learning to learn; Ledford et al., 2021; Ledford & Wolery, 2013). In this case, outcomes from different intervention conditions in a subsequent comparison could become undifferentiated even if a procedure had previously produced faster skill acquisition than the other alternatives. Similarly, because replications may introduce new target sets, skills, or other contexts, the difference in these experimental arrangements between the comparisons could also affect replicability.

In addition to external validity, the inconsistent findings between comparisons could also be related to threats to internal validity. For example, single-case comparisons addressing skill acquisition require researchers to assign different stimuli or behaviors to the comparison conditions because the behavioral outcomes are unlikely to reverse back to the baseline levels (Holcombe et al., 1994; Sindelar et al., 1985). If differentiated difficulty levels are inadvertently introduced during comparison, the findings could be influenced by these uncontrolled variables, affecting replication. Although strategies such as logical analysis and randomization have been suggested to provide control for differences in characteristics among target sets (Cariveau et al., 2022; Ledford et al., 2021), their relation with within-participant replicability is unclear. Similarly, when assessing efficiency, a priori mastery criteria are often included in the comparisons as the basis for data evaluation (e.g., Holcombe et al., 1994; Ledford et al., 2021). A higher performance level and longer observation requirement (e.g., 100% across five sessions vs. 80% across three sessions) could increase the probability that the performance temporarily decreases below the criterion level under one condition, resulting in decreased or reversed efficiency between comparisons. However, this requirement could also negate the influence of outliers when determining efficiency.

As it is unclear how these variables influence within-participant replication in single-case comparisons, examining their associations is warranted. Findings could help understand the conditions under which within-participant replications may or may not occur and refine experimental arrangements when conducting comparisons, potentially allowing the use of single-case comparison to select and develop efficient, individualized interventions. Thus, we addressed the below questions in this review:

Research Question 1 (RQ1): Have single-case comparative studies produced consistent intervention efficiency results within the same participant (i.e., within-participant replication)?
Research Question 2 (RQ2): What variables relating to internal and external validity are associated with within-participant replicability?

Method

Search

A search for peer-reviewed and non-peer-reviewed sources was conducted on June 8, 2022, using three search strategies: electronic, ancestral, and forward searches. First, the first two authors conducted an electronic search, using ProQuest Dissertations and Theses and three ProQuest databases—PsycINFO, ERIC, and PsycARTICLES—with the following search string: (autis* OR retard* OR disabilit* OR delay OR handicap) AND ("alternating treatment*" OR "multielement" OR "multi-element" OR "simultaneous treatment*" OR "parallel treatment*" OR "repeated acquisition"). The inclusion of dissertations and theses in the search was to reduce the possibility of publication bias (Shadish et al., 2016). Following this step, ancestral and forward searches were conducted using four single-case comparison systematic reviews (Cariveau et al., 2021, 2022; Ledford et al., 2021; Shabani & Lam, 2013). The search procedures produced a total of 1,636 articles and 14,026 dissertations.

Following the search, two screening steps were conducted. During the initial screening, the abstract, title, participants, and figure were screened, followed by the full-text screening using the below criteria. First, articles had to be written or transcribed in English, but the timeframe was not restricted. Second, studies had to include participants with disabilities. The age and the disability type were not restricted. Studies with parents, teachers, therapists, clinicians, or paraprofessionals as implementers must also assess the skill acquisition of the participants with disabilities. Third, the studies must have compared at least two different interventions or the same intervention with different parameters (e.g., one-to-one vs. small-group instruction) that targeted skill acquisition. We excluded studies without a comparison between two interventions (e.g., comparing an intervention to a baseline control condition). Fourth, we excluded studies that targeted reversible behavior or did not arrange distinctive sets of targets or behaviors for the comparison conditions. Fifth, studies had to use a single-case comparative design with rapid iterative alternation (e.g., adapted alternating treatments and parallel treatments design). Sixth, studies had to include at least one within-participant replication (i.e., at least two comparisons of the same interventions). We excluded those that had only one comparison. Seventh, studies had to include a graphical representation of the dependent measures for data extraction. Studies were excluded if data could not be extracted. Last, each study had to include a mastery criterion for the dependent variable to determine relative efficiency. We eliminated 14,728 studies during the initial screening, and the full-text screening yielded 104 articles that met all criteria for subsequent coding. See Figure 1 for our screening and search procedures and results using the Preferred Reporting Items for Systematic Reviews and Meta-Analysis flow diagram (Moher et al., 2009).

Coders for this review included the first four authors and a graduate student. They were required to achieve 100% reliability for the first 200 initial and full-text screenings before continuing. Each article was screened by two coders, with one being either of the first two authors. In cases of discrepancies, the first two authors discussed until they reached a consensus. The intercoder reliability for screening was 95.6%.

Coding

We coded predictors and outcome values. The predictors were categorized into (a) target selection and assignment methods, (b) mastery criterion dimensions, and (c) contextual change types (see Table 1 for the definitions). The first two categories were coded for each comparison, while the last was coded between each pair of comparisons. Outcome values were coded for the consistency of within-participant comparisons. The same five coders completed the coding. The first two authors trained the other three coders. Coders demonstrated a minimum of 95% intercoder reliability before starting coding. Studies were evenly and randomly assigned to each coder, with two coders per study. Similar to screening, one of the two coders was either the first or the second author, and discrepancies were discussed by the first two authors until consensus. Mean intercoder reliability was 98.54% for all comparisons.

Predictors

Target Selection and Assignment Methods. The target selection and assignment methods refer to the strategies controlling for differences among the targets: random assignment and equating procedure. *Random assignment* was coded if a

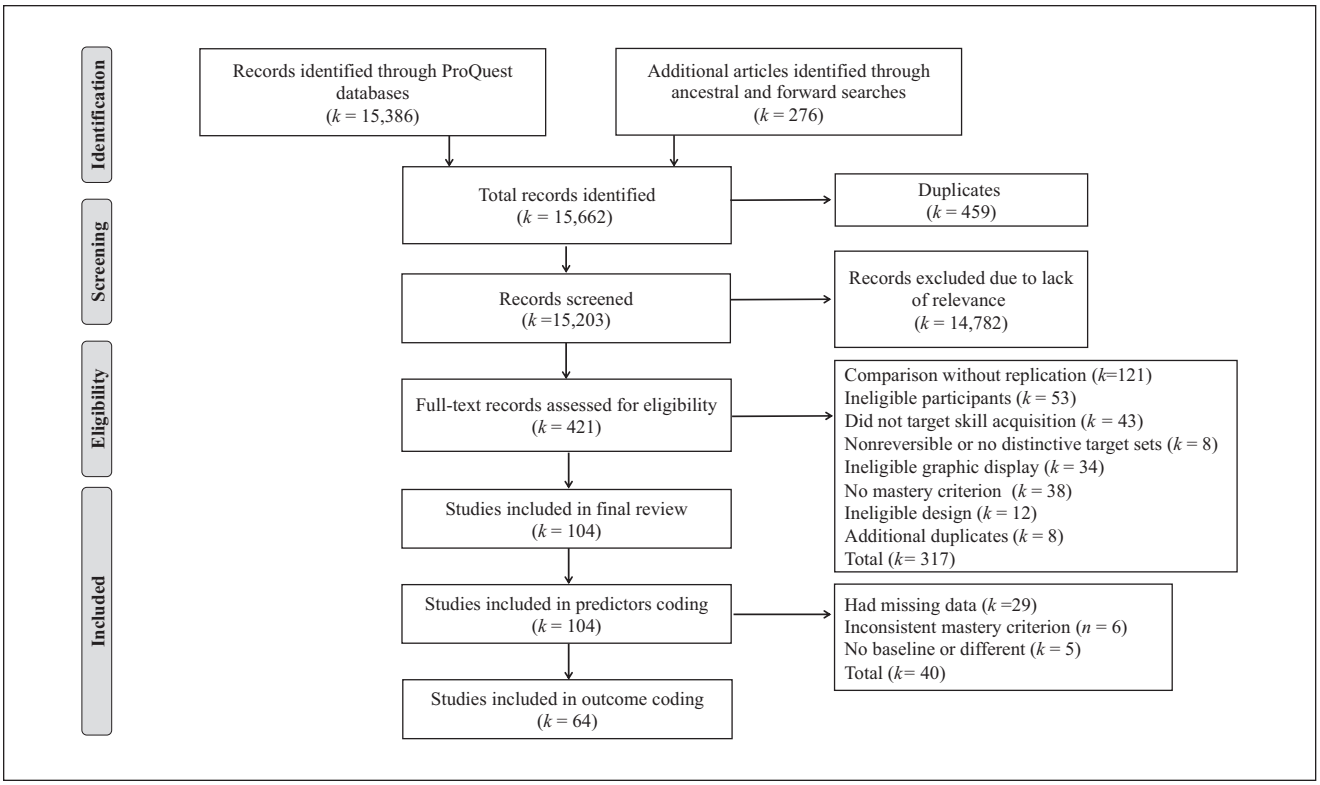

Figure 1. PRISMA Flow Diagram for Search and Screening Procedures.
PRISMA = Preferred Reporting Items for Systematic Reviews and Meta-Analysis.

Table 1. Predictor Definitions.

Predictor	Definition
Target selection and assignment methods	Strategies used to control for differences in target sets assigned to different conditions.
Random assignment	Randomly or quasi-randomly assigning targets to the conditions.
Equating procedure	Analyzing target characteristics (e.g., using logical analysis or expert rating) and assigning targets of equivalent difficulty to the conditions.
Mastery criterion dimension	Two dimensions in the criterion used to determine mastery in a comparison.
Performance level	The value that performance must meet (e.g., 100% correct).
Frequency	The number of performance-level observations (e.g., three consecutive sessions).
Contextual change type	Systematic changes in any experimental arrangements between comparisons
Target	Different target sets without an additional systematic difference (e.g., skill difference) between the comparisons (e.g., sets of visual stimuli when teaching labeling common objects, sets of completely different social behaviors between the comparisons).
Skill	Target sets between the comparisons selected from different skills (e.g., comparisons targeted motor imitation and following instructions).
Setting	Comparisons conducted in different settings (e.g., classroom and home).
Implementer	Procedures conducted by different individuals between comparisons (e.g., Teacher A vs. Teacher B).
Timing	Procedures conducted at different points of time between comparisons (e.g., procedures between comparisons conducted at different times of a day such as a morning and an afternoon, procedures between comparisons conducted at different times relative to the response such as 5 s vs. 10 s after the response).

comparison was reported to have randomly or quasi-randomly assigned targets across the conditions. *Equating procedure* was coded when the comparison was reported to have equated the targets (e.g., using logical analysis, expert rating, etc.) between the conditions. If a comparison did not report using either method, *neither* was coded. If a

comparison was reported to have used both strategies in combination, we coded it as *combined*. Coders recorded the target selection and assignment methods for each within-participant comparison.

Mastery Criterion Dimensions. For each comparison, we coded the mastery criterion using two dimensions, the performance-level requirement and the frequency of observations requirement at that performance level. Take the mastery criterion of 100% accuracy across three sessions as an example, the performance level was coded as 100% and the observation frequency was coded as 3. For the performance level, percentage was used for coding. If a percentage was not reported but the number of performance opportunities (e.g., 10 trials in each session) and response frequency requirement (e.g., 10 correct responses) were reported, we calculated the percentage for that comparison. For comparisons that only reported the response frequency requirement without a denominator (e.g., opportunities), we treated them as missing data (3.5%, $n = 29$). For the observation frequency, we selected the session-length requirement for coding, as the majority of the comparisons (95.69%, $n = 799$) reported the session-length requirement in their mastery criteria. For comparisons that did not report the number of sessions in their mastery criteria, we coded them as missing data (1.4%, $n = 12$) even if they had used other units (e.g., trials). Coders recorded performance level and observation frequency for each within-participant comparison.

Contextual Change Types. Contextual change types refer to how contexts were different between the two comparisons. For example, if the two comparisons used different sets of targets without additional systematic difference between the two comparisons, the contextual change was coded as *target* change between that pair of comparisons. Although all replications include different targets across the comparisons, additional systematic differences, such as different skills, settings, implementors, and timings, could exist. Thus, we predetermined some plausible contextual change types: *target, skill, setting, implementor*, and *timing* (see definitions and examples in Table 1). If any other types of contextual changes were observed between a pair of comparisons, the coders needed to describe the specific contextual change type used in the study. Coders recorded the contextual change type for each pair of comparisons.

In addition to predictor coding, we also assessed the baseline phase for all comparisons. We coded a comparison as *no baseline* when at least one condition lacked a baseline phase or a baseline included fewer than three data points per condition. We further compared baseline data between the conditions and coded *not different* or *different*. Baselines of two conditions were judged as *not different* if (a) neither baseline had an increasing trend, (b) the baseline levels (medians) were within a maximum of 20% difference, and (c) overlapping data points between the baselines were at least 33.3%. Otherwise, baselines were considered *different* between the conditions. After coding the predictors, we found 821 comparisons from 104 studies.

Outcome Values

Data Extraction. Within each comparison, we manually counted the data points to derive the total session number to mastery criterion for each condition. If the intervention was terminated before mastery, we coded it as early termination (ET).

Determining Efficiency. To determine relative efficiency, we first calculated the differential rate to mastery between the two conditions in each comparison by dividing the difference in sessions to criterion between the two conditions by the larger session number reported in either condition. For example, if a participant required eight sessions to reach mastery for one condition and 10 sessions for the other, the difference would be two sessions, and the differential rate would be two sessions divided by 10 sessions, multiplied by 100, yielding 20%.

We used a differential rate of 10%, proposed in Ledford et al. (2021), as the cutoff point when determining differentiated efficiency. As Ledford et al. pointed out, a 10% difference could produce a meaningful difference in practice over a long time. Using this cutoff point, if the differential rate between the two conditions was equal to or above 10% in one comparison, we considered the results of the two conditions to be differentiated. The condition with fewer sessions to mastery was deemed more efficient than the other condition for that comparison. If the differential rate was below 10%, we considered the effects of the two conditions undifferentiated. If both conditions were coded as ET where neither condition produced mastery, we considered that the efficiency could not be determined for this comparison. However, if one condition was coded as ET while the other produced mastery, the latter was considered a more efficient intervention.

Outcome Coding. Each outcome value was coded as either *replicated* or *not replicated.* We did not assign one outcome value for each participant across all comparisons and all interventions compared. When multiple comparisons are conducted (or multiple interventions are compared), the results between each pair of comparisons (or interventions) may differ. For example, Comparisons 1 and 2 may yield replicated results, while Comparisons 1 and 3 may not. In this case, assigning one value (*not replicated*) across all comparisons and interventions may not be sensitive in detecting replication. The chances of replication can decrease as the number of comparisons and interventions increases. Instead, we assigned each outcome value for two

comparisons of two interventions for each participant. If one participant had multiple comparisons, outcomes were coded for all pairs of comparisons. For example, if one participant experienced three comparisons of two interventions, we coded outcomes for each pair (i.e., Comparisons 1 and 2, 2 and 3, and 1 and 3). Likewise, if one participant had more than two interventions in the comparisons, outcome values were coded for all pairs of interventions in all pairs of comparisons. For example, if two comparisons included three interventions (e.g., Interventions A, B, and C), outcome values were coded separately for Interventions A and B, A and C, and B and C. This procedure may be more sensitive to detect replication and is more likely to produce sufficient data for analysis.

Outcomes were coded using the following criteria. An outcome was coded as *replicated* (coded as 0) if the same condition was consistently more efficient between two comparisons. Similarly, *replicated* was also coded if a pair of comparisons both showed undifferentiated results for their conditions. If two comparisons yielded different efficiency results, *not replicated* was coded (coded as 1) for the outcome. However, if both conditions in both comparisons were coded as ET, this outcome was removed as we could not determine intervention efficiency when no condition had resulted in mastery.

We used the following criteria to determine eligible outcomes for further logistic regression analysis. An outcome was excluded if (a) it had missing data, (b) the mastery criterion was not consistent between a pair of comparisons, and (c) at least one comparison in a pair was coded as *no baseline* or *different*. A total of 487 eligible outcomes were yielded from 64 studies.

Outcome Interobserver Agreement and Procedural Fidelity

We examined if the outcome interobserver agreement (IOA) and procedural fidelity (PF) in each comparison were scored for at least 20% of sessions and reached a threshold of at least 80% agreement for IOA and 80% accuracy for PF (Barton et al., 2018; Reichow et al., 2018). We coded the IOA and PF for each outcome. As each outcome value was derived from two comparisons, we used the lower value from the comparisons, when different, to determine if IOA and PF criteria were met for each outcome.

Multilevel Mixed-Effects Logistic Regression Analysis

We conducted a multilevel mixed-effects logistic regression analysis to identify the associations between predictors and the outcome values. Logistic regression describes the relation between one binary variable (i.e., outcome values of *replicated* or *not replicated*) and one or more predictors (i.e., target selection and assignment methods, mastery criterion dimensions, and contextual change types). We included both categorical and continuous predictors. Categorical predictors were the target selection and assignment methods and contextual change types, while continuous predictors were the two mastery criterion dimensions.

Consisting of both fixed and random effects, the multilevel mixed-effects model allows us to statistically incorporate nesting issues using homogeneity of variance (Snijders & Bosker, 2012). In this review, as each study could have multiple participants and each participant multiple outcomes, we used a three-level hierarchical nested data structure in multilevel mixed-effects logistic regression analysis. The three levels included outcome level as the Level 1 group, participant level as the Level 2 group, and study level as the Level 3 group. This three-level mixed-effects model included all predictive variables as fixed effects and incorporated random intercepts of Levels 2 and 3 groups as random effects. Akaike information criterion (AIC; Akaike, 1974) was calculated for the model fit. Marginal and conditional R^2 in the mixed-effects model measure the proportion of variance explained by the predictors (Nakagawa & Schielzeth, 2013). The value of marginal R^2 represents the contribution of the fixed effects (i.e., all predictors), while the value of conditional R^2 accounts for the variance explained by the entire model, including fixed and random effects. We used the intraclass correlation coefficient (ICC) to assess the degree of outcome homogeneity within participant and study levels (Wu et al., 2012). The value of ICC ranged from 0 to 1, with 0 representing outcomes unrelated to the group levels and 1 indicating all variance explained by group levels and not other factors (i.e., no variation within each level).

The three-level mixed-effects logistic regression analysis yielded a comparative statistic, the odds ratio (*OR*), that provided an effect size. An $OR > 1$ indicates an increase in the odds of within-participant replication due to a one-unit increase in the predictor, whereas an $OR < 1$ represents a decrease in the odds of replication due to a one-unit increase in the predictor (e.g., Rumberger, 1995). An *OR* of 1 indicates no change in the odds of replication in relation to the predictor. An *OR* was interpreted as a significant outcome when the 95% confidence interval (CI) of the *OR* did not include the value of 1.0 (Szumilas, 2010). In contrast, when the 95% CI included the value of 1.0, the outcome was considered not significant. Supplemental materials are available on the Open Science Framework at https://osf.io/9vuc2/?view_only=1aa8b334560f42e0b9fcc34bdd294358. Analyses were conducted in R with mixed models fit using the lme4 package (Version 3.3.1; Bates et al., 2015).

Results

A total of 64 studies were included in outcome coding and analysis, with 180 participants with disabilities and 487

Table 2. Descriptive Data for the Categorical Data.

	Total outcomes		Outcomes coded as *replicated*	
Predictors	Frequency	Percent total frequency[a]	Frequency	Percent at predictor-level[b]
Target selection and assignment methods				
Random assignment	86	17.7	48	55.83
Equating procedure	133	27.3	46	34.59
Combined	156	32.0	92	58.97
Neither	112	23.0	33	29.46
Total	487		219	
Contextual change types				
Target	437	89.70	204	46.68
Skill	32	6.6	13	40.63
Timing	18	3.7	2	11.11
Total	487		219	

[a]Percentage calculated with total frequency of all predictors in the respective category as the denominator.
[b]Percentage calculated at predictor level with total outcome frequency of that predictor as the denominator.

Table 3. Descriptive Data for Continuous Data.

	Total outcomes			
Predictors	Total valid number	*M*	Range	*SD*
Performance level	487	94.02	(75, 100)	8.34
Frequency of observation	487	2.64	(1, 5)	0.76
	Outcomes coded as *replicated*			
	Total valid number	*M*	Range	*SD*
Performance level	219	93.98	(75, 100)	8.35
Frequency of observation	219	2.64	(1, 5)	0.76

outcomes indicating whether replication occurred between two comparisons for each participant. Replication accounted for 44.97% ($n = 219$) of the outcomes, and the remaining 55.03% ($n = 268$) did not demonstrate replication. A total of 404 (82.96%) outcomes reported IOA measures at 80% or higher for at least 20% of the sessions, while 124 (17.04%) outcomes did not achieve the reliability criterion or report data. Similarly, 404 (82.96%) outcomes achieved the PF level of at least 80% for at least 20% of sessions, and the remaining 124 (17.04%) did not report data or achieve this level. Tables 2 and 3 present the descriptive data for all predictors, and Table 4 presents the results using the three-level mixed-effects model for the predictors as they relate to replication.

Target Selection and Assignment Methods

The target selection and assignment methods included four codes: *random assignment, equating procedure, neither*, and *combined*. The distribution of the four codes was as follows: *random assignment* in 17.7% ($n = 86$), *equating procedure* in 27.3% ($n = 133$), *neither* in 23.0% ($n = 112$), and *combined* in 32.0% ($n = 156$) of the outcomes. The code, *combined*, presented the highest percentage of replication (58.97%; $n = 92$), followed by *random assignment* (55.83%, $n = 48$), *equating procedure* (34.59%; $n = 46$), and *neither* (29.46%, $n = 33$). In the mixed model, when accounting for other predictors, the methods coded as *combined* ($OR = 5.09$, $p < .001$, 95% CI [2.13, 12.16]) and *random assignment* ($OR = 3.27$, $p < .05$, 95% CI [1.15, 9.33]) were statistically significant, indicating a higher likelihood for them to be related to replication than *neither* procedure. The *equating procedure* alone was not significantly associated with replication compared with *neither* procedure.

Mastery Criterion Dimensions

The mastery criterion dimensions included the performance level and the frequency of observations (as measured in session number). The outcomes had an average

Table 4. Multilevel Mixed-Effects Logistic Regression Analysis Results.

	Replication	
Predictors	*OR*	95% CI
(Intercept)	0.06	[0.00, 3.79]
Target selection and assignment methods		
Random assignment	3.27*	[1.15, 9.33]
Equating procedure	1.71	[0.66, 4.44]
Combined	5.09**	[2.13, 12.16]
Neither	1 (Ref)	
Mastery criterion dimensions		
Performance level	1.00	[0.96–1.04]
Frequency of observation	1.21	[0.77–1.90]
Contextual change types		
Target	4.63	[0.48–44.93]
Skill	4.08	[0.35–46.97]
Timing	1 (Ref)	
Random effects		
ICC	.17	
$N_{Participants}$	183	
$N_{Studies}$	64	
$N_{Outcomes}$	487	
Marginal R^2/conditional R^2	.127/.272	

Note. OR = odds ratio; CI = confidence interval; Ref = reference; ICC = intraclass correlation coefficient; *N* = number.
*$p < .05$. **$p < .001$.

performance-level requirement of 94.02% (range: 75%–100%, *SD* = 8.34) with a mean session-length requirement of 2.64 sessions (range: 1–5 sessions, *SD* = 0.76). Replication presented in 219 performance-requirement values with an average of 93.98% (range: 75%–100%, *SD* = 8.35) and 219 frequency-of-observation values with a mean of 2.64 sessions (range: 1–5 sessions, *SD* = 0.76). In the mixed model, when adjusting for other predictors, neither performance level nor the frequency of observation requirement was significantly associated with replication.

Contextual Change Types

Three contextual change types were coded: target, skill, and timing changes. Most between-comparison contextual changes were target change, with 89.7% (*n* = 437) outcomes associated with using different targets between two comparisons. Fewer were related to skill (6.6%, *n* = 32) and timing (3.7%, *n* = 18) changes. Although the target change (46.68%, *n* = 204) had a higher percentage of outcome coded as *replicated* than skill (40.63%, *n* = 13) and timing (11.11%, *n* = 2) changes, it was not significantly associated with replication compared with the reference (i.e., timing change) when controlling for other variables in the mixed model.

Model Fit

The value of ICC in the current model was .17, indicating 17% of the variance in outcomes could be explained by the difference between groups. Relating to model fit, the AIC for our three-level mixed-effects logistic regression model was 634.2. The conditional R^2 was .272, indicating 27.2% of the variance in the outcomes could be explained by the entire model that included both fixed and random effects, while the marginal R^2 of .127 represented that 12.7% of the variance could be explained by fixed effects (i.e., all the predictors).

Discussion

Given the potential for single-case comparisons to guide the selection and development of efficient, individualized interventions in special education, we examined within-participant replicability of relative efficiency results in single-case comparison studies and explored the predictive variables that may be associated with within-participant replication. Similar to the results in Ledford et al. (2021), we found inconsistent within-participant replication between comparisons; we identified fewer than half of the total outcomes that represented replication. However, our logistic regression results delineated variables that may be associated with within-participant replication.

Target Selection and Assignment Methods

Our review found that random assignment was reported for 17.7%, equating procedure for 27.3%, combined procedure for 32%, and neither method for 23% of the total outcomes. Although the studies with the *neither* coding could have included either or both methods but did not report them, this large proportion of outcomes without using or reporting either method is still concerning, especially given the repeated recommendations in the literature relating to the need to control for the difference between target sets when conducting single-case comparisons (e.g., Holcombe et al., 1994; Wolery et al., 2018). Future researchers should always explicitly address if and what methods were used in their studies to ensure equivalent target sets.

The methods reported in the outcomes included procedures to equate target difficulty, such as a logical analysis (Cariveau et al., 2021, 2022) and random assignment of targets (Ledford, 2018; Ledford et al., 2021) to distribute confounding characteristics among the conditions. Using the logistic regression analysis, we found that random assignment, when used independently, is significantly more likely to predict within-participant replication than neither procedure. Surprisingly, when an equating procedure was used independently, the procedure was not statistically significant in predicting within-participant

replication compared with neither method. In other words, between random assignment and an equating procedure, random assignment appears more likely to be associated with within-participant replication than an equating procedure. As discussed, successful replications may be related to experimental control in which changes in the dependent variable are due to the independent variable and not extraneous influences. Because both methods are used to control target-related influences among the conditions, our results seemed to indicate that randomly assigning targets among the conditions may have a greater ability to distribute and control for these confounds among the conditions than an equating procedure.

When an equating procedure is used by itself, it is plausible that the targets between conditions perceived as having equal difficulty by researchers may still differ for the learner. For example, if the learner already had some experience with or a particular preference for one of the targets (Cariveau et al., 2021; Wolery et al., 2018), the acquisition rate for the target could be affected by this idiosyncrasy. Although a baseline could be used to determine if the performance is equivalent between the targets, it may not be adequate to reveal information on learner history. Even if a learner does not emit any target responses during baseline, it is still possible that they have had exposure to some of the targets (Cariveau et al., 2022), find the target sets different in difficulty levels (Ledford et al., 2021), or prefer some of the targets. As a result, the equating procedure may not be sufficient to prevent uncontrolled idiosyncratic target characteristics from influencing the outcome, potentially reducing the likelihood of replication. However, multiple methods to equate targets do exist (see Wolery et al., 2018). Because specific analysis per procedure may not produce a sufficient number of eligible outcomes, we did not analyze different equating procedures. Further empirical analysis of these procedures is warranted to examine if and how specific procedures (or their combinations) and their interactions with the comparison context (e.g., participant characteristics) are associated with the likelihood of within-participant replication.

Even though an equating procedure itself did not appear to have an association with within-participant replication, including an equating procedure seemed to have additive effects. When using an equating procedure in conjunction with random assignment, not only was this combined method significantly more likely to be associated with within-participant replication than neither procedure, but it also had greater odds of replication ($OR = 5.09$) than using random assignment without an equating procedure ($OR = 3.27$). This difference between the combined method and random assignment alone may be expected. Although an equating procedure may not account for all possible confounds relating to targets, it could still provide protection from some threats to internal validity as researchers need to conduct at least one assessment, such as a logical analysis or expert rating, to promote target equality. As a result, the combined method may further strengthen experimental control.

Taken together, among various methods, random assignment of targets seems critical when conducting single-case comparisons that require target equivalence. When using random assignment, researchers should consider including a large set of targets for an adequate distribution of extraneous influences among the conditions to strengthen internal validity (e.g., Ledford, 2018). As Cariveau et al. (2022) point out, random assignment alone may not ensure equivalent difficulty without an adequate equating procedure such as logical analysis. Researchers should consider using both strategies whenever possible, especially given the larger odds of replication for the combined procedure than random assignment alone.

Mastery Criterion Dimensions

In analyzing the two mastery criterion dimensions—performance level and frequency of observations—we found a mean of 94.02% accuracy for the mastery performance-level requirement with a range from 75% to 100%. More than half of the total outcomes ($n = 298$) had a 100% accuracy requirement. The remaining outcomes included 65 with a 90% to 95% accuracy requirement, 113 with an 80% to 89% requirement, and 11 with a 75% requirement. The mean requirement for the observation frequency was 2.64 sessions, with a range between one and five sessions. The logistic regression analysis for the mastery criterion dimensions revealed that neither performance-level values (e.g., % correct) nor observation frequencies at the performance level were associated with within-participant replication. Nevertheless, educators and researchers are recommended to use higher criteria as they have been found to be associated with better skill maintenance (e.g., Fuller & Fienup, 2018; Wong et al., 2022). In addition, while a relation between the observation frequency and replicability was not found, extended data collection at the performance level could allow researchers to assess performance stability, increasing confidence when drawing inferences.

Differences Between Comparisons

Across the studies, three types of between-comparison contextual changes were coded. Studies typically conducted multiple comparisons across different targets, skills, and timings. Although we expected the extent to which the contexts differed between the comparisons could be related to replicability (e.g., comparisons using different skills vs. comparisons using different targets), we did not find an association. Notably, the outcome distribution was skewed toward one category, target change, which was presented in

89.7% of the between-comparison changes, while the other two categories were each presented in only 6.6% or less. This skewed distribution may not permit a meaningful interpretation of the outcomes. Future comparisons should consider systematically varying the different contexts during replications.

Unexplained Outcome Variance

Although we found that random assignment and a combined procedure in target selection and assignment were significantly associated with within-participant replication, a large proportion of the outcome variance was not captured by our logistic regression model, indicating other variables relating to within-participant replication may exist. For example, participant characteristics (e.g., cognitive functioning level, learning history, etc.) could affect replication (Fabrigar et al., 2020; Kazdin, 2011). As most reviewed studies did not report sufficient participant information that can be compared and summarized, it may be necessary for future comparative research to provide adequate information on the participant characteristics to further examine if and how they relate to replication.

Furthermore, participant experience in an initial comparison should affect the results in a subsequent comparison (e.g., carryover effects), reducing the likelihood of replication. As previously discussed, in "learning-to-learn," even if a procedure had produced faster acquisition in an initial comparison, the participant's acquisition rate may increase in the subsequent comparisons regardless of the procedure, resulting in undifferentiated outcomes in the latter comparisons (Ledford et al., 2021; Ledford & Wolery, 2013). Similarly, skill generalization from one comparison to the next could also result in undifferentiated outcomes. In this case, examining the order in which comparisons were conducted may be necessary. For example, because these effects seem to emerge in successive comparisons, comparisons conducted simultaneously could be associated with a higher likelihood of replication than successive comparisons. Examining comparison order (e.g., simultaneous vs. successive) could help reveal the association. In this review, we were only able to establish comparison orders in 25 of the 64 studies included in outcome coding, none of which reported conducting simultaneous comparisons. Future research should consider arranging and explicitly reporting different comparison orders to examine these effects to help better understand the conditions under which within-participant replication may or may not occur.

Limitations

Further limitations should be considered when interpreting the current results. Specifically, it should be emphasized that our results only revealed the association between within-participant replicability and the specific predictors included in this review using aggregated data from studies with varying purposes. We reasoned that the selected predictors were likely included in most single-case comparisons regardless of the purposes, and separating studies by specific characteristics (e.g., purposes, target behaviors) may not yield an adequate number of outcomes for this analysis, especially given the relatively fewer single-case comparisons that included replications. However, in doing so, the nuances among different studies may not have been captured. For example, a study may include a specific efficiency measure appropriate for its purpose. Thus, attempts to examine within-participant replication when conducting single-case comparisons should be systematically programmed so that sufficient data can be generated to examine replications in the context of specific research.

Relatedly, although we selected the number of sessions to mastery because most studies included this measure and only a few studies reported their results using other measures, efficiency could be measured using the number of errors, sessions, trials, or time to mastery (Wolery et al., 1991). Although a high correspondence among some measures (i.e., sessions, trials, and time to criterion) has been reported (e.g., Carroll et al., 2015, 2018), relative efficiency outcomes could still differ based on the different measures. Researchers could consider reporting their efficiency data using multiple measures or indicate if they maintain the same or similar instructional time between the conditions to allow conversion among the measures. Similarly, the accuracy-based mastery criterion was selected due to its extensive use in educational settings (Fuller & Fienup, 2018), but some researchers note that fluency (e.g., number of correct responses per unit time) could be a more reliable option (e.g., Burns et al., 2006).

In addition, when evaluating our data, we used a cutoff point of 10% difference (Ledford et al., 2021) to determine the efficiency between two interventions. Although this level of difference over a long time could produce meaningful differences, empirical investigation and discussion among the stakeholders are necessary to validate this criterion. Finally, while our analysis was limited to within-participant replication, we assume a similar analysis could be conducted for the research investigating efficiency across different participants to provide insight and clarifications when investigating other types of replications in the context of single-case comparative designs.

Research and Practical Implications

Central to single-case research, replicability is indicative of not only the possibility for special educators to use single-case designs to inform individualized intervention

selection (McComas & Burns, 2009; McGhan & Lerman, 2013) but also the generalizability of findings and the level of confidence in experimental control (Ledford & Gast, 2018). Our review and logistic regression analysis extended previous literature on single-case comparative studies by examining variables related to within-participant replication. Although our model did not capture all outcome variance—potentially indicating missing predictive variables—we identified variables associated with within-participant replication. Specifically, random assignment of targets among the comparison conditions could increase the likelihood that targets assigned to conditions do not systematically differ. We further recommend incorporating at least one equating procedure, along with random assignment, when selecting and assigning targets. Previous reviews have also suggested using multiple equating procedures and logical analysis methods to strengthen the control for target differences before random assignment (Cariveau et al., 2021, 2022).

Although the current review identified some variables with significant association with within-participant replication compared with other variables, empirical assessment of conditions under which replication occurs is still limited. As a large proportion of variance was not explained by our model, it seems crucial for researchers to identify additional variables associated with successful replication, particularly those related to various types of validity (Fabrigar et al., 2020). To do so requires researchers to conduct replication attempts with various experimental arrangements (e.g., arranging simultaneous and successive comparisons) so that sufficient data can be generated. Although successful replications increase confidence in intervention effects, we argue that data demonstrating nonreplication are equally critical for delineating possible variables responsible for differing replicability, thereby providing further guidelines on arranging single-case comparisons for learners with disabilities. In addition, including both within- and between-participant replication could further allow analyses of how various participant characteristics, contexts, and experimental arrangements relate to replicability.

For practitioners, we recommend prioritizing effective procedures over assessing efficiency, given the lack of replication in more than half of the outcomes in our and previous reviews (Ledford et al., 2021). However, evaluating efficiency may be justified under some circumstances (e.g., an excessively slow acquisition rate). In this case, practitioners may identify a target skill and compare the procedures to assess if one procedure would produce faster acquisition. This comparison should include sufficient experimental control elements to safeguard internal validity, such as those recommended above and in the previous literature (Cariveau et al., 2022; Ledford et al., 2021).

Authors' Note

LW and CY contributed equally to this work, as did SA and QAZ, and are listed alphabetically.

Acknowledgments

The authors would like to thank Katherine Nguyen for assisting with screening and coding procedures and the anonymous reviewers for valuable suggestions.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Chengan Yuan https://orcid.org/0000-0002-6316-0146

Supplemental Material

Supplemental material is available on the *Remedial and Special Education* webpage with the online version of the article.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Barton, E. E., Meadan-Kaplansky, H., & Ledford, J. R. (2018). Independent variables, fidelity, and social validity. In D. L. Gast & J. R. Ledford (Eds.), *Single case research methodology: Applications in special education and behavioral sciences* (3rd ed., pp. 133–156). Routledge.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Burns, M. K., VanDerHeyden, A. M., & Jiban, C. L. (2006). Assessing the instructional level for mathematics: A comparison of methods. *School Psychology Review*, *35*(3), 401–418. https://doi.org/10.1080/02796015.2006.12087975

Cariveau, T., Batchelder, S., Ball, S., & La Cruz Montilla, A. (2021). Review of methods to equate target sets in the adapted alternating treatments design. *Behavior Modification*, *45*(5), 695–714. https://doi.org/10.1177/0145445520903049

Cariveau, T., Helvey, C. I., Moseley, T. K., & Hester, J. (2022). Equating and assigning targets in the adapted alternating treatments design: Review of special education journals. *Remedial and Special Education*, *43*(1), 58–71. https://doi.org/10.1177/0741932521996071

Carroll, R. A., Joachim, B. T., St Peter, C. C., & Robinson, N. (2015). A comparison of error-correction procedures on skill acquisition during discrete-trial instruction. *Journal of Applied Behavior Analysis*, *48*(2), 257–273. https://doi.org/10.1002/jaba.205

Carroll, R. A., Owsiany, J., & Cheatham, J. M. (2018). Using an abbreviated assessment to identify effective error-correction procedures for individual learners during discrete-trial instruction. *Journal of Applied Behavior Analysis*, *51*(3), 482–501. https://doi.org/10.1002/jaba.460

Chow, J. C., & Hampton, L. H. (2022). A systematic review of sequential multiple-assignment randomized trials in educational research. *Educational Psychology Review*, *34*, 1343–1369. https://doi.org/10.1007/s10648-022-09660-x

Fabrigar, L. R., Wegener, D. T., & Petty, R. E. (2020). A validity-based framework for understanding replication in psychology. *Personality and Social Psychology Review*, *24*(4), 316–344. https://doi.org/10.1177/1088868320931366

Fuller, J. L., & Fienup, D. M. (2018). A preliminary analysis of mastery criterion level: Effects on response maintenance. *Behavior Analysis in Practice*, *11*(1), 1–8. https://doi.org/10.1007/s40617-017-0201-0

Gast, D. L., & Wolery, M. (1988). Parallel treatments design: A nested single subject design for comparing instructional procedures. *Education and Treatment of Children*, *11*(3), 270–285.

Holcombe, A., Wolery, M., & Gast, D. L. (1994). Comparative single-subject research: Description of designs and discussion of problems. *Topics in Early Childhood Special Education*, *14*(1), 119–145. https://doi.org/10.1177/027112149401400111

Johnston, J. M. (1988). Strategic and tactical limits of comparison studies. *The Behavior Analyst*, *11*(1), 1–9. https://doi.org/10.1007/BF03392448

Jones, K. M., Wickstrom, K. F., Noltemeyer, A. L., Brown, S. M., Schuka, J. R., & Therrien, W. J. (2009). An experimental analysis of reading fluency. *Journal of Behavioral Education*, *18*(1), 35–55. https://doi.org/10.1007/s10864-009-9082-9

Kasari, C., Shire, S., Shih, W., & Almirall, D. (2021). Getting SMART about social skills interventions for students with ASD in inclusive classrooms. *Exceptional Children*, *88*(1), 26–44. https://doi.org/10.1177/00144029211007148

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

Kirby, M. S., Spencer, T. D., & Ferron, J. (2021). How to be RAD: Repeated acquisition design features that enhance internal and external validity. *Perspectives on Behavior Science*, *44*(2–3), 389–416. https://doi.org/10.1007/s40614-021-00301-2

Kodak, T., & Halbur, M. (2021). A tutorial for the design and use of assessment-based instruction in practice. *Behavior Analysis in Practice*, *14*(1), 166–180. https://doi.org/10.1007/s40617-020-00497-w

Ledford, J. R. (2018). No randomization? No problem: Experimental control and random assignment in single case research. *American Journal of Evaluation*, *39*(1), 71–90. https://doi.org/10.1177/1098214017723110

Ledford, J. R., Chazin, K. T., Gagnon, K. L., Lord, A. K., Turner, V. R., & Zimmerman, K. N. (2021). A systematic review of instructional comparisons in single-case research. *Remedial and Special Education*, *42*(3), 155–168. https://doi.org/10.1177/0741932519855059

Ledford, J. R., & Gast, D. L. (2018). *Single case research methodology: Applications in special education and behavioral sciences* (3rd ed.). Routledge.

Ledford, J. R., & Wolery, M. (2013). Peer modeling of academic and social behaviors during small-group direct instruction. *Exceptional Children*, *79*(4), 439–458. https://doi.org/10.1177/001440291307900404

McComas, J. J., & Burns, M. K. (2009). Brief experimental analyses of academic performance: Introduction to the special series. *Journal of Behavioral Education*, *18*(1), 1–4. https://doi.org/10.1007/s10864-009-9078-5

McGhan, A. C., & Lerman, D. C. (2013). An assessment of error-correction procedures for learners with autism. *Journal of Applied Behavior Analysis*, *46*(3), 626–639. https://doi.org/10.1002/jaba.65

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. & The PRISMA Group (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLOS Medicine*, *6*(7), Article e1000097. https://doi.org/10.1371/journal.pmed.1000097

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x

Reichow, B., Barton, E. E., & Maggin, D. M. (2018). Development and applications of the single-case design risk of bias tool for evaluating single-case design research study reports. *Research in Developmental Disabilities*, *79*, 53–64. https://doi.org/10.1016/j.ridd.2018.05.008

Rumberger, R. W. (1995). Dropping out of middle school: A multilevel analysis of students and schools. *American Educational Research Journal*, *32*(3), 583–625. https://doi.org/10.3102/00028312032003583

Shabani, D. B., & Lam, W. Y. (2013). A review of comparison studies in applied behavior analysis. *Behavioral Interventions*, *28*(2), 158–183. https://doi.org/10.1002/bin.1361

Shadish, W. R., Zelinsky, N. A. M., Vevea, J. L., & Kratochwill, T. R. (2016). A survey of publication practices of single-case design researchers when treatments have small or large effects. *Journal of Applied Behavior Analysis*, *49*(3), 656–673. https://doi.org/10.1002/jaba.308

Sindelar, P. T., Rosenberg, M. S., & Wilson, R. J. (1985). An adapted alternating treatments design for instructional research. *Education and Treatment of Children*, *8*(1), 67–76. https://doi.org/10.1002/bin.1865

Snijders, T. A., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Sage. https://us.sagepub.com/en-us/nam/multilevel-analysis/book234191

Szumilas, M. (2010). Explaining odds ratios. *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, *19*(3), 227–229.

VanDerHeyden, A. M., & Burns, M. K. (2009). Performance indicators in math: Implications for brief experimental analysis of academic performance. *Journal of Behavioral Education*, *18*(1), 71–91. https://doi.org/10.1007/s10864-009-9081-x

Wolery, M., Doyle, P. M., Ault, M. J., Gast, D. L., Meyer, S., & Stinson, D. (1991). Effects of presenting incidental information in consequent events on future learning. *Journal of Behavioral Education*, *1*(1), 79–104. https://doi.org/10.1007/BF00956755

Wolery, M., Gast, D. L., & Ledford, J. R. (2018). Comparison designs. In D. L. Gast & J. R. Ledford (Eds.), *Single case research methodology: Applications in special education and behavioral sciences* (3rd ed., pp. 283–334). Routledge.

Wong, K. K., Fienup, D. M., Richling, S. M., Keen, A., & Mackay, K. (2022). Systematic review of acquisition mastery criteria and statistical analysis of associations with response maintenance and generalization. *Behavioral Interventions*, *37*(4), 993–1012. https://doi.org/10.1002/bin.1885

Wu, S., Crespi, C. M., & Wong, W. K. (2012). Comparison of methods for estimating the intraclass correlation coefficient for binary responses in cancer prevention cluster randomized trials. *Contemporary Clinical Trials*, *33*(5), 869–880. https://doi.org/10.1016/j.cct.2012.05.004

Yuan, C., & Zhu, J. (2020). An evaluation of prompting procedures in error correction for children with autism. *Behavioral Interventions*, *35*(4), 581–594. https://doi.org/10.1002/bin.1733