Single Case in the Social Sciences 1(1)
DOI: 10.15763/issn.1936-9298.2024.1.1.61-83
http://singlecasejournal.com/

Systematic Review of Gaps in Single Case Design Research: Evaluation of Study Quality and Rigor Using the Single Case Analysis and Review Framework

Tyler-Curtis C. Elliott[1], Kevin M. Ayres[1], Joseph B. Hart[1], and Jennifer R. Ledford[2]

Abstract
As the use of single case research design (SCRD) to answer socially significant research questions increases, so too should the rigor and quality of those designs. Higher rigor and quality decreases the chance of threats to internal validity and increases the chance of replication, both of which are critical to determining the effectiveness of an intervention. We conducted a systematic review of systematic literature reviews (k =18) that scored the quality and rigor of SCRD papers using the Single Case Analysis and Review Framework (SCARF; Ledford et al., 2023). On a continuous rigor/quality scale of 0-4, the 1,251 articles in all included reviews obtained an average of 2.14 with a standard deviation of 0.62 suggesting large gaps in the rigor/quality of SCRD. We discuss the implications of these gaps and offer recommendations for researchers to increase the rigor and quality of their SCRD studies.

Keywords
replication, internal validity, single case analysis and review framework

[1]Center for Autism and Behavioral Education Research, University of Georgia, Athens, GA, USA
[2]Department of Special Education at Peadbody College, Vanderbilt University, Nashville, TN, USA

Corresponding Author:
Tyler-Curtis C. Elliott, Rivers Corssing, 850 College Station Rd Bldg. 2, Athens, GA 30605. E-mail: tce62651@uga.edu

Notes:
Joseph B. Hart is now at Bluesprig Pediatrics.

Although many consider randomized control trials (RCT) the gold standard for determining a causal relation between experimental independent and dependent variables (Hariton & Locascio, 2018), the research question should guide selection of the research design (Creswell & Creswell, 2023). For example, if answering a research question involving analysis at the individual level, group aggregated information in an RCT may prove inadequate. As Skinner (1953) wrote, "A prediction of what the average individual will do is often of little or no value in dealing with a particular individual" (p. 19). In addition, if evaluating the effect of an intervention on a rare dependent variable (e.g., pica), researchers are unlikely to be able to recruit large samples for evaluation. In these cases, single case research design (SCRD) allows the best means to answer the research questions. Generally, SCRD is approached from an inductive lens which involves participants serving as their own control. Using formative data collection, participants' performance is evaluated across at least two conditions in a replicated manner to allow for evaluation of a causal conclusion. Because the logic of SCRD differs inherently from group design, there are many misconceptions about SCRD (Aeschleman, 1991; Dermer & Hoch, 1999). In addition, some research textbooks (Privitera, 2020) refer to SCRD as a quasi-experimental process (due to lack of randomization; McDonnell & O'Neill, 2003) suggesting inadequate methodological rigor for drawing conclusions about the functional relation. Although there are some instances in which researchers may elect to use randomization (see Tanious & Onghena, 2019), the logic of SCRD does not rely on randomization to control for threats to internal validity.

For decades, researchers have identified ways to strengthen the rigor of SCRD and urged their colleagues to take affirmative steps to raise the standard (Wolery & Ezell, 1993; Wolery, 1994). Although use of SCRD (Gast & Ledford, 2018) and reviews of SCRD continue to increase (Maggin et al., 2011), the calls for higher standards in SCRD are still ongoing (Ganz & Ayres, 2018; Ledford et al., 2023). For example, researchers have documented limited data on the implementation fidelity of independent variables (Barnett et al., 2014; Falakfarsa, et al., 2022; Ledford & Wolery, 2013; Preas et al., 2024; Sanetti et al., 2012; Swanson et al., 2011), social validity (Ennis et al., 2013; Snodgrass et al., 2022; Wellons et al., 2023), and need for better descriptions of factors critical to replication (Lane et al., 2007; Wolery et al., 2011). This desire for high-quality SCRD studies has led to the development of multiple reporting standards, such as calls for increased rigor in peer reviewed journals (Ganz & Ayres, 2018; Horner, 2005) and by professional organizations (CEC, 2014; WWC, 2020). Specifically designed tools have also been developed such as the Risk of Bias in N-of-1 Trials scale (Tate et al., 2014), Risk of Bias (Reichow et al., 2018), Single-Case Reporting Guideline In Behavioural Interventions (Tate et al., 2016), and Single Case Analysis and Review Framework (SCARF; Ledford et al. 2016; Ledford et al., 2023).

SCARF is one quality indicator framework that allows a comprehensive analysis of multiple related studies and provides a quantifiable metric for both study rigor (controlling for threats to internal validity), quality (adequate reporting), and outcomes (effectiveness of the experimental manipulation). The rigor/quality score is a continuous variable with a range of 0-4 that is calculated such that the average rigor scores are doubled then averaged with the quality scores to obtain the combined rigor/quality score (Ledford et al., 2018). The outcome scores are ordinal on a scale of 0-4 and are determined using visual analysis standards. These scores (rigor/quality and outcomes) are graphed on a scatterplot that shows the relationship between quality/rigor of designs and the outcomes for a large set of designs (Ledford et al., 2018). Since SCARF includes a continuous variable for rigor/quality between a range of 0-4, it allows more nuanced analysis beyond just meeting standards or not (Hardy et al., 2022). Most often, SCARF is used in a narrative review to evaluate the

believability and replicability of papers. For example, Zimmerman and Ledford (2017) evaluated the evidence for social narratives for students without autism and used SCARF to determine which studies had adequate rigor/quality to be believable and replicable. Their results determined that although there were studies evaluating social narratives for people without ASD, it is not yet considered an evidence-based practice and the authors noted specific methodological concerns with the studies (e.g., explicit procedures).

Since SCARF includes the majority of the indicators from other major standards and also includes its own unique indicators (Hardy et al., 2022), a large-scale evaluation of SCRD studies with the SCARF tool may elucidate gaps that exist in the SCRD literature that large scale reviews using other standards may not identify. To date, we know of no comprehensive methodological evaluations of SCRD across multiple domains (i.e., dependent variables, populations, and independent variables). However, some researchers have evaluated SCRD studies focused on a specific set of populations (e.g., Hott et al. 2023), or focused on specific aspects of research quality (e.g., procedural fidelity; Barnett et al., 2014; Falakfarsa, et al., 2022; Ledford & Wolery, 2013; Preas et al., 2024; Sanetti et al., 2012; Swanson et al., 2011). Thus, the field could benefit from a systematic review that includes a large sample of articles across dependent variables, journals, populations, and independent variables).

Using a sample of SCRD in published and unpublished literature reviews, we intended to answer the following research questions:

1. What is the average (and standard deviation) rigor/quality score achieved by this sample of SCRD studies when measured using the SCARF?
2. What are the common "gaps" in this sample of SCRD that result in lower rigor/quality scores?

Method

Language

In this manuscript "SCARF articles" or "articles" refers to the review papers that used SCARF to code for rigor/quality. "Cases" refers to the designs coded within the SCARF articles. Note that SCARF is coded on the design level of analysis rather than the study. For example, if a study using a multiple baseline across behaviors replicated across three participants, would be coded as three separate experiments (i.e., one for each participant); a multiple baseline design across three participants would be coded as a single experiment. We consider a design to be the smallest unit of a study that can stand alone to evaluate a causal relation (e.g., a study that includes three participants with their own demonstrations of effect using a withdrawal design would be coded as three separate designs). If researchers used a combination of design elements to answer multiple research questions (e.g., demonstrative and comparative question answered with a multiple baseline with an embedded alternating treatment design) each comparison the researcher was interested in (defined a priori) would be scored separately.

We refer to the different SCARF indicators as either quality indicators or rigor indicators in this manuscript. Rigor refers to indicators that focus on the experimental rigor (controlling for threats to internal validity) while quality refers to the quality of reporting (writing descriptions that are replicable). Of course, if authors have high experimental rigor but do not report this, then they will receive a lower score for rigor.

Screening

We conducted a systematic search of the literature by typing *"Single Case Analysis and Review Framework"* into Google Scholar on November 7, 2023. We used Google Scholar because prior research has shown that it is considered more scholarly and comprehensive than other databases (Gusenbaur, 2018; Howland et al., 2009). Once we removed duplicates, Google Scholar profiles, and files (e.g., SCARF itself), we were left with 70 articles to screen. The first author obtained the full text for each result to ensure that they (a) were available in English or Spanish (so they could be read by one of the authors) (b) were a review of literature (published or unpublished), (c) used SCARF to code for both rigor and quality (d) collected interrater agreement data on at least 30% of the included articles, and (e) reported average agreement between raters as at least 90% or more. This inclusion criterion resulted in identification of 18 review papers. See Figure 1 for a PRISMA flowchart.

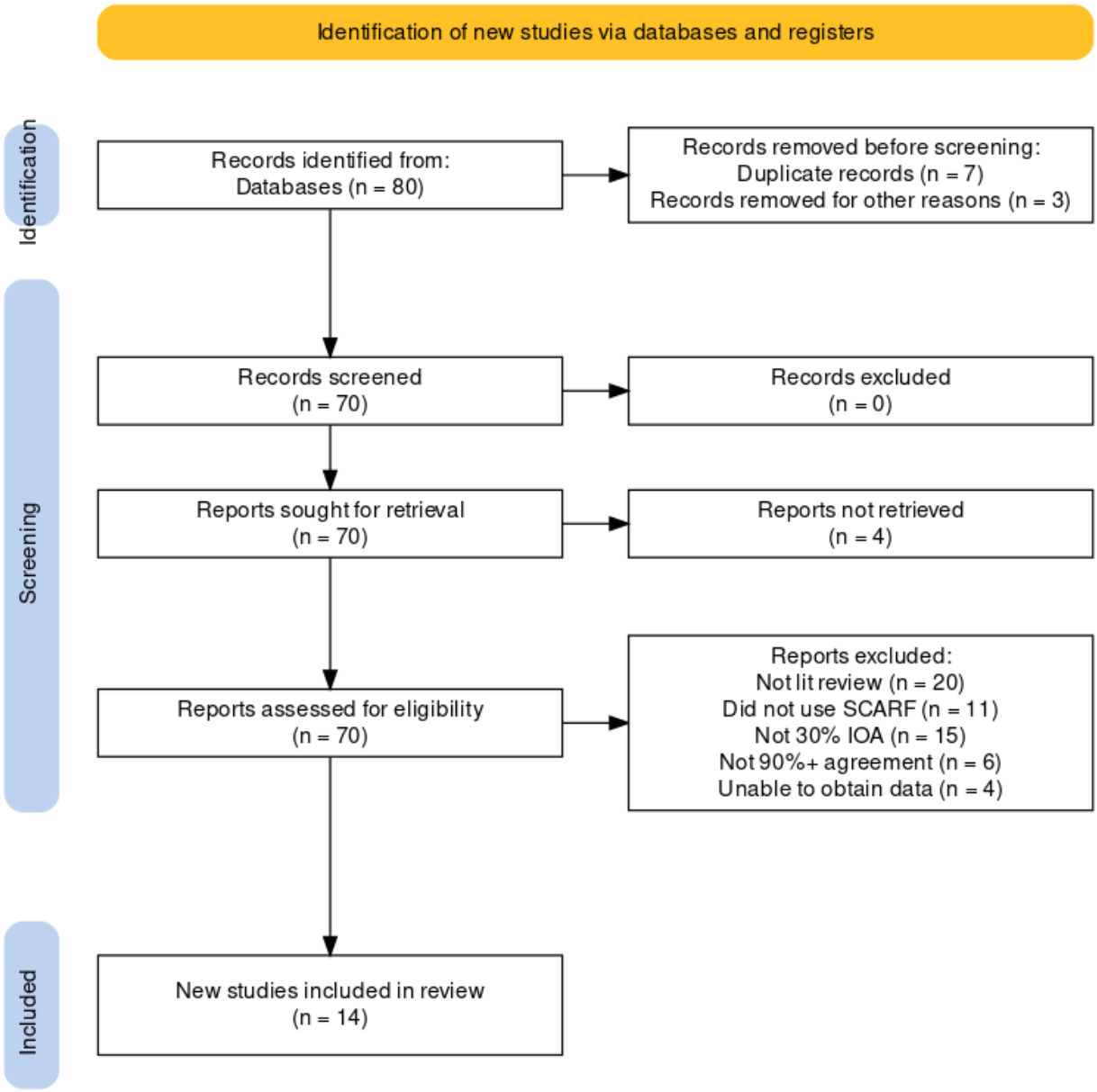
To ensure accuracy of screening and inclusion, the third author replicated the primary search's predefined strategy, which included the same database and search terms. This reproducibility check led to a 98.6% agreement rate between the authors using a point-by-point evaluation method for the initial search results. After a discussion on the one article with disagreement, both authors reached a consensus to exclude the study due to it being written in Chinese (thus, no member of the research team was able to interpret it). Subsequently, each remaining article was independently evaluated by both authors against the established inclusion criteria, culminating in unanimous agreement to include 18 articles in the final review. We did not pre-register this study prior to screening, but still conducted the study in a systematic manner, following our proposed methods.

To ensure that Google Scholar was a valid database for finding all of the literature relevant to this review, the first author conducted a search in PsycInfo by searching for *"Single Case Analysis and Review Framework"*. Since this search term only resulted in only three results, we then searched *Single Case Analysis and Review Framework* (without quotes) resulting in 204 articles to screen. Using the same screening criteria and the same timeframe as used in Google Scholar, only four papers met final inclusion for review. Thus, we used the results from Google Scholar in this review as it resulted in a more comprehensive sample of the literature.

Obtaining SCARF Data

For each article, we searched for any open access SCARF Excel files uploaded by the authors. This included looking in the paper for reference to open access material, searching for the name of the paper in Google, and looking at the authors Center for Open Science accounts. If we did not find any open access SCARF files for that publication, we emailed the corresponding author (and/or primary advisor for gray literature) asking for access to their data (email available via Supplemental Materials 1). We did not re-code the original studies nor did we collect interrater agreement on these papers. Our inclusion criteria required that interrater data be collected for at least 30% of the included cases and with 90% or more agreement between observers, thus we assumed the data sent by authors was accurate.

Figure 1. PRISMA Flow Chart.



Note. Created with Haddaway et al., 2022.

Rigor/Quality Scores

Of the 18 included reviews, we found open access SCARF Excel files for three and researchers shared the data by responding to our email for 11 (See Table 1). Thus, we were able to obtain datasets for 77.77% of the reviews that met inclusion criteria. These 14 studies sum 531 cases using SCARF 1.0 and 720 cases coded using SCARF 2.0. The included SCARF reviews spanned across 12 journals and included gray literature. For each of the 14 included reviews, we extracted the combined rigor/quality score for each case from the raw SCARF excel sheets. This variable is a continuous variable with a possible range of 0-4. We calculated both the average and standard deviation for cases coded with SCARF 1.0 and 2.0 separately. Supplemental Materials 2 provides a list of differences between SCARF 1.0 and 2.0 (according to the SCARF authors). The primary difference is that some SCARF 2.0 items include more detail.

Individual Item Scores

SCARF is organized such that individual items are grouped into the categories of participant descriptions (quality), dependent variable descriptions (rigor), dependent variable reliability (quality), condition descriptions (quality), independent variable reliability (rigor), social/ecological validity (quality), and sufficiency of data (rigor). We aggregated data based on each SCARF quality/rigor item (indicator). For example, for the SCARF question "do authors report formal test results (e.g., IQ, language competence, achievement)?" we extracted the coded responses (yes/no/NA) from all 1,251 cases. In SCARF, "not applicable (N/A)" is coded if the prior question was answered as a "no" making the current question inapplicable (and thus, does not add points towards rigor/quality). For example, if no data are collected to evaluate dependent variable reliability (i.e., interobserver agreement data), "N/A" is scored for questions such as whether those data were collected frequently and whether reliability outcomes were favorable. For the majority of questions, SCARF uses the response "yes" to indicate higher quality/rigor (e.g., Do authors report any data related to fidelity of implementation?). For these questions, we calculated the average item score using the following formula: yes/(yes+no+N/A). For example, if 657 of the 720 cases from SCARF 2.0 put "yes" for the question "Do authors report any data related to fidelity of implementation?", we would divide 657 by 720 to identify that 91.25% of those articles met that quality/rigor indicator.

Some scores in SCARF are reverse coded, that is, an answer of "no" indicates higher quality/rigor (e.g., Do authors report the use of self-report fidelity only). In this case, we used the following formula: no/(no+yes+NA). In this case, if 657 of 720 cases using SCARF 2.0 marked "no", we calculated that 8.75% of those articles met that indicator. Lastly, the question "Did data collection begin simultaneously during initial baseline or probe conditions for at least three tiers?" only applies if the previous question (i.e., Is the design a multiple baseline or multiple probe design?) was answered as "yes". Thus, for this question, we used the formula yes/(yes+no). Relevant formulas can be found via Supplemental Materials 3. For example, if 50 of the 720 cases using SCARF 2.0 identified that a multiple baseline was used, and 25 of those cases answered "yes" to the indicator "did data collection begin simultaneously", we calculated that 50% of the cases met that indicator.

Results for each individual item were calculated from the total number of cases included within each SCARF iteration (n = 531 and 720, respectively). For example, the percentage of experiments with acceptably high IOA data is reported out of a total number of experiments rather than out of the number of experiments that included IOA data collection.

Table 1. List of Review Papers that Met Inclusion Criteria.

Paper	Journal	Focus	Method	Obtained
Zimmerman, Ledford, Severini, Pustejovsky, Barton & Lloyd (2018)	Research in Developmental Disabilities	Antecedent sensory-based interventions for young children	Email: Zimmerman	No
Eyler & Ledford (2023)	Journal of Early Intervention	Time delay for young children	Center for Open Science	Yes
Nanda (2021)	Gray Literature	Social skills to young students with ASD	Email: Nanda and Missall	No
Zimmerman & Ledford (2017)	Journal of Early Intervention	Social narratives for children without ASD	Email: Zimmerman	Yes
Gibbs & Tullis (2021)	Review Journal of Autism and Developmental Disorders	Untrained stimulus relations for students with ASD and other DD	Email: Tullis	Yes
Ledford & Windsor (2022)	Topics in Early Childhood Special Education	Imitation for young children with DD	Email: Ledford	Yes
Gibbs, Tullis, Conine, & Fulton (2024)	Journal of Developmental and Physical Disabilities	Derived relational responding (beyond coordination) for students with ASD and other DD	Email: Tullis	Yes
Barton, Murray, O'Flaherty, Sweeney, & Gossett (2020)	American Journal on Intellectual and Developmental Disabilities	Object play to young children with DD	Email: Barton	Yes
Chazin, Ledford, & Pak (2021)	Journal of Speech-Language Pathology	Augmented input for people using AAC device	Email: Chazin	Yes
Rubio, McMahon, & Volkert (2021)	Journal of Applied Behavior Analysis	Physical guidance as an open-mouth prompt for children with feeding disorder	Email: Volkert	No
Chazin, Velez, & Ledford (2022)	Journal of Behavioral Education	Interventions for escape function with people with DD	Email: Chazin	Yes
Ledford, Trump, C., Chazin, Windsor, Eyler, & Wunderlich (2023)	Behavioral Interventions	Interruption and redirection procedures for people with ASD	Center for Open Science	Yes

Scheibel, Chen, Zaeske, Wills, & Zimmerman (2022)	Journal of Positive Behavior Interventions	Teacher self-monitoring to increase fidelity	Center for Open Science	Yes
Scott (2022)	Gray Literature	Escape-maintained inappropriate mealtime behavior	Email: Saini	Yes
Snyder (2023)	Gray Literature	Indirect, direct and functional analysis assessment methods	Email: Snyder	Yes
Rubio (2021)	Gray Literature	Behavioral interventions for avoidant/restrictive food intake disorder in young children	Email: Rubio & Roach	No
Ledford & Pustejovsky (2023)	Behavior Interventions	Stay-play-talk for social behaviors of young children	Email: Ledford	Yes
Herrod, Snyder, Hart, Frantz & Ayres (2023)	Behavior Modification	Evaluations of Premack Principle	Email: Herrod	Yes

Results

The 531 cases that used SCARF 1.0 received an average rigor/quality score of 1.89 and a standard deviation of 0.55. For the 720 cases coded using SCARF 2.0, the average rigor/quality score was 2.32 with a standard deviation of 0.60. All data (individual by case) can be found via Supplemental Materials 3. Aggregated information on each indicator can be found in Table 2 and visually depicted in Figure 2.

Participant Descriptions (Q)

For the SCARF participant questions, an average of 98.49% (1.0) and 96.11% (2.0) of cases included demographic information about the participants (e.g., age, eligibility category). Only 64.72% (1.0) and 60.42% (2.0) reported formal test results for participants. Approximately 56.42% (1.0) and 76.53% (2.0) reported "general" information about participants (e.g., relevant language skills, achievement levels). Lastly, 47.74% (1.0) and 49.44% (2.0) reported inclusion criteria for participants.

Dependent Variable Descriptions (Q)

For SCARF dependent variable questions, 69.25% (1.0) and 96.25% (2.0) of cases used operational definitions. Approximately 62.83% and 74.31% of cases used examples and non-examples to supplement the operational definitions. An average of 91.32% (1.0) and 96.53% (2.0) of cases adequately described a dependent variable measurement system, and approximately 70.51% (1.0) and 60.83% (2.0) of cases described the use of that system (e.g., training of data collectors and how the data were collected).

Dependent Variable Reliability (R)

For SCARF dependent variable reliability data, 97.36% (1.0) and 97.50% (2.0) of cases reported some reliability data taken throughout the case regardless of condition. Approximately 83.02% (1.0) and 87.78% (2.0) of cases reported aggregate reliability data in all applicable conditions (e.g., baseline, treatment A, treatment B, etc.). A point-by-point reliability measurement system was reported in 55.58% (1.0) and 67.78% (2.0) of cases. Finally, the reliability data reporters were described as "naïve observers," or observers that had little to no experience with the experiment, in approximately 1.13% (1.0) and 1.39% (2.0) of cases.

Condition Descriptions (Q)

Authors reported descriptions of each condition in 65.47% (1.0) and 93.33% (2.0) of cases. In those descriptions, 46.69% (1.0) and 58.59% (2.0) of cases described the dosage of the intervention package used, and 56.04% (1.0) and 89.58% (2.0) adequately described the setting of the study. Lastly, 52.45% (1.0) and 33.47% (2.0) of cases adequately described the intervention implementers and their background/skills.

Independent Variable Reliability (R)

Independent variable reliability, more commonly known as procedural fidelity, can be defined as the measure of the implementation of the procedures as initially intended (Barton et al., 2018). Within this SCARF review, authors reported collecting procedural fidelity in 51.51% (1.0) and 50.97% (2.0) of cases with 49.81% (1.0) and 50.14% (2.0) of the total cases using a second observer to collect procedural fidelity (rather than using the interventionist to collect their own fidelity data).

While recent advances in SCRD have called the unofficial acceptable threshold of procedural fidelity into question (Jones & St. Peter, 2022; Ledford et al., 2023), researchers have used the 80% point as a benchmark for the quality of treatment implementation. Authors reported procedural fidelity as being greater than or equal to 80% in 42.08% (1.0) and 45.28% (2.0) of cases. Approximately 24.72% (1.0) and 37.08% (2.0) of cases reported collecting procedural fidelity in all relevant conditions, and 23.02% (1.0) and 33.89% (2.0) of cases obtained procedural fidelity in at least 20% of sessions (another unofficial, but widely recognized, single-case design standard).

Another area assessed in SCARF regarding the independent variable reliability is whether the reliability reported is separated out across relevant conditions and/or implementers. Of the cases included in this review, 6.42% (1.0) and 17.92% (2.0) of cases separated the independent variable reliability data. Finally, agreement on the independent variable reliability is also assessed and 1.7% (1.0) and 6.25% (2.0) of cases included this data.

Social/Ecological Validity (Q)

For social and ecological validity, 10.75% (1.0) and 20.83% (2.0) of cases reported acceptability ratings of the intervention from the relevant stakeholders (e.g., parents, caregivers, teachers, etc.). Approximately 1.51% (1.0) and 1.25% (2.0) of cases reported any psychometric data for the participants, and 5.28% (1.0) and 4.31% (2.0) of cases reported the importance or the social significance of the results obtained. Finally, 55.66% (1.0) and 49.86% (2.0) of cases reported interventions used were in the subjects' typical environment (i.e., school or home).

Sufficiency of Data (R)

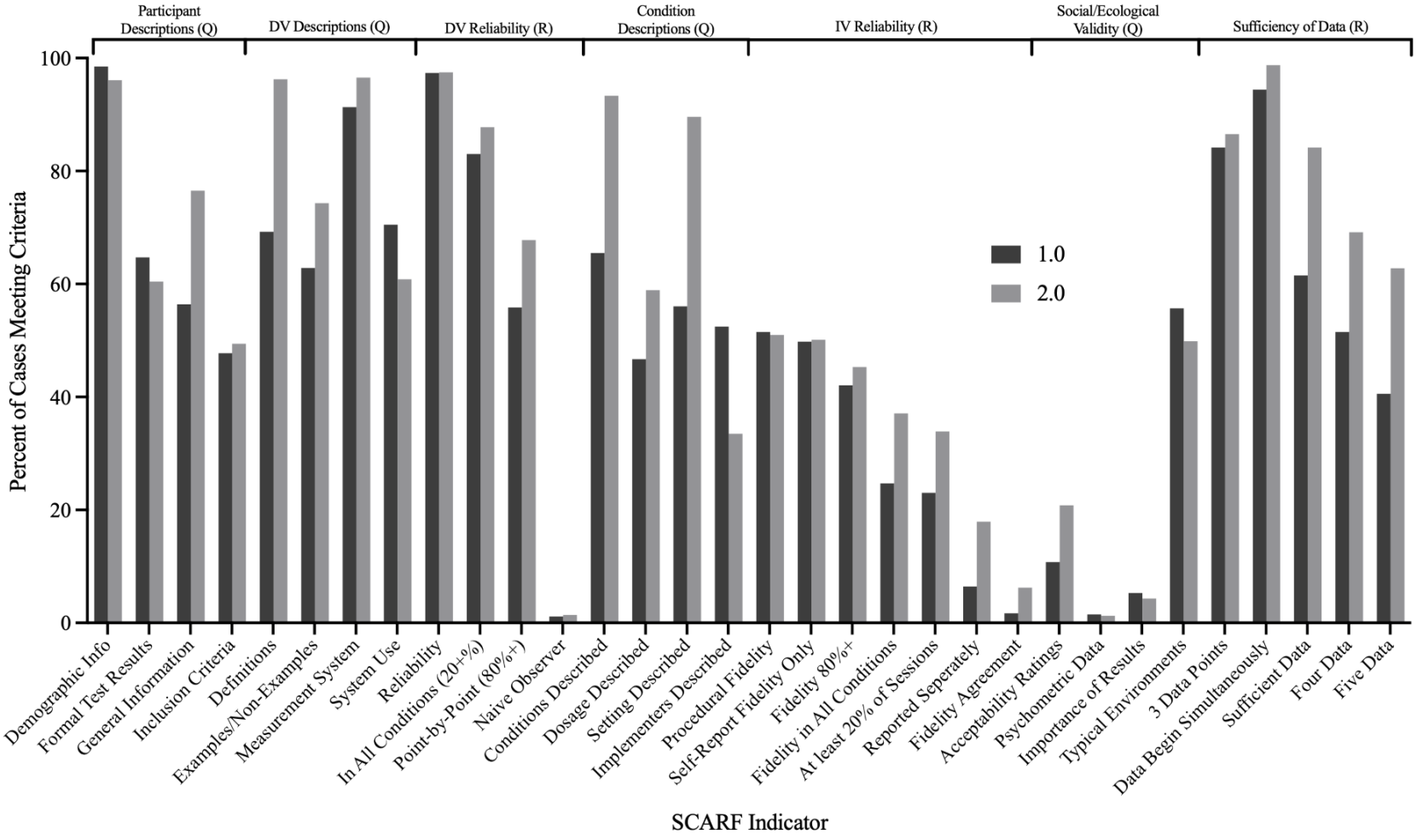
To gauge the effectiveness of an independent variable on some dependent variable(s), sufficient data is needed to verify the results. Authors reported at least three data points and three opportunities to demonstrate an effect in 84.15% (1.0) and 86.53% (2.0) of cases. For those who used a time-staggered design (e.g., multiple baseline), data collection began simultaneously in 94.44% (1.0) and 98.77% (2.0) of cases. The designs contained sufficient data to detect all threats to internal validity in 61.51% (1.0) and 84.17% (2.0) of cases. At least four data points were reported in 51.51% (1.0) and 69.17% (2.0) of cases and at least five data points in 40.57% (1.0) and 62.78% (2.0) of cases.

Table 2. Percentage of Cases Meeting Criteria for Each SCARF Indicator in Table Format.

Section	Indicator	SCARF 1.0	SCARF 2.0
Participant Descriptions	Demographic information	98.49	96.11
	Formal test results	64.72	60.42
	General information	56.42	76.53
	Inclusion Criteria	47.74	49.44
Dependent Variable Descriptions	Definitions	69.25	96.25
	Examples/nonexamples	62.83	74.31
	Measurement system	91.32	96.53
	System Use	70.51	60.83
Dependent Variable Reliability	Collected reliability	97.36	97.50
	Reliability in all conditions (20+%)	83.02	87.78
	Point-by-point	55.58	67.78
	Naive Observer	1.13	1.39
Condition descriptions	Conditions described	65.47	93.33
	Dosage described	46.69	58.59
	Setting described	56.04	89.58
	Implementers described	52.45	33.47
Independent Variable Reliability	Procedural fidelity	51.51	50.97
	Self-report fidelity only	49.81	50.14
	Fidelity at 80%+	42.08	45.28
	Fidelity in all conditions	24.72	37.08
	For at least 20% of sessions	23.02	33.89
	Reported separately	6.42	17.92
	Fidelity agreement	1.7	6.25
Social/Ecological Validity	Acceptability ratings	10.75	20.83
	Psychometric data	1.51	1.25
	Importance of results	5.28	4.31
	Typical environments	55.66	49.86

Sufficiency of data	Three data points and effect opportunities	84.15	86.53
	Data begin simultaneously	94.44	98.77
	Sufficient data	61.51	84.17
	Four data points	51.51	69.17
	Five data points	40.57	62.78

Figure 2. Percentage of Cases Meeting Criteria for Each SCARF Indicator in Graphed Format.



Note. In this table, we use simplified names that cover the theme of the indicator. See SCARF 2.0 spreadsheet (available open access online) for full descriptions of indicators.

Discussion

As use of SCRD continues to grow (Gast & Ledford, 2018), so do standards for SCRD (CEC, 2014; Ganz & Ayres, 2018; Horner, 2005; Kratochwill, 2010; Ledford et al. 2016; Ledford et al., 2023; Reichow et al., 2018; Tate et al., 2014; Tate et al., 2016; WWC, 2020). One specific tool for evaluating SCRD, SCARF, uses a continuous metric to score study rigor/quality. Since SCARF uses a continuous metric, it is possible to take a large sample of studies and aggregate them based on their SCARF rigor/quality score.

We conducted a systematic literature review of other SCRD reviews that used SCARF to code for rigor/quality. We found 18 reviews that met our inclusion criteria and were able to obtain databases for 14. Of the 1251 cases, we found an average of 1.98 rigor/quality score for SCARF 1.0 and 2.32 average for SCARF 2.0. Since a score of two or less is considered to be less valid for drawing conclusions (Zimmerman & Ledford, 2017), these results suggest that there are many areas in which the rigor and quality of SCRD research could be improved. We evaluated scores on individual SCARF indicators and found that there were some indicators that researchers infrequently met, contributing to this low quality score. Our findings were similar to that of small-scale or narrow reviews focused on the quality and/or rigor of SCRD. For example, Hott et al. (2023) found a large variance (33-77%) in how studies aligned with Single-Case Reporting Guideline In BEhavioral interventions (SCRIBE) standards. Although the SCRIBE standards are different from those of SCARF and the authors had a much smaller sample size (n=74), our findings corroborate theirs: Much work still needs to be done to increase the quality and rigor of SCRD. Given these findings, we make suggestions for the field across two categories: Quick fixes and strategic solutions.

Low Effort Solutions

Although these solutions are a parsimonious first attempt to increase the quality and rigor of SCRD, each of these suggestions should be taken within the context of each study, as not all of these solutions may be applicable (e.g., instances in which families are unable to provide comprehensive diagnostic information about the participant).

Experimental Descriptions. Many of the SCARF indicators absent from the cases in this review relate to reporting standards. Researchers could quickly and easily address the deficiencies and communicate more explicitly and completely about their studies. These simple fixes do not require changing any aspects of the study design, rather they require changes in reporting about methods. Accurate reporting of all aspects of an experiment is critical because if descriptions are unclear or not operationalized, interpretability of the study becomes difficult (Lane et al., 2017) and also may lead to implementers unintentionally using different procedures (Morgan & Morgan, 2008). Having procedures clearly described is so critical to SCRD, that the field of applied behavior analysis (which includes SCRD design as one of their practitioner competencies; BACB, 2023) includes it as one of their fields core dimensions (Baer et al., 1968).

Based on the results of this review of SCARF studies, we suggest that researchers adequately describe all experimental conditions including a detailed and replicable description of the baseline characteristics. In addition, researchers should report information about session durations, pace, and how frequently sessions occurred, such that readers can calculate some proxy of dosage and replicate or adapt accordingly (Ma & Travers, 2022). Lastly, researchers should report the setting with sufficient detail that readers can determine if the intervention in question can apply to their settings and report the implementers such that readers can determine if their implementers have the knowledge, skill, and capacity to be able to replicate the procedures. Each of these suggestions will increase the SCARF quality/rigor ratings of a manuscript and, more importantly, increase its replicability. As with all descriptions, if researchers are unable to provide detail about critical information due to journal page limits, most journals now have the option to include these descriptions as supplemental information. Authors also have the option to upload supplemental information to open science platforms or even preregister their study with these platforms. Preregistering has the added benefit of explaining how researchers planned quality and rigor from the onset of the study. Regardless of whether authors enter information through preregistration of supplemental information, authors can embed these open science links into their manuscripts.

Participant Descriptions. In group design, random selection of subjects results in a variable range of participant characteristics, hypothetically increasing the likelihood generalizability (Birnbrauer, 1981). However, in SCRD, participant selection is often intentional, making detailed reporting of participant variables even more critical to determine the replicability to other subjects (Wolery & Ezell, 1993). For example, simply describing students as Autistic, although an official diagnostic label, is unlikely to capture the idiosyncrasies of the participant and their learning history nor information about how this label became assigned (i.e., medical diagnosis, educational diagnosis, or self diagnoses). This lack of information will make it extremely difficult to determine for whom the intervention is effective (Wolery, 2013). Researchers should report inclusion criteria to provide a systematic way to select participants who are likely to have similar characteristics and skill repertoires. In addition, by describing relevant participant characteristics (e.g., information about participants that relate to the effectiveness of the intervention given the researchers theory of change) researchers can help determine which participants are likely to (or not) benefit from the intervention.

Strategic Solutions

Naive Observers. Prior research has demonstrated that using naive observers can result in different interpretations of outcomes (Chazin et al., 2018). Yet results of this review suggest that researchers almost never used observers that were naive to the experimental conditions. For experiments in which the data collection context (e.g., data collection probes) is different from the teaching context, the research team can keep the second observer masked to the experimental condition by only bringing them into the experimental context only during data collection. For example, Elliott (2024) used a feedback package to increase staff usage of evidence-based practices during structured play. The data collection involved watching the staff for 15 minutes play with students and was distinctly separate from the feedback package (which was delivered after the observation). Because of this, the secondary observer stayed naive to which participant was in intervention or baseline by entering the experimental context for the data collection sessions and leaving before the primary researcher delivered (or in baseline did not deliver) the feedback package.

When the data collection is embedded into the intervention context, keeping the secondary observer completely masked from the procedural variables may not be possible. However, one other solution is to video/audio record the experimental conditions and have the secondary observer code them in a random order. Although the secondary observer will see the procedural variables, they will be masked to key information such as if the participant is in baseline or intervention, what are the control variables/experimental variables, and how many sessions of baseline/intervention the student has received. For example, Zimmerman et al. (2020) evaluated social narratives and visual supports as an antecedent intervention and recorded all experimental conditions such that the data collector could be masked. The authors note that complete masking was not possible because the visual support was clearly present in the video. However, this system still likely reduced expectancy bias beyond what would have occurred if the data collector coded each session live and in order.

Of course, there are instances in which masking observers is not possible (Petursdottir & Carr, 2018). For example, when conducting a study on interval-based toilet training in schools, it would be illegal and unethical to record sessions and impossible to mask observers because the independent variables are blatant (e.g., praise, delivery of reinforcer). Even though masking observers is not possible in some experimental contexts, only 1.13% (1.0) and 1.39% (2.0) of the cases included in this review used masked observers, making this an area with opportunity for substantial improvement.

Acceptability Ratings

One other area for improvement amongst SCRD researchers is the use of acceptability ratings. Only 10.75% (1.0) and 20.83% (2.0) of researchers evaluated feasibility and acceptability ratings. Even less (1.51%; 1.25%) used psychometrically validated acceptability ratings. Researchers recommend using psychometrically validated rating because it is considered to be less "subjective" (Snodgrass et al, 2022). There are many established and psychometrically validated acceptability rating tools that researchers can adapt and use in their studies. This includes the Behavior Intervention Rating Scale (Elliot & Von Borck Treuting, 1991), Consultation Acceptability Satisfaction Scale (Dufrene & Ware, 2018), Intervention Rating Profile (Martens & Witt, 1982), Children Intervention Rating Profile (Kratochwill, 1985), and Treatment Acceptability Rating Form (Reimers et al., 1991).

One other measure of social validity is how well we actually make a difference in clients lives (Kazdkin, 1977). Barton et al. (2018) suggest that measures that are less susceptible to bias should be used to measure the importance of goals, procedures, and outcomes. Still, very few researchers (5.28%; 4.31%) evaluate importance of outcomes through less subjective measures such as normative comparisons or blind ratings of outcomes (i.e., acceptability, feasibility, importance of results). These findings replicate others, such as Ennis et al. (2013) who found few examples of researchers using normative comparisons as an objective measure of social validity. One example that researchers can use as a guide, is Ennis et al. (2021) who used peer comparisons as an objective measure of the significance of increases in academic engagement. Lastly, researchers can and should use assessments for evaluating acceptability with direct consumers. For example, Hanley et al. (2005) used colored pieces of paper to represent their experimental conditions and once the experimental evaluation was complete, they allowed participants to select the condition they received.

Procedural Fidelity

Collecting procedural fidelity data on both control and independent variables in an experiment is of critical importance. During an experiment, this data allows researchers to detect implementation drifts or changes (Wolery, 1994), so that researchers can retrain implementers. In addition, this information is critical to ensuring that implementers actually did what they reported in their procedures. Given that different levels of procedural fidelity can influence treatment outcomes (Holcombe et al. 1994), this information is critical to determining the relationship between the independent variable dosage and the achieved outcomes. Similar to other reviews (Barnett et al., 2014; Falakfarsa, et al., 2022; Ledford & Wolery, 2013; Preas et al., 2024; Sanetti et al., 2012; Swanson et al., 2011), we found that researchers inconsistently collect/report procedural fidelity data (51.52%; 50.97%). However, the SCARF guidelines go beyond just reporting procedural fidelity and look at the specifics (e.g., if self report only, reporting each condition, at least 20% of overall sessions, obtaining agreement data for procedural fidelity). For example, higher scores on SCARF are coded if direct observation is used to assess fidelity rather than self report. This is relevant given that previous research has demonstrated that implementors overestimate their own performance when using self-report measures (Martino et al., 2009).

Some researchers have evaluated barriers to collecting and reporting procedural fidelity in the fields of behavior analysis (St. Peter et al., 2023), school psychology (Sanetti et al., 2012), and clinical psychology (Perepletchikova et al., 2009). Although a detailed analysis of these studies is beyond the scope of this paper, we will attempt to provide suggestions for mitigating the impact of two barriers (lack of knowledge and lack of resources). For knowledge, many researchers report that they are unaware of the best practices or resources for collecting procedural fidelity (Perepletchikova et al., 2009; St. Peter et al., 2023; Sanetti et al., 2012). This is somewhat unavoidable since comparisons of which procedural fidelity practices are the most accurate/valid are ongoing (e.g., Bergmann et al., 2023). In addition, some researchers may not know what critical independent variables (e.g., delivering reinforcer) to include in the intervention and which arbitrary variables to exclude (e.g., giving the child a pencil; St. Peter et al., 2023). One solution to this problem would be to use procedural fidelity forms from similar research as a model. This could include checking similar published studies for supplemental material, searching open access research repositories, or directly email researchers who have conducted work in that area.

One other frequently documented barrier to collecting procedural fidelity data is the lack of resources. Researchers have reported that the process is time and labor intensive (Sanetti et al., 2012) and inexperienced data collectors may have difficulty balancing interobserver agreement and procedural fidelity at the same time. If possible, the most parsimonious solution would be to record sessions so that when staff are not available, the data can be collected later. In addition, researchers can set up systems for their data collectors such as collecting interobserver agreement and procedural fidelity on alternating days. If possible, having two people collect procedural fidelity for some sessions will allow researchers to report agreement data for procedural fidelity. If researchers are unable to record, they may find it beneficial to prioritize measurement of the most critical independent variables (e.g., picking the 1-2 most important variables). Lastly, when reporting results, researchers should collect and report fidelity for all conditions (not just intervention).

Limitations and Future Directions

Although this review includes a large sample of designs (1,251 designs), this sample may not be representative of the population due to sampling variation (i.e., the questions addressed in various SCARF reviews represent only the interests of those authors and not the entirety of research using SCRD). However, the sample in this analysis was split into two different data sets (those coded with SCARF 1.0 and 2.0), and both data sets yielded similar results for the majority of items. In addition, since SCARF focuses on features unique to SCRD, it is not possible to make comparisons between the quality of SCRD and group design. Thus, although this review shows that there are many areas in which SCRD researchers can improve the rigor and quality of their studies, inherent differences limit drawing conclusions about the comparative quality of group and SCRD. We also relied on original researcher coding of SCARF data; it is possible that variation in coding decisions exists across teams. For example, if researchers included design specific inclusion criteria (e.g., must have at least 5 data points), relevant SCARF items are likely to be biased.

Some of the SCARF items resulted in very different scores from those reviews that used SCARF 1.0 versus 2.0 (e.g., setting described). This calls into question the concurrent validity of these assessments. Given that the item differences between SCARF 1.0 and 2.0 are minimal (see Supplemental Materials 2), it is unlikely that these minor changes caused this difference in scores. It is also unlikely that this difference is due to research practices changing over time since the release date of SCARF has no effect on the data range of studies included in reviews. One other explanation is that the difference in scores is due to natural sampling variability. One way to systematically evaluate the cause of the variability between SCARF versions would be to systematically code SCARF 1.0 and 2.0 using the same set of articles. This process may be necessary for future researchers as since the completion of this project, SCARF now has a 3.0 version.

Although SCARF does include quality indicators around generalization and maintenance, we elected not to include those in this review. First, the generalization and maintenance quality indicators are optional components of SCARF (i.e., researchers only use those indicators in their review if they have a specific research question surrounding generalization or maintenance). Not all of the SCARF articles in our review had research questions about generalization and/or maintenance, leaving us with a smaller sample size for those indicators. Second, those SCARF articles that did include generalization and/or maintenance indicators could only do so for cases in which maintenance or generalization occurred. For example, in a review of 27 articles only four may have included maintenance data. This left us with an even smaller sample size for the maintenance and generalization indicators, making it difficult to draw conclusions about the quality and rigor of maintenance and generalization evaluations as a whole. Our recommendation is that researchers revisit the literature once more review papers are published using SCARF so that a more adequate sample can better answer questions about the quality and rigor of generalization and maintenance evaluations.

This study utilized secondary data analysis methods, which has its own unique sets of strengths and weaknesses. One strength is the large sample of data we were able to evaluate; realistically, our research team would not have been able to code the quality/rigor of 1,251 designs to be able to conduct this largescale analysis. Using data that had already been collected allowed for an aggregation of information across journals, topics, and populations. The major limitation of this secondary data analysis is that not all researchers who used SCARF to code quality and rigor of designs used the same labeling systems. For example, some researchers only included the first author's name as the case identifier (e.g., Skinner), while other researchers used multiple

descriptors (e.g., Skinner, 2024, unpublished, ABAB). Of course, this is not a limitation to answering the research questions of this study (i.e., what is the average rigor/quality score; what are common gaps in SCRD).

However, since our team was unable to assign categorical descriptors to each paper (e.g., published/unpublished, year conducted, design used, etc.), this prevented us from answering other interesting research questions (e.g., have rigor/quality scores changed over time?). Our research team would be unable to locate this categorical information for all 1251 designs, leaving us unable to make these types of comparisons. We hope that future researchers may employ other data collection methods that allow for answering these other research questions. For example, researchers could code a sample of papers from 2004 and 2024 and directly make a comparison on the change in quality/rigor scores over time."

Although our review verified the adequate scope of the review by conducting the same with another database (i.e., PsycInfo), we still only used one database and did not include an ancestral or forward review of the articles. Thus, this review is not comprehensive and may not have included all reviews that included SCARF. However, we believe that this sample of 1,251 cases is sufficient for drawing conclusions. Last, to our knowledge, no research has been conducted on SCARF itself. Thus, although SCARF is commonly used, this tool may not be adequate as a dependent variable in this review. We suggest that future researchers evaluate the technical adequacy of SCARF through psychometric evaluations and obtaining field consensus on which indicators are critical to SCRD.

Conclusions

The results of this review suggest that, by and large, researchers that employ SCRD methodology to answer their research questions have many areas of improvement. Of course, this statement does not place blame on those researchers as they (we) are all responding to the environmental conditions presented to them (us). Journal page limits, limited resources, and lack of training may contribute to these low factors. In this paper, we offer specific strategies for mitigating the impact of these barriers. We hope that if someone conducts a similar review in 5 years, the quality and rigor of these new studies will be higher. It would be erroneous to assume that because there is room for the improvement and quality of SCRD studies, those data are invalid. SCRD is still a valid research mechanism for answering a variety of research questions.

References

**Met inclusion criteria and included in the review.

*Met inclusion criteria but unable to access data.

Barnett, D., Hawkins, R., McCoy, D., Wahl, E., Shier, A., Denune, H., & Kimener, L. (2014). Methods used to document procedural fidelity in school-based intervention research. *Journal of Behavioral Education, 23*(1), 89–107. https://doi.org/10.1007/s10864-013-9188-y

Barton, E.E., Meadan-Kaplansky, H., & Ledford, J.R. (2018). Independent Variables, Fidelity, and Social Validity. In J.R. Ledford & D.L. Gast (Eds.), *Single case research methodology* (3rd ed., pp. 133-156). Routledge.

**Barton, E. E., Murray, R., O'Flaherty, C., Sweeney, E. M., & Gossett, S. (2020). Teaching object play to young children with disabilities: A systematic review of methods and rigor. *American Journal on Intellectual and Developmental Disabilities, 125*(1), 14–36. https://doi.org/10.1352/1944-7558-125.1.14

Barton, E.E., Meadan-Kaplansky, H., & Ledford, J.R. (2018). Research Approaches in Applied Settings. In J.R. Ledford & D.L. Gast (Eds.), *Single case research methodology* (3rd ed., pp. 133-156). Routledge.

Bergmann, S., Niland, H., Gavidia, V. L., Strum, M. D., & Harman, M. J. (2023). Comparing multiple methods to measure procedural fidelity of discrete-trial instruction. *Education & Treatment of Children*, *46*, 201-220. https://doi.org/10.1007/s43494-023-00094-w

Chazin, K. T., Ledford, J. R., Barton, E. E., & Osborne, K. C. (2018). The effects of antecedent exercise on engagement during large group activities for young children. *Remedial and Special Education, 39*(3), 158–170. https://doi.org/10.1177/0741932517716899

**Chazin, K. T., Ledford, J. R., & Pak, N. S. (2021). A systematic review of augmented input interventions and exploratory analysis of moderators. *American Journal of Speech-Language Pathology, 30*(3), 1210–1223. https://doi.org/10.1044/2020_AJSLP-20-00102

**Chazin, K. T., Velez, M. S., & Ledford, J. R. (2022). Reducing escape without escape extinction: A systematic review and meta-analysis of escape-based interventions. *Journal of Behavioral Education, 31*(1), 186–215. https://doi.org/10.1007/s10864-021-09453-2

Creswell, J. W., & Creswell, J. D. (2023). Chapter 1 the selection of a research approach. In *Research design: Qualitative, quantitative, and mixed methods approaches* (6th ed., pp. 40–61). SAGE Publications, Inc.

Elliott, T.C., Morgan, G. & Ayres, K.M. (2024). Techincal Adequacy of the Resaerch Informed Classroom Evlauation - Play (RICE-P) Tool for Evaluating Special Education Teachers. [Manuscript in preparation].

Ennis, R. P., Lane, K. L., & Flemming, S. C. (2021). Empowering teachers with low-intensity strategies: Supporting students at-risk for EBD with instructional choice during reading. *Exceptionality, 29*(1), 61–79. https://doi.org/10.1080/09362835.2020.1729766

Ennis, R. P., Jolivette, K., Fredrick, L. D., & Alberto, P. A. (2013). Using comparison peers as an objective measure of social validity: Recommendations for researchers. *Focus on Autism and Other Developmental Disabilities, 28*(4), 195–201. https://doi.org/10.1177/1088357612475078

**Eyler, P. B., & Ledford, J. R. (2023). Systematic review of time delay instruction for teaching young children. *Journal of Early Intervention*. Advanced online publication. https://doi.org/10.1177/10538151231179121

Falakfarsa, G., Brand, D., Jones, L., Godinez, E. S., Richardson, D. C., Hanson, R. J., Velazquez, S. D., & Wills, C. (2022). Treatment integrity reporting in behavior analysis in practice 2008–2019. *Behavior Analysis in Practice, 15*(2), 443–453. https://doi.org/10.1007/s40617-021-00573-9

Gast, D.L. & Ledford, J.R. (2018). Research Approaches in Applied Settings. In J.R. Ledford & D.L. Gast (Eds.), *Single case research methodology* (3rd ed., pp. 1-26). Routledge.

Ganz, J. B., & Ayres, K. M. (2018). Methodological standards in single-case experimental design: Raising the bar. *Research in Developmental Disabilities, 79*, 3–9. https://doi.org/10.1016/j.ridd.2018.03.003

**Gibbs, A. R., & Tullis, C. A. (2021). The emergence of untrained relations in individuals with autism and other intellectual and developmental disabilities: A systematic review of the recent literature. *Review Journal of Autism and Developmental Disorders, 8*(2), 213–238. https://doi.org/10.1007/s40489-020-00211-0

**Gibbs, A. R., Tullis, C. A., Conine, D. E., & Fulton, A. A. (2024). A systematic review of derived relational responding beyond coordination in individuals with autism and intellectual and developmental disabilities. *Journal of Developmental and Physical Disabilities*, *36*, 1–36. https://doi.org/10.1007/s10882-023-09901-z

Gusenbauer, M. (2019). Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databas*es. Scientometrics, 118*(1), 177–214. https://doi.org/10.1007/s11192-018-2958-5

Han, J. B., Bergmann, S., Brand, D., Wallace, M. D., St. Peter, C. C., Feng, J., & Long, B. P. (2023). Trends in reporting procedural integrity: A comparison. *Behavior Analysis in Practice, 16*(2), 388–398. https://doi.org/10.1007/s40617-022-00741-5

Hardy, J. K., McLeod, R. H., Sweigart, C. A., & Landrum, T. (2022). Comparing and contrasting quality frameworks using research on high-probability requests with young children. *Infants and Young Children, 35*(4), 267–284. https://doi.org/10.1097/iyc.0000000000000223

Haddaway, N. R., Page, M. J., Pritchard, C. C., & McGuinness, L. A. (2022). PRISMA2020: An R package and shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and open synthesis. Campbell Systematic Reviews, *18*(2), e1230. https://doi.org/10.1002/cl2.1230

Hariton, E., & Locascio, J. J. (2018). Randomised controlled trials – the gold standard for effectiveness research: Study design: randomised controlled trials. *BJOG: An International Journal of Obstetrics and Gynaecology, 125*(13), 1716–1716. https://doi.org/10.1111/1471-0528.15199

**Herrod, J. L., Snyder, S. K., Hart, J. B., Frantz, S. J., & Ayres, K. M. (2023). Applications of the Premack principle: A review of the literature. *Behavior Modification, 47*(1), 219–246. https://doi.org/10.1177/01454455221085249

Holcombe, A., Wolery, M., & Snyder, E. (1994). Effects of two levels of procedural fidelity with constant time delay on children's learning. *Journal of Behavioral Education, 4*(1), 49–73. https://doi.org/10.1007/bf01560509

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children, 71*(2), 165–179. https://doi.org/10.1177/001440290507100203

Howland, J. L., Wright, T. C., Boughan, R. A., & Roberts, B. C. (2009). How scholarly is Google Scholar? A comparison to library databases. *College and Research Libraries*, *70*(3), 227–234. https://doi.org/10.5860/0700227

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

Lane, J. D., Ledford, J. R., & Gast, D. L. (2017). Single-case experimental design: Current standards and applications in occupational therapy. *The American Journal of Occupational Therapy, 71*(2), 7102300010p1-7102300010p9. https://doi.org/10.5014/ajot.2017.022210

Lane, K., Wolery, M., Reichow, B., & Rogers, L. (2007). Describing baseline conditions: Suggestions for study reports. *Journal of Behavioral Education, 16*(3), 224–234. https://doi.org/10.1007/s10864-006-9036-4

Ledford, J. R., & Wolery, M. (2013). Procedural fidelity: An analysis of measurement and reporting practices. *Journal of Early Intervention, 35*(2), 173–193. https://doi.org/10.1177/1053815113515908

Ledford, J. R., Lane, J. D., Zimmerman, K. N., Chazin, K. T., & Ayres, K. A. (2016/2023). *Single case analysis and review framework* (SCARF). http://ebip.vkcsites.org/scarf/

Ledford, J. R., Lambert, J. M., Pustejovsky, J. E., Zimmerman, K. N., Hollins, N., & Barton, E. E. (2023). Single-case-design research in special education: Next-generation guidelines and considerations. *Exceptional Children, 89*(4), 379–396. https://doi.org/10.1177/00144029221137656

**Ledford, J. R., & Windsor, S. A. (2022). Systematic review of interventions designed to teach imitation to young children with disabilities. *Topics in Early Childhood Special Education, 42*(2), 202–214. https://doi.org/10.1177/02711214211007190

Ledford, J.R., Lane, J.D., Tate, R. (2018). Evaluating Quality and Rigor in Single Case Research. In J.R. Ledford & D.L. Gast (Eds.), *Single case research methodology* (3rd ed., pp. 365-392). Routledge.

**Ledford, J. R., & Pustejovsky, J. E. (2023). Systematic review and meta-analysis of stay-play-talk interventions for improving social behaviors of young children. *Journal of Positive Behavior Interventions, 25*(1), 65–77. https://doi.org/10.1177/1098300720983521

**Ledford, J. R., Trump, C., Chazin, K. T., Windsor, S. A., Eyler, P. B., & Wunderlich, K. (2023). Systematic review of interruption and redirection procedures for autistic individuals. *Behavioral Interventions, 38*(1), 198-218. https://doi.org/10.1002/bin.1905

Ma, Z., & Travers, J. C. (2022). Adjusting intervention intensity to support students with autism spectrum disorder. *Intervention in School and Clinic, 57*(5), 291–297. https://doi.org/10.1177/10534512211032596

Maggin, D. M., O'Keeffe, B. V., & Johnson, A. H. (2011). A quantitative synthesis of methodology in the meta-analysis of single-subject research for students with disabilities: 1985–2009. *Exceptionality, 19*(2), 109–135. https://doi.org/10.1080/09362835.2011.565725

McDonnell, J., & O'Neill, R. (2003). A perspective on single/within subject research methods and "scientifically based research." *Research and Practice for Persons with Severe Disabilities, 28*(3), 138–142. https://doi.org/10.2511/rpsd.28.3.138

Morgan, D. L., & Morgan, R. K. (2012). *Single-case research methods for the behavioral and health sciences*. SAGE Publications.

*Nanda, S. (2021). Peer-mediated instruction and interventions to increase social skills of children with ASD without intellectual impairments in inclusive preschool and elementary school settings: A meta-analysis [Doctoral dissertation, University of Washington]. ResearchWorks Archive. http://hdl.handle.net/1773/48012

Perepletchikova, F., Hilt, L. M., Chereji, E., & Kazdin, A. E. (2009). Barriers to implementing treatment integrity procedures: survey of treatment outcome researchers. *Journal of Consulting and Clinical Psychology, 77*(2), 212–218. https://doi.org/10.1037/a0015232

Petursdottir, A. I., & Carr, J. E. (2018). Applying the taxonomy of validity threats from mainstream research design to single-case experiments in applied behavior analysis. *Behavior Analysis in Practice, 11*(3), 228–240. https://doi.org/10.1007/s40617-018-00294-6

Preas, E. J., Halbur, M. E., & Carroll, R. A. (2024). Procedural fidelity reporting in the analysis of verbal behavior from 2007–2021. *The Analysis of Verbal Behavior, 40*, 1-12. https://doi.org/10.1007/s40616-023-00197-w

Privitera, G. J. (2020). *Research Methods for the Behavioral Sciences* (3rd Edition). Sage Publications.

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

*Rubio, E. K., McMahon, M. X. H., & Volkert, V. M. (2021). A systematic review of physical guidance procedures as an open-mouth prompt to increase acceptance for children with pediatric feeding disorders. *Journal of Applied Behavior Analysis, 54*(1), 144–167. https://doi.org/10.1002/jaba.782

*Rubio, E.K. (2021). Evaluation of intensive behavioral interventions for severe avoidant/restrictive food intake disorders in young children [Doctoral dissertation, Georgia State University]. ScholarWorks. https://doi.org/10.57709/24097625

**Scheibel, G., Chen, P.-Y., Zaeske, L. M., Wills, H. P., & Zimmerman, K. N. (2023). Improving implementation fidelity with teacher-directed self-monitoring interventions: A systematic review. *Journal of Positive Behavior Interventions, 25*(4), 253–269. https://doi.org/10.1177/10983007221137368

**Scott, V. (2022). On the efficacy of treating escape-maintained inappropriate mealtime behaviour with and without escape extinction: A meta-analysis of escape-based interventions [Doctoral dissertation, Brock University]. DSpace Software. http://hdl.handle.net/10464/17040

Shepley, C., Zimmerman, K. N., & Ayres, K. M. (2021). Estimating the impact of design standards on the rigor of a subset of single-case research. *Journal of Disability Policy Studies, 32*(2), 108–118. https://doi.org/10.1177/1044207320934048

**Snyder, S. K. (2023). A comparison of outcomes of indirect, direct and functional analysis assessment methods and implications for classroom based treatment for challenging behavior [Doctoral dissertation, University of Georgia]. Exlibris.

St. Peter, C. C., Brand, D., Jones, S. H., Wolgemuth, J. R., & Lipien, L. (2023). On a persisting curious double standard in behavior analysis: Behavioral scholars' perspectives on procedural fidelity. *Journal of Applied Behavior Analysis, 56*(2), 336–351. https://doi.org/10.1002/jaba.974

Sanetti, L. M. H., Dobey, L. M., & Gritter, K. L. (2012). Treatment integrity of interventions with children in the Journal of Positive Behavior Interventions from 1999 to 2009. *Journal of Positive Behavior Interventions, 14*(1), 29–46. https://doi.org/10.1177/1098300711405853

Snodgrass, M. R., Chung, M. Y., Kretzer, J. M., & Biggs, E. E. (2022). Rigorous assessment of social validity: A scoping review of a 40-year conversation. *Remedial and Special Education, 43*(2), 114–130. https://doi.org/10.1177/07419325211017295

Swanson, E., Wanzek, J., Haring, C., Ciullo, S., & McCulley, L. (2013). Intervention fidelity in special and general education research journals. *The Journal of Special Education, 47*(1), 3–13. https://doi.org/10.1177/0022466911419516

Tanious, R., & Onghena, P. (2019). Randomized single-case experimental designs in healthcare research: What, why, and how? *Healthcare, 7*(4), 143. https://doi.org/10.3390/healthcare7040143

Tate, R. L., Perdices, M., McDonald, S., Togher, L., & Rosenkoetter, U. (2014). The design, conduct and report of single-case research: Resources to improve the quality of the neurorehabilitation literature. *Neuropsychological Rehabilitation, 24*(3–4), 315–331. https://doi.org/10.1080/09602011.2013.875043

Tate, R. L., Perdices, M., Rosenkoetter, U., Shadish, W., Vohra, S., Barlow, D. H., Horner, R., Kazdin, A., Kratochwill, T., McDonald, S., Sampson, M., Shamseer, L., Togher, L., Albin, R., Backman, C., Douglas, J., Evans, J. J., Gast, D., Manolov, R., … Wilson, B. (2016). The single-Case Reporting guideline In BEhavioural interventions (SCRIBE) 2016 Statement. *Physical Therapy, 96*(7), e1–e10. https://doi.org/10.2522/ptj.2016.96.7.e1

Wellons, Q. D., Roach, A. T., & Sanchez-Alvarez, S. (2023). Is social validity an afterthought in single-case design studies in school psychology research? *Contemporary School Psychology*. Advanced online publication. https://doi.org/10.1007/s40688-023-00460-w

What Works Clearinghouse. (2020). What Works Clearinghouse Standards Handbook, Version 4.1. Washington, DC: U.S. *Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance*. https://ies.ed.gov/ncee/wwc/handbooks

Wolery, M. (1994). Procedural fidelity: A reminder of its functions. *Journal of Behavioral Education, 4*(4), 381–386. https://doi.org/10.1007/bf01539539

Wolery, M., Dunlap, G., & Ledford, J. R. (2011). Single-case experimental methods: Suggestions for reporting. *Journal of Early Intervention, 33*(2), 103–109. https://doi.org/10.1177/1053815111418235

Zimmerman, K. N., Ledford, J. R., Gagnon, K. L., & Martin, J. L. (2020). Social stories and visual supports interventions for students at risk for emotional and behavioral disorders. *Behavioral Disorders, 45*(4), 207–223. https://doi.org/10.1177/0198742919874050

**Zimmerman, K. N., & Ledford, J. R. (2017). Beyond ASD: Evidence for the effectiveness of social narratives. *Journal of Early Intervention, 39*(3), 199–217. https://doi.org/10.1177/1053815117709000

*Zimmerman, K. N., Ledford, J. R., Severini, K. E., Pustejovsky, J. E., Barton, E. E., & Lloyd, B. P. (2018). Single-case synthesis tools I: Comparing tools to evaluate SCD quality and rigor. *Research in Developmental Disabilities, 79*, 19–32. https://doi.org/10.1016/j.ridd.2018.02.003