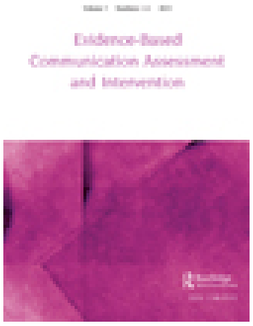
Taylor & Francis
Taylor & Francis Group



Evidence-Based Communication Assessment and Intervention

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tebc20

Evaluating the design and evidence of single-case experimental designs using the What Works Clearinghouse standards

Mariola Moeyaert & Marzieh Dehghan-Chaleshtori

To cite this article: Mariola Moeyaert & Marzieh Dehghan-Chaleshtori (2022): Evaluating the design and evidence of single-case experimental designs using the What Works Clearinghouse standards, Evidence-Based Communication Assessment and Intervention, DOI: 10.1080/17489539.2022.2139148

To link to this article: https://doi.org/10.1080/17489539.2022.2139148

Evidence-Based Communication Assessment and Intervention, 2022
https://doi.org/10.1080/17489539.2022.2139148

Evaluating the design and evidence of single-case experimental designs using the What Works Clearinghouse standards

Mariola Moeyaert & Marzieh Dehghan-Chaleshtori
Department of Educational and Counseling Psychology, University at Albany-SUNY, Albany, NY, USA

Abstract
In 2020, the What Works Clearinghouse (WWC) for single-case experimental design (SCED) studies released an updated version of their standards and procedures handbook (Version 4.1). Because of these updates, there is a need to understand the implications for the field in terms of quality rating of the design, and subsequent synthesis of evidence at the study and meta-analytic level. This study provides a comparison between the previous SCED design and evidence standards, and the updated ones published in 2020 as Version 4.1. We are interested in whether Version 4.1 results in differences in terms of (1) quality rating of the design and (2) analysis and meta-analysis of research evidence. The results indicate no differences related to quality rating of the design, but there are notable differences in terms of how evidence is analyzed, synthesized, and reported. This is further illustrated using a selected publication, namely Pivotal Response Training in which research evidence is meta-analyzed related to the effectiveness of Pivotal Response Training as an intervention to increase communication of children (aged 2–18) with autism. Based on the findings, recommendations and implications are discussed.

Keywords: *What Works Clearinghouse Design and Evidence Standards; Single-Case Experimental Design; Meta-analysis.*

INTRODUCTION

The What Works Clearinghouse (WWC) is an important part of the Institute of Education Sciences (IES) that contributes to evaluating scientific research evidence on education interventions. The aim of the WWC is to systematically identify and review scientific studies, assess the quality of each identified study, and summarize findings of high-quality studies. In that way, the education system can be improved by identifying the "source of scientific evidence for what works in education" (see What Works Clearinghouse, 2020, p. 1). The eligible designs for WWC reviews are randomized controlled trials (RCT), quasi-experimental designs (QED), regression discontinuity designs (RDD), and single-case experimental designs (SCED). In the current study, the focus is on SCEDs. In SCEDs, repeated measurement of the case under

For correspondence: Mariola Moeyaert, School of Education, Department of Educational and Counseling Psychology, Division of Educational Psychology & Methodology, The University at Albany - SUNY, 1400 Washington Ave, Albany, NY 12222. Email: mmoeyaert@albany.edu

Source of funding and declaration of interests: This research was supported by the Institute of Education Sciences, U.S. Department of Education, through grant R305D190022. The content is solely the responsibility of the author and does not necessarily represent the official views of the Institute of Education Sciences, or the U.S. Department of Education.

investigation (i.e. unit, subject, or participant) is the fundamental design tactic. Each case is repeatedly measured before and during (and sometimes after) the intervention. Therefore, the individual cases serve as their own control which is a unique design strength to document the effectiveness of interventions (Horner et al., 2014).

The WWC uses a set of standards to evaluate education research studies and assign them one of the following three ratings: Meets WWC Standards Without Reservations, Meets WWC Standards With Reservations, and Does Not Meet WWC Standards. WWC integrates the results of the SCED studies into systematic reviews. For studies to be eligible to be included in a WWC systematic review, a rigorous and high-quality research methodology needs to be employed. This ensures that the synthesis of research evidence across the studies (i.e. at the meta-analytic level) is reliable and valid. This is important as policy makers, politicians, researchers, clinicians, and practitioners rely on meta-analytic findings to identify effective interventions and programs (i.e. what is working well in education, and what is not working well, and why is that the case?). For this purpose, the WWC developed a set of rigorous design standards, and provided recommendations about how evidence can best be synthesized. To provide a context for the WWC design and evidence standards, an overview of available critical appraisal tools to evaluate the quality of SCED studies in the field of Education and Psychology is provided.

Critical quality appraisal tools for single-case experimental design studies

A literature search was conducted to identify critical appraisal tools suitable to evaluate the methodological quality of SCEDs. We searched the three most used systematic databases within the fields of Education and Psychology, namely PsychINFO, ERIC, and Web of Science using the search term "single-case quality appraisal." In PsychINFO as well as ERIC, the option "anywhere" was selected. For the Web of Science, "All fields" was selected. The search resulted in 14 resources (12 articles and 2 dissertations) in PsychINFO, 4 articles in ERIC, and 34 articles in Web of Science. After removing the duplicates, a total of 36 reports were maintained for full text screening. Twenty articles were excluded because they did not discuss a quality appraisal tool for SCEDs, which resulted in a final selection of 16 reports. Next, we identified the critical appraisal tools to evaluate SCEDs using these reports. An overview of identified tools (together with references to the original reports, and number of citations to the reports), including the WWC SCED standards, is provided in Table 1. The most widely cited tool, across all years, is the Quality Indicators for SSEDs (Horner et al., 2005), closely followed by the WWC SCED. A closer look at the number of citations per year revealed that the number of citations to the WWC SCED exceeds the number of citations to the Quality Indicators for SSED since 2017. Note that the Quality Indicators for SSED was established 5 years prior to the WWC SCED. In sum, taken all together, we can confidently conclude that the WWC SCED is the most used SCED quality appraisal tool. Therefore, it is of crucial importance to understand how the updated 2020 WWC SCED design and evidence standards can be applied to rate the design quality, and to synthesize SCED scientific evidence (at the study and meta-analytic level), and how this is different from previous versions.

What works clearinghouse design and evidence standards

In 2009, the WWC team assembled a panel of recognized experts to draft design and

Table 1. Quality appraisal tools for sceds list from highest number of citations to lowest.

Appraisal tool	Reference[a]	Number of citations[b]
The quality indicators for within single-subjects research	Horner et al. (2005)	4008
What Works Clearinghouse Standards	Kratochwill et al. (2010), Kratochwill et al. (2013)	3267
Kmet appraisal checklist	Kmet et al. (2004)	1410
The Single-Case Experimental Design (SCED) Scale	Tate et al. (2008)	636
Risk of Bias in *N*-of-1Trials (RoBiN-T) scale	Tate et al. (2013) Tate et al. (2015)	
Evaluative Method	Reichow et al. (2008)	471
Single-Case Reporting guideline In BEhavioural interventions (SCRIBE)	Tate et al. (2014) Tate et al. (2016) Tate et al. (2016b)	497
NTACT Standards	Test et al. (2009)	497
CEC Standards for Evidence- Based Practices in Special Education	Cook et al. (2014), Cook et al. (2015)	449
Model for developing, validating, and disseminating interventions for children with autism (ASD)	Smith et al. (2007)	446
Critical Appraisal Skills Programme (CASP) (2014) criteria	Critical Appraisal Skills Programme (CASP) (2014) Singh (2013)	265
Logan Scale/SSRD scale	Logan et al. (2008)	241
Critical appraisal guidelines for single case study research	Atkins and Sampson (2002)	105
American Academy for Cerebral Palsy and Developmental Medicine (AACPDM) (2008) – Conduct Questions for Single Subject Design Studies	American Academy for Cerebral Palsy and Developmental Medicine (AACPDM) (2008)	75
Adapted AACPDM Scale	Smith et al. (2009)	
Certainty Framework	Simeonsson and Bailey (1991)	70
The Single-Case Design Risk of Bias Tool	Reichow et al. (2018)	37
Single Case Analysis and Review Framework (SCARF)	Ledford et al. (2016)	34
Evidence in Augmentative and Alternative Communication Scales (EVIDAAC)	Schlosser et al. (2009) Schlosser (2011)	17
Comparative and Single- Subject Experimental Design Rating Scale (CSSEDARS)		
The Quality Indicator Checklist for Single-Case Research in ASD	Wang and Parrila (2008)	12
Quality Rating Scale for Single Subject Research	Smith et al. (2010)	7
Protocol for assessing single-subject research quality (PASS-RQ)	Maggin and Chafouleas (2010)	5
The Comparative Single-Case Experimental Design Rating System (CSCEDARS, "cedars")	Schlosser et al. (2018)	5
Task Force on Evidence-Based Interventions in School Psychology	Task Force on Evidence-Based Interventions in School Psychology (2003)	0

[a]The full reference is provided in the reference list. [b]Citation search through Google scholar on December 18, 2021.

evidence standards for SCEDs (called Version 1.0 pilot, Kratochwill et al., 2010). Since the establishment of the pilot standards, many SCED researchers relied on this resource to develop their SCED study and evaluate whether their study results provide evidence in support of an effective intervention. The number of citations to this document keeps on increasing over the years. In 2013, Kratochwill and colleagues wrote a research paper, introducing the Version 1.0 pilot WWC standards for SCEDs. Their paper is indicated as a "highly cited paper" using the "InCites Essential Science Indicators" through the Web of Science. As of July 2021, this paper received sufficient citations to be placed in the top 1% of the academic field of social sciences. In 2014, an accompanying document to the Pilot Single-Case Standards (Kratochwill et al., 2010, 2013) known as WWC *Study Review Guide* (SRG) was first published. The SCED study review guide "represents a database where all the relevant aspects of a study are first documented in a systematic manner" (Hitchcock et al., 2015, p. 461). In January 2020, the WWC released an updated standards handbook (Version 4.1) and a procedures handbook (Version 4.1) in which it incorporated the SCED standards published earlier with additions and changes. In March 2021, the latest version of SCED SRG (Version S4 V2) was published and is available to researchers through the WWC website.

Evolution of the what works clearinghouse SCED standards

In September 2011, standards to evaluate SCEDs were added to the WWC standards and procedures handbook (Version 2.1) as an appendix (called SCED Version 1.0 pilot). The pilot SCED standards remained in the appendix section in subsequent versions of the WWC standards, which are Version 3.0 (published in March 2014) and Version 4.0 (published in October 2017). Finally, in 2020, when Version 4.1 of the WWC standards handbook was published, there were significant revisions (i.e. including modifications and extensions) to the SCED standards, along with removal of the "pilot" designation. Removal of the pilot designation also meant that the SCED standards were placed in the body of the handbook, rather than in appendix. One last thing to notice is that Versions 4.0 and 4.1 have separated the standards handbook from the procedures handbook. In prior versions, standards and procedures were included in one single handbook.

Comparison of the what works clearinghouse standards version 1.0 pilot and version 4.1 for single-case designs

The WWC standards for SCEDs contain two sections: (1) evaluating the quality of the design and (2) synthesizing scientific evidence at the study level and the meta-analytic level. Therefore, the comparison of Version 1.0 pilot and Version 4.1 will be discussed according to these two sections.

Critical quality appraisal of the design. There are only a few updates to the Version 1.0 pilot standards. First, Version 4.1 of the WWC standards handbook requires that the raw data be provided in graphical or tabular format to permit readers to perform their own examination of the data. Second, the reviewer guidance around residual treatment effects, multiple probe designs, changing criterion designs, training phases, and more complex variations of basic design tactics were removed from secondary review documents and imbedded into the standards handbook with no changes. Regarding the residual treatment effect, if an intervention is

judged to be likely to have a residual treatment effect, the study is rated *Does Not Meet WWC Standards*. This was added to Version 4.1 of the WWC standards handbook. Residual treatment effect mostly happens in alternating treatment designs when the responses within phases and conditions are caused by interventions in previous phases and conditions. As stated in Version 4.1 of the WWC standards handbook, "the reversal or withdrawal (AB) design standards should be applied to changing criterion designs" (p. 80). A set of additional criteria were added and must be met for multiple probe; otherwise, they receive the rating of *Does Not Meet WWC Standards*: "initial preintervention data collection sessions must overlap vertically, probe points must be available prior to introducing the independent variable, and each case not receiving the intervention must have a probe point in a session where another case either first receives the intervention or reaches the prespecified intervention criterion" (What Works Clearinghouse, 2020b, p. 81).

Synthesizing research evidence. Research evidence can be synthesized at the primary or study level and at the meta-analytic level. The changes to the WWC standards handbook regarding these levels of synthesis are described in this section.

Primary study level. First, a new procedure for quantifying the intervention effect has been added to the WWC standards handbook Version 4.1. This effect size is known as the Design-Comparable Effect Size (D-CES), also called the between-case standardized mean difference (BC-SMD) (Pustejovsky et al., 2014). This effect size is assumed to be on the same scale as Hedges' g effect size from group design studies. Therefore, the D-CES for SCEDs can potentially be combined with Hedges' g from group design studies in one meta-analysis. First, we introduce the D-CES for SCED studies, followed by Hedges' g for group design studies.

The D-CES requires at least three individual cases within an SCED study, and is only available for certain SCED types, namely treatment reversal, multiple baseline, and multiple probe designs. The effect size can be estimated using a two-level hierarchical linear model that accommodates the nested structure of the SCED data. The D-CES parameter (i.e. δ) has two components: (a) the numerator representing the unstandardized intervention effect across cases (i.e. θ_1) and the denominator representing the standardizing component. The standardizing factor captures both the within-case (i.e. σ) and the between-case (τ) standard deviation, see Equation 1.

$$\delta = \frac{\theta_1}{\sqrt{\sigma^2 + \tau^2}} \tag{1}$$

The estimated D-CES is biased in conditions containing a relatively small number of participants, which is traditionally the case in SCED studies. Therefore, it is recommended to apply Hedges' small sample size bias correction (Hedges, 1981; Pustejovsky et al., 2014) which is obtained by multiplying the estimated D-CES (i.e. $\hat{\delta}$) with the factor $(1 - \frac{3}{4v-1})$:

$$\hat{\delta}\_c = \left(1 - \frac{3}{4v - 1}\right) \times \hat{\delta}, \tag{2}$$

with v referring to the estimated degrees of freedom, $\hat{\delta}$ to the estimated D-CES before correction, and $\hat{\delta}\_c$ to the estimated bias-corrected D-CES. The correction factor also needs to be applied to the estimated standard error of $\hat{\delta}$. The *scdhlm* Shiny App tool (Version 0.5.2; Pustejovsky et al., 2021) provides the estimated D-CES and standard error corrected for small sample bias. All

estimated parameters are obtained based on the restricted maximum likelihood estimation procedure. The basic modeling approach to estimate D-CES assumes changes in level and constant intervention effects across cases. A more general and complex approach to estimate the D-CES allows for the intervention effect to vary across cases and it can include linear/polynomial time trends (see Pustejovsky et al., 2014).

For group design studies, the effect size can be obtained by calculating the difference between the mean outcome for the intervention group (y_i) and the mean outcome for the comparison group (y_c) divided by the pooled within-group standard deviation of the outcome measure. Suppose *ni* and *nc* indicate the sample sizes of the intervention and control condition respectively; and *si* and *sc* the standard deviations of the outcome data in both conditions, then the formula for Cohen's d is as follow:

$$d = \frac{(y_i - y_c)}{\sqrt{\frac{(n_i-1)s_i^2+(n_c-1)s_c^2}{n_i+n_c-2}}} \quad (3)$$

And the standard error for Cohen's d is as follow:

$$\mathrm{SE}(d) = \sqrt{\frac{n_i + n_c}{n_i n_c} + \frac{d^2}{2(n_i + n_c)}} \quad (4)$$

Like the D-CES, Cohen's d is biased when there are a small number of participants included in the group design study. Therefore, it is recommended to apply Hedges' small sample size correction factor, which can be obtained by multiplying Cohen's d with the factor (1- $\frac{3}{4N-9}$):

$$g = d * 1 - \frac{3}{4N - 9} \text{ (with N} = ni + nc) \quad (5)$$

Subsequently, the standard error of Cohen's d also needs to be corrected for small sample size bias. Unlike the D-CES, Cohen's d and SE(d) can be calculated directly using closed-form equations (instead of using the restricted maximum likelihood estimation procedure).

Meta-analytic level. The first notable change to the WWC Standards and Procedures (Version 4.1) is the removal of the 5 March 2020 rule. Previously, at least five SCED studies conducted by at least three independent research teams and at least 20 cases across all studies were required to synthesize the effectiveness of the intervention across studies. Second, in Version 4.1, a fixed-effects meta-analytic approach is recommended to synthesize effect sizes across studies. Previously, vote counting procedures were suggested, which involves tallying the significance of primary-level evidence. Using vote counting, the significance at the meta-analytic level is concluded if the majority of primary-level study results are reporting significant findings. This ignores completely the size of the intervention effect at the primary level. When there is no information provided at the primary level to calculate effect sizes (i.e. the raw data in tabular or graphical format is missing and cannot be retrieved), vote-counting procedures can come in handy. Given the new WWC Procedures, WWC reviews do not rely anymore on vote-counting procedures as all studies eligible for inclusion in the synthesis must provide the raw data either in tabular or graphical format. In the recommended fixed-effects meta-analytic approach, effect sizes are weighted by the inverse of its variance. A larger study results in a smaller variance, which in turn results in a larger weight. Consequently, larger studies are given a larger weight in the meta-analysis and have a larger impact on the average effect size estimate across studies. The underlying assumption for

using a fixed-effect meta-analysis is that the primary studies are sampled from one population (given the inclusion criteria). Deviations of effect sizes between studies are not systematic and are due to random sampling variability (Borenstein et al., 2010; Deeks et al., 2019). A last, and significant, change is the removal of the visual analysis requirement. Previously, visual analysis was required prior to calculating effect sizes and synthesizing effect sizes across studies. Primary studies rated as "no evidence in support of intervention effectiveness," based on the visual analysis, were not considered for effect size calculation. This has been mentioned in the WWC procedures handbook (Version 4.1, p.4) as "the WWC no longer uses visual analysis to characterize SCD findings." This has considerable implications as all studies are now eligible for inclusion in the meta-analysis, independent of visual analysis findings.

Aims of current study

The aim of this study is two-folded. First, we compare the old WWC standards (Version 1.0 pilot, which is the same as Version 3.0) to the new version of the WWC standards (Version 4.1) in terms of the critical appraisal of the design. For this purpose, we apply the standards to a previously published WWC intervention report in which Version 3.0 of the standards was applied. We evaluate whether the quality appraisal of SCED studies included in the intervention report receive different ratings using the new standards compared to what is reported in the intervention study. Second, we compare how evidence related to intervention effectiveness is evaluated under both versions, both at the primary study level (i.e. for individual SCED studies) and at the meta-analytic level (i.e. synthesis across studies). In sum, we compare (1) the critical quality appraisal of the design and (2) how scientific evidence is synthesized at the primary and meta-analytic level. Our hypothesis is that there will be no major differences between the old and the new standards in terms of the quality appraisal of the design, but that there will be differences in how evidence is synthesized and reported at both the primary and meta-analytic level. Given the important role the IES WWC plays in designing and synthesizing intervention effectiveness in education research, this study is timely. With this study, we aim to further disseminate Version 4.1 of the WWC standards for SCEDs, and further educate the field about how to apply these standards in practice.

Table 2. Number of reports used in empirical demonstration and WWC standards quality rating.

Quality rating	Number of reports
Meet the standards without reservation	3 SCED journal articles 2 group design journal articles
Meet the standards with reservation	1 SCED dissertation
Do not meet the standards	22 SCED journal articles 1 group design journal article 11 SCED dissertations

Table 3. Reasons for not meeting the WWC standards.

Reason not meeting WWC standards	Number of reports
Insufficient data to evaluate the attempts to demonstrate an intervention effect.	22
The eligible outcomes do not meet WWC requirements.	7
The measures of effectiveness cannot be attributed solely to the intervention.	4

Table 4. Studies meeting WWC standards version 3.0 in PRT intervention report.

Study	Design	Rating	Outcome domain
Feldman and Matos (2012)	Single case (multiple baseline) design	Meet WWC single-case design standards without reservations	Social-emotional Competence
Schreibman et al. (2009)	Single case (multiple baseline) design	Meet WWC single-case design standards without reservations	Communication/language competencies (prompted vocalizations and spontaneous vocalizations)
Sherer and Schreibman (2005)	Single case (multiple baseline) design	Meet WWC single-case design standards without reservations	Communication/language competencies (appropriate communication)
Kim's (2015)	Single case (multiple baseline) design	Meet WWC single-case design standards with reservations	Communication/language competencies (responses to peer communication initiations and initiating social communication with peers)
Hardan et al. (2015)	Group Design (randomized controlled trial)	Meet WWC group design standards without reservations	Communication/language competencies (Imitative utterances, Nonverbally prompted utterances, Spontaneous utterances, Unintelligible utterances, Verbally prompted utterances)
Schreibman and Stahmer (2014)	Group Design	Meet WWC group design standards without reservations	Communication/language competencies (Mullen Scales of Early Learning (MSEL), Expressive Language Scale)

METHODOLOGY

Identification and description of intervention report

A literature search was conducted (February 2021) using the WWC Website repository (What Works Clearinghouse, 2021b) to identify intervention reports (i.e. systematic reviews of interventions) published by the WWC. Out of the 598 identified intervention reports, one intervention report was selected to demonstrate the application of the updated WWC standards (Version 4.1) and compare this to the previous version (Version 3.0, which is the same as Version 1.0 pilot). The selected intervention report needed to include research evidence from both SCED and group design studies. The rationale for this is that Version 4.1 recommends combining SCED and group design studies if possible, as this enhances the generalizability of the findings. Because we want to demonstrate this possibility, we only considered an intervention report including both design types. Out of the intervention reports satisfying this criterion, we selected the review that included the largest amount of SCED and group design studies. The selected report, entitled Pivotal Response Training (U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse, 2016), synthesizes research evidence related to the effectiveness of Pivotal Response Training (PRT) as an intervention to help children (aged 2–18) with

autism in areas they are struggling in (i.e. in the domain of communication). The full report can be accessed through: https://ies.ed.gov/ncee/wwc/InterventionReport/668. In this selected intervention report, the extent of scientific evidence related to the influence of PRT on autistic students' communication domain appeared to be small. A total of 40 studies were considered for inclusion, including 37 SCED studies, and three group design studies. For the SCED studies, the 5 March 2020 rule was not satisfied as only three SCED studies met the WWC with or without reservations. Across these three studies there were less than 20 participants (16 participants). Therefore, the evidence from the SCED studies was not synthesized. Only two group design studies met the WWC standards. Evidence for two group design studies was then synthesized in the intervention report, based on data from 85 Autistic children.

Data for the empirical illustration

We re-reviewed the SCED studies included in the WWC's PRT review using the new WWC standards (version 4.1). For the purposes of the paper, we assumed that the results of the group design studies remained the same. The two group design studies that received a rating *Meets WWC Standards With or Without Reservations* will be included in our empirical demonstration. Out of the 40 studies included in the PRT review, 28 are journal articles and the remaining 12 are doctoral dissertations. A summary of the studies with their assigned quality rating by WWC raters (as reported in the PRT review) under Version 3.0 and the major reason for not meeting the standards is provided in Tables 2 and 3. A full list of the studies included in the PRT review can be found in Appendix A. Next, Table 4 shows the list of studies meeting the WWC standards with or without reservations using Version 3.0 as reported by the intervention report. All these studies investigated the outcome domain of communication/language competencies except from Feldman and Matos (2012).

For the second part of current study, we demonstrate the application of the WWC standards for the synthesis of research evidence at the primary and meta-analyst level. For this reason, the five studies (listed in Table 4) investigating the effectiveness of PRT on communication were selected (three SCEDs and two group design studies). We extracted the raw data from the graphs of the three SCED studies meeting WWC standards with or without reservations using the data retrieval program called WebPlotDigitizer (Rohatgi, 2020) as this is needed to synthesize research evidence under the new WWC standards (Version 4.1). For the group design studies, the required information for synthesis was already reported in the PRT study and readily available.

Materials, software, and procedure

We followed the WWC's Standards and Procedures for reviewing studies using the publicly available materials including the procedures handbook (Version 4.1), the standards handbook (Version 4.1 and Version 3.0), and the public version of the WWC Study Review Guide (SRG) for SCED studies (version S4 V2, retrieved from https://ies.ed.gov/ncee/wwc/StudyReviewGuide).

The data retrieval program WebPlotDigitizer (Version 4.4, Rohatgi, 2020) was used to retrieve raw data from SCED graphs, which allows estimation of the D-CES. The point-and-click Shiny tool *scdhlm* (Pustejovsky et al., 2021) was used to estimate the D-CES together with its standard error. The obtained D-CES at the primary level is compared to the results based on the visual analysis in the original PRT intervention report. In the next step, the

Table 5. Single case design findings for the communication/language competence domain copied from PRT intervention (2016).

	Study characteristics			WWC summary
Outcome measure	**Sample size (case)**	**Age**	**Design type**	**Evidence level**
- Kim (2015)				
Responses to peer initiations	4 (all students)	5-10	Multiple baseline across participants	Strong (+)
Social initiations	4 (all students)	5-10	Multiple baseline across participants	No evidence
- Schreibman et al. (2009)				
Cued vocalizations (% of intervals)	3 (HTC subgroup)	2	Multiple baseline across participants	Strong (+)
Cued vocalizations (% of intervals)	3 (LSA subgroup)	2-4	Multiple baseline across participants	No evidence
Spontaneous vocalizations (% of intervals)	3 (HTC subgroup)	2	Multiple baseline across participants	No evidence
Spontaneous vocalizations (% of intervals)	3 (LSA subgroup)	2-4	Multiple baseline across participants	No evidence
- Sherer and Schreibman (2005)				
Appropriate communication (% of intervals)	3 (responder subgroup)	3	Multiple baseline across participants	No evidence
Appropriate communication (% of intervals)	3 (nonresponder subgroup)	3-5	Multiple baseline across participants	No evidence

The WWC does not calculate effect sizes for single-case design (SCD) research. Characterizations of *Strong* and *Moderate* evidence, based on WWC visual analysis, indicate that the experiment demonstrated an effect of the intervention. Characterizations of *No evidence* indicate that the experiment did not provide at least three demonstrations of an intervention effect in the same direction. + = a positive (favorable effect) in the desired direction. The evidence from the SCD studies on PRT does not reach the threshold to include SCD evidence in the effectiveness ratings for the communication/language competencies domain. HTC = high toy contact. LSA= low social avoidance

D-CESs of the primary studies are synthesized across studies using the fixed effects meta-analysis approach. The fixed effects meta-analytic findings are compared with the vote-counting synthesis under the old standards. Lastly, for demonstration purposes, we run a fixed effects meta-analysis combining the D-CES for SCED with Hedges' g for group design studies. The package metafor (Viechtbauer, 2010) in the statistical software program R version 4.2.1 (R Core Team, 2020) was used to run the fixed effects meta-analyses.

RESULTS

Quality appraisal of the design

Forty articles and doctoral dissertations (37 SCEDs and 3 group designs) previously receiving a quality rating by the WWC review team (in 2016, by applying Version 3.0 of WWC standards handbook) were reevaluated using the updated standards Version 4.1. No changes in terms of quality rating were identified for all 37 SCED studies. This is not surprising as changes between Version 3.0 and Version 4.1

Table 6. Effect size, variance, and weight for SCED and group design studies of the PRT intervention report (2016) meeting the WWC standards.

Outcome measure	Effect size	Variance	Weight
Single Case Experimental Design Studies			
- Kim (2015)			
Responses to peer initiations	3.048	.192	5.19
Social initiations	0.114	.054	18.42
- Schreibman et al. (2009)			
Spontaneous vocalizations	−0.157	.022	46.28
Cued vocalization	0.811	.034	29.54
- Sherer and Schreibman (2005)			
Appropriate communication	0.455 .026	38.58	
Group Design Studies			
- Hardan et al. (2015)			
Imitative utterances	0.73	.088	11.34
Nonverbally prompted utterances	1.02	.093	10.75
Spontaneous utterances	0.80	.088	11.26
Unintelligible utterances	−0.18	.083	12.06
Verbally prompted utterances	0.10	.082	12.14
- Schreibman and Stahmer (2014)			
Mullen Scales of Early Learning (MSEL) Expressive Language Scale	−0.22	.102	9.83

include the requirement of displaying raw data (graphically or in tabular format) and are related to additional criteria to evaluate changing criterion designs, alternating treatments designs, and multiple probe designs. The designs in the PRT report that are meeting the standards with or without reservations were all multiple baseline designs. There were a few other design types in the PRT report (i.e. an alternating treatments design and a multiple probe design) that did not meet the WWC standards. Because of the additional criteria for these designs, the rating remained the same. This confirms our hypothesis that no changes in terms of quality appraisal of the design are found.

Synthesizing evidence at the primary level

Under Version 3.0 of the WWC standards, research evidence at the primary level was synthesized using visual analysis. The PRT report provided the results of the visual analysis of the three SCED studies meeting the WWC standards and focused on the communication/language competencies domain. The results of the visual analysis of Kim (2015) indicated strong evidence for *responses to peer initiation* outcome measure while no evidence was detected for *social initiations*. The evidence level for Schreibman et al. (2009) was strong for one outcome measure, namely *cued vocalization*, and only for the high toy contact (HTC) subgroup (no evidence was detected for the low social avoidance [LSA] subgroup). There was no evidence for *spontaneous vocalization*. In Sherer and Schreibman (2005) no evidence was found for the two subgroups looking into *appropriate communication* outcome measure (see Table 5). These results are similar to what we can conclude based on estimating the D-CES (see Table 6). The estimated effect size for *cued vocalization* measure in Schreibman et al.

(2009) equals 0.811 with standard error of 0.184. This reflects a large and statistically significant intervention effect. The calculated effect size for Kim's (2015) *responses to peer initiation* outcome is also large and statistically significant, indicating strong evidence in support of an effective intervention ($\hat{\delta}_c = 3.048, SE = .438$). The effect sizes calculated for other outcome measures in Schreibman et al. (2009), Kim's (2015), and Sherer and Schreibman (2005) are small to moderate (see Table 6).

In group design studies, the intervention was rated as "none of the studies shows a statistically significant or substantively important effect, either positive or negative" (see p. 35 of U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse, 2016, for WWC Rating Criteria used in PRT intervention report). Also, the extent of evidence has been determined as small as the total sample size ($n = 85$) is fewer than 350 students (which is a minimum requirement to have a medium or large extent of evidence in group design studies). The effect sizes as reported in the PRT study for group design studies were readily available which was not the case for the standard errors. Therefore, we calculated the standard errors by using Equation 4 as both the effect size and standard errors (see Table 7) are needed for subsequent synthesis across studies (i.e. meta-analysis, see next section). Note that there are five effect sizes reported in Table 7 for Hardan et al. (2015), as they investigated five different outcome measures.

Synthesizing evidence at the meta-analytic level

Since the three SCED studies meeting the WWC standards with or without reservations did not satisfy the requirements (meeting the threshold 5 March 2020 rule), they were not synthesized at the meta-analytic level in the PRT report. However, in this study, we re-analyzed the data of the SCED studies and could include the results from those studies as the 5 March 2020 rule is no longer applicable under the new standards. The procedure for calculating D-CES using *scdhlm* package is

Table 7. Group design subscale findings for communication/language competencies domain.

		Mean (std)		
	Sample size	**Intervention group**	**Comparison group**	**Hedges' g (SE)**
- Hardan et al. (2015)	47 children ($n_i = 25$, $n_c = 22$)	15.80 (14.20)	7.10 (7.80)	0.73 (0.297)
Imitative utterances				
Nonverbally prompted utterances	47 children ($n_i = 25$, $n_c = 22$)	3.00 (3.80)	0.10 (0.30)	1.02 (0.305)
Spontaneous utterances	47 children ($n_i = 25$, $n_c = 22$)	1.30 (1.40)	0.40 (0.60)	0.80 (0.298)
Unintelligible utterances	47 children ($n_i = 25$, $n_c = 22$)	21.16 (14.90)	24.60 (23.30)	−0.18 (0.288)
Verbally prompted utterances	47 children ($n_i = 25$, $n_c = 22$)	17.70 (15.20)	16.00 (17.00)	0.10 (0.287)
- Schreibman and Stahmer (2014) Mullen Scales of Early Learning (MSEL), Expressive Language Scale	38 children ($n_i = 18$, $n_c = 20$)	25.50 (11.20)	28.70 (16.50)	−0.22 (0.319)

Table 8. Summary of the results obtained from scdhlm package.

Study	Sample size	$\hat{\delta}\_c$ (SE)	Auto-correlation	Outcome
Schreibman et al. (2009)	6	−0.157 (0.147)	−0.003	Spontaneous vocalizations
		0.811 (0.184)	0.168	Cued vocalization
Sherer and Schreibman (2005)	6	0.455 (0.161)	0.745	Appropriate communication
Kim (2015)	4	3.048 (0.439)	0.508	Responses to peer initiations
		0.114 (0.233)	0.756	Social initiations

discussed in a tutorial paper by Valentine et al. (2016) and a video published by IES and available on YouTube titled "design-comparable effect size for single-case design studies" (https://www.youtube.com/watch?v=uXTbL8QkNvY, Institute of Education Sciences, 2020). The procedure, together with text and screenshots, are documented in Appendix B as well. The obtained results can be found in Table 8. Note that the studies by Kim's (2015) and Schreibman et al. (2009) investigated multiple outcome domains. Therefore, multiple D-CESs were calculated for these studies, one per outcome domain. Estimates for the autocorrelation parameter are also provided in Table 8 as this is estimated by default using the scdhlm package. It is recommended to include an autocorrelation parameter when modeling repeated measures as this accounts for serial dependency (Ferron, 2002). Using a fixed effects meta-analysis, the pooled estimated effect size across the three SCED studies (including a total of five effect sizes) is 0.38, with a standard error of 0.085 [$\hat{\delta}\_c = 0.380$, 95% CI (0.213, 0.548), $Z = 4.46$, $p < .0001$]. This is a statistically significant but rather small intervention effect. These meta-analytic results cannot be compared with the results from the original published PRT intervention report as the results in that report were not synthesized due to the violation of the 5 March 2020 rule.

Finally, the D-CES for SCEDs and Hedges' g for group design studies are combined using a fixed-effects meta-analysis. Following WWC instruction on estimating the fixed-effect meta-analytic effect size, study-level effect sizes were weighted by the inverse of their variances (variances and weights for effect sizes for both SCEDs and group design studies can be found in Table 6). The computed meta-analytic effect size across the three SCED studies and two group design studies is .375 with a standard error of .0700. Using Cohen's d metric, the calculated meta-analytic effect size can be considered as small. Cohen (1988, 1992) provided guidelines for the interpretation of these values: values of 0.20, 0.50, and 0.80 for Cohen's d and Hedges' g are commonly considered to be indicative of small, medium, and large effects. Statistical significance of the meta-analytic result was evaluated using a nondirectional Z test and a type I error rate of $\alpha = .05$ as suggested in WWC Standards Handbook (Version 4.1). The results indicate that across the five studies, the overall average effect size of 0.375 (SE = 0.0700) is statistically significant [95% CI (0.24, 0.51), $Z = 5.41$, $p < .0001$]. These meta-analytic results cannot be compared with the results from the original published PRT intervention report as results from SCED studies and group

design studies were not synthesized together.

DISCUSSION

Summary

The WWC SCED design and evidence standards (Kratochwill et al., 2010, 2013; What Works Clearinghouse, 2011, 2014, 2017, 2017b, 2020, 2020b) is a frequently referenced and influential source for designing SCED studies and for evaluating intervention effectiveness. In 2020, the Institute of Education Sciences released updated WWC design and evidence standards. However, little is known about the implications of the updates for the SCED field. This study was designed to provide an overview and understanding of the updates, and to demonstrate how the new design and evidence standards can be applied to 1) rate the design quality and 2) synthesize evidence at the study and meta-analytic level. In terms of rating the design quality, the new WWC SCEDs are more inclusive as quality rating criteria for changing criteria designs and multiple probe designs are also included. This is a significant improvement compared to previous versions of the WWC SCED standards. Previously there were no specific criteria for these designs and therefore these design types could either not be evaluated or inappropriate criteria were applied. Consequently, these design types are typically excluded from meta-analyses. In terms of evaluating and synthesizing intervention effectiveness at the study level (i.e. synthesizing intervention effectiveness across participants from one SCED study), major differences were identified. The old WWC SCEDs required a visual analysis of the graphical display of individual participant SCED data to infer intervention effectiveness. In contrast, the new WWC SCEDs removed this requirement and endorsed the use of the D-CES to estimate intervention effectiveness across participants within one SCED study. However, Kratochwill et al. (2021) argue for the inclusion of visual-analysis to document design standards and evidence criteria. They also provided limitations to the D-CES: the requirement of at least three individuals/cases per study ignores the effect size evidence from studies with only one or two participants, exclusive reliance on the D-CES limits the scope of potential moderator analyses, only applicable for certain SCED design types, and it should not be applicable to context where the construct underlying the outcomes measured in the SCDs is qualitatively different from the construct underlying the outcomes measured in the group comparison studies.

The D-CES is obtained by running a two-level hierarchical linear model (using the restricted maximum likelihood estimation procedure), which is a parametric test. Therefore, a standard error and statistical significance of the D-CES estimate are obtained. Consequently, an effect size together with its statistical significance can be reported to evaluate intervention effectiveness. The new WWC standards recommend using a fixed-effects meta-analysis to combine D-CESs across studies. Therefore, an overall estimate of intervention effectiveness across SCED studies together with its statistical significance can be reported. This was not possible under previous standards, which recommended vote-counting procedures. In addition, the D-CES is assumed to be on the same scale as Cohen's/Hedges'g from group design studies. Therefore, under the WWC SCEDs, it is possible to combine D-CES and Hedges'g using a fixed effects meta-analysis. This results in more research evidence available for synthesis, and more generalizable conclusions. It is also assumed that Cohen's/Hedges' g metric indicative for small,

medium, and large effect sizes can be applied to the combined evidence which helps making substantive interpretations.

The use of the D-CES

Researchers considering using the D-CES to synthesize and meta-analyze research evidence should be aware of some critical issues. The D-CES is estimated using a two-level hierarchical linear modeling approach. This approach, in general, works well when a large set of level two units (i.e. participants) are included (given the asymptotic assumptions), which is commonly not the case for SCED studies (Shadish & Sullivan, 2011). Under small sample sizes, the estimated D-CES might still be unbiased; however, biased standard errors are anticipated resulting in invalid statistical inference. This has implications for meta-analyzing D-CESs as the precision (i.e. inverse of the squared standard error) will be influenced. This can result in a biased estimate of D-CES (assigning an inappropriate weight to some studies) and its corresponding standard error. Therefore, caution is needed when interpretating meta-analytic findings and overly replying on statistical significance needs to be avoided. Next, SCED researchers should also be aware that the D-CES can only be used for reversal/withdrawal and multiple baseline designs. Therefore, certain design types such as alternating treatments designs and changing criterion designs are systematically excluded from the meta-analysis, resulting in a biased representation of the evidence. In addition, the D-CES requires at least three individuals/cases per SCED study. Therefore, SCED studies including one or two individuals are excluded from the analysis. Next, the D-CES summarizes the research evidence at the study level. Consequently, participant moderators cannot be investigated. Therefore, in certain conditions, other additional effect sizes need to be considered and might be more appropriate (Kratochwill et al., 2021). Finally, researchers considering combining D-CES and group design studies make the strong (and perhaps not realistic) assumption that both effect sizes are on the same scale. The two effect sizes use a different standardization factor and consequently might not be comparable and suitable to be meta-analyzed using one fixed-effects model. The D-CES uses an estimate of the within-participant and between-participant standard deviation whereas Cohen's d/Hedges' g uses the pooled within-group standard deviation. Traditionally, values of 0.20, 0.50, and 0.80 for Cohen's d and Hedges' g are indicative of small, medium, and large effects. However, this scale is unlikely applicable for the combined effect size across SCED studies and group design studies. It is recommended to evaluate the magnitude of effect based on different factors such as the type of effect size used, the research context and its practical or clinical value, confidence interval around the point estimate, research design, alignment of the measurement and the intervention, and sample size (Bakker et al., 2019; Durlak, 2009).

Future research

A first suggestion for future research is to further study under which SCED design conditions the use of D-CES is appropriate and recommended. Second, research is needed to study the underlying scale of the D-CES for SCEDs, and whether this is comparable to the underlying scale of Cohen's d/Hedges' g. Finally, it is not only important that primary studies are of high quality, but also meta-analyses of primary studies. Jamshidi et al. (2018) conducted a systematic review in which methodological quality of 178 SCED meta-analyses studies (using the modified *R*-AMSTAR tool) was evaluated. They concluded that the methodological quality of SCED meta-

analyses increased over time but remained at a low level. Therefore, researchers should not only rely on the peer review process but also investigate certain checklists to evaluate the quality of SCED meta-analyses. The WWC design and evidence standards do not provide insights related to the critical quality appraisal for SCED meta-analysis. Research is needed to identify and/or develop valid and reliable tools for the critical appraisal of SCED meta-analyses.

Acknowledgement

The authors sincerely acknowledge Daniel Swan for providing initial feedback to this manuscript.

Funding

This research was supported by the Institute of Education Sciences, U.S. Department of Education, through grant [R305D190022]. The content is solely the responsibility of the author and does not necessarily represent the official views of the Institute of Education Sciences, or the U.S. Department of Education.

DISCLOSURE STATEMENT

The commentary authors report no conflicts on interest and are solely responsible for the content and writing of this commentary.

SUPPLEMENTARY MATERIAL

Supplemental data for this article can be accessed online at https://doi.org/10.1080/17489539.2022.2139148.

ORCID

Mariola Moeyaert http://orcid.org/0000-0003-1453-8162

REFERENCES

American Academy for Cerebral Palsy and Developmental Medicine (AACPDM). (2008). *Methodology to develop systematic reviews of treatment interventions*. https://www.aacpdm.org/UserFiles/file/systematic-review-methodology.pdf.

Atkins, C., & Sampson, J. (2002). Critical appraisal guidelines for single case study research. *ECIS 2002 Proceedings* (pp. 1–15). Poland.

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

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods, 1*(2), 97–111. https://doi.org/10.1002/jrsm.12

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

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Cook, B., Buysse, V., Klingner, J., Landrum, T., McWilliam, R., Tankersley, M., & Test, D. W. (2014). Council for exceptional children: Standards for evidence-based practices in special education. *Teaching Exceptional Children, 80*(4), 504–511. https://doi.org/10.1177/0014402914531388

Cook, B. G., Buysse, V., Klingner, J., Landrum, T. J., McWilliam, R. A., Tankersley, M., & Test, D. W. (2015). CEC's standards for classifying the evidence base of practices in special education. *Remedial and Special Education, 36*(4), 220–234. https://doi.org/10.1177/0741932514557271

Critical Appraisal Skills Programme (CASP). (2014). *CASP Checklists*. http://www.casp-uk.net/casp-tools-checklists

Deeks J. J., Higgins J. P. T., & Altman D. G. (Eds). (2019). Chapter 10: Analysing data and undertaking meta-analyses. In Higgins J. P. T., & Thomas J. (Eds.), *Cochrane Handbook for Systematic Reviews of Interventions* version 6.3. https://www.training.cochrane.org/handbook

Durlak, J. A. (2009). How to select, calculate, and interpret effect sizes. *Journal of Pediatric Psychology, 34*(9), 917–928. https://doi.org/10.1093/jpepsy/jsp004

Feldman, E. K., & Matos, R. (2012). Training paraprofessionals to facilitate social interactions between children with autism and their typically developing

peers. *Journal of Positive Behavior Interventions*, *15*(3), 169–179. https://doi.org/10.1177/1098300712457421

Ferron, J. (2002). Reconsidering the use of the general linear model with single-case data. *Behavior Research Methods, Instruments, & Computers*, *34*(3), 324–331. https://doi.org/10.3758/BF03195459

Hardan, A. Y., Gengoux, G. W., Berquist, K. L., Libove, R. A., Ardel, C. M., Phillips, J., Frazier, T. W., & Minjarez, M. B. (2015). A randomized controlled trial of Pivotal Response Treatment Group for parents of children with autism. *Journal of Child Psychology and Psychiatry*, *56*(8), 884–892. https://doi.org/10.1111/jcpp.12354

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*(2), 107–128. https://doi.org/10.3102/10769986006002107

Hitchcock, J. H., Kratochwill, T. R., & Chezan, L. C. (2015). What works clearinghouse standards and generalization of single-case design evidence. *Journal of Behavioral Education*, *24*(4), 459–469. https://doi.org/10.1007/s10864-015-9224-1

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

Horner, R. H., & Odom, S. L. (Eds.). (2014). Constructing single-case research designs: Logic and options. In Kratochwill, T. R. & Levin, J. R. Eds, *Constructing single-case research designs: Logic and options. Single-case intervention research: Methodological and statistical advances* (pp. 27–51). American Psychological Association https://doi.org/10.1037/14376-002

Jamshidi, L., Heyvaert, M., Declercq, L., Fernández-Castilla, B., Ferron, J. M., Moeyaert, M., Beretvas, S. N., Onghena, P., & Van den Noortgate, W. (2018). Methodological quality of meta-analyses of single-case experimental studies. *Research in Developmental Disabilities*, *79*, 97–115. https://doi.org/10.1016/j.ridd.2017.12.016

Kim, S. (2015). *Implementing a pivotal response social skills intervention with Korean American children with autism*. [ProQuest Dissertations and Theses database] (UMI No. 1680833240)

Kmet, L. M., Lee, R. C., & Cooks, L. S. (2004). *Standard quality assessment criteria for evaluating primary research papers from a variety of fields*. Alberta Heritage Foundation for Medical Research.

Kratochwill, T. R., Hitchcock, J. J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). *Single-case designs technical documentation: Version 1.0 (Pilot)*. Institute of Education Sciences, What Works Clearinghouse.

Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2013). Single-case intervention research design standards. *Remedial and Special Education*, *34*(1), 26–38. https://doi.org/10.1177/0741932512452794

Kratochwill, T. R., Horner, R. H., Levin, J. R., Machalicek, W., Ferron, J., & Johnson, A. (2021). Single-case design standards: An update and proposed upgrades. *Journal of School Psychology*, *89*, 91–105. https://doi.org/10.1016/j.jsp.2021.10.006

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

Logan, L. R., Hickman, R. R., Harris, S. R., & Heriza, C. B. (2008). Single-subject research design: Recommendations for levels of evidence and quality rating. *Developmental Medicine & Child Neurology*, *50*(2), 99–103. https://doi.org/10.1111/j.1469-8749.2007.02005.x

Maggin, D. M., & Chafouleas, S. M. (2010). PASS-RQ: Protocol for assessing single-subject research quality. Unpublished research instrument.

Institute of Education Sciences. [Moeyaert, M., Swan, D. M. & Tanner-Smith, E.]. (2020, December 22). Design-comparable effect size for single case design studies. [video]. https://www.youtube.com/watch?v=uXTbL8QkNvY

Pustejovsky, J. E., Chen, M., & Hamilton, B. (2021). *Scdhlm: A web-based calculator for between-case standardized mean differences (version 0.5.2) Web application*. https://jepusto.shinyapps.io/scdhlm

Pustejovsky, J. E., Hedges, L. V., & Shadish, W. R. (2014). Design-comparable effect sizes in multiple baseline designs: A general modeling framework. *Journal of Educational and Behavioral Statistics*, *39*(5), 368–393. https://doi.org/10.3102/1076998614547577

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

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

Reichow, B., Volkmar, F. R., & Cicchetti, D. V. (2008). Development of the evaluative method for evaluating and determining evidence-based practices in autism. *Journal of Autism and Developmental Disorders*, *38*(7), 1311–1319. https://doi.org/10.1007/s10803-007-0517-7

Rohatgi, A. (2020). *WebPlotdigitizer (version 4.4) Web application*. https://automeris.io/WebPlotDigitizer

Schlosser, R. W. (2011). *EVIDAAC single-subject scale*. http://www.evidaac.com/ratings/Single_Sub_Scale.pdf.

Schlosser, R. W., Belfiore, P. J., Sigafoos, J., Briesch, A. M., & Wendt, O. (2018). Appraisal of comparative single-case experimental designs for instructional interventions with non-reversible target behaviors: Introducing the CSCEDARS ("Cedars"). *Research in Developmental Disabilities*, *79*, 33–52. https://doi.org/10.1016/j.ridd.2018.04.028

Schlosser, R. W., Sigafoos, J., & Belfiore, P. (2009). *EVIDAAC comparative single-subject experimental design scale (CSSEDARS)*. https://www.academia.edu/55680770/EVIDAAC_1_Comparative_Single_Subject_Experimental_Design_Scale_CSSEDARS_.

Schreibman, L., & Stahmer, A. C. (2014). A randomized trial comparison of the effects of verbal and pictorial naturalistic communication strategies on spoken language for young children with autism. *Journal of Autism and Developmental Disorders*, *44*(5), 1244–1251. https://doi.org/10.1007/s10803-013-1972-y

Schreibman, L., Stahmer, A. C., Barlett, V. C., & Dufek, S. (2009). Brief report: Toward refinement of a predictive behavioral profile for treatment outcome in children with autism. *Research in Autism Spectrum Disorders*, *3*(1), 163–172. https://doi.org/10.1016/j.rasd.2008.04.008

Shadish, W. R., & Sullivan, K. J. (2011). Characteristics of single-case designs used to assess intervention effects in 2008. *Behavior Research Methods*, *43*(4), 971–980. https://doi.org/10.3758/s13428-011-0111-y

Sherer, M. R., & Schreibman, L. (2005). Individual behavioral pro les and predictors of treatment effectiveness for children with autism. *Journal of Consulting and Clinical Psychology*, *73*(3), 525–538. https://doi.org/10.1037/0022-006X.73.3.525

Simeonsson, R. J., & Bailey, D. B. (1991). Evaluating programme impact: Levels of certainty. In Mitchell, D.M., Brown, R.I. (Eds.). *Early intervention studies for young children with special needs* (pp. 280–296). Rehabilitation Education: Springer. https://doi.org/10.1007/978-1-4899-3292-1_11

Singh, J. (2013). Critical appraisal skills programme. *Journal of Pharmacology & Pharmacotherapeutics*, *4*(1), 76. https://doi.org/10.4103/0976-500X.107697

Smith, V., Jelen, M., & Patterson, S. (2009). Video modeling to improve play skills in a child with autism: A procedure to examine single-subject experimental research. *Evidence-Based Practice Briefs*, *4*(1), 1–13.

Smith, V., Jelen, M., & Patterson, S. (2010). Video modeling to improve play skills in a child with autism: A procedure to examine single-subject experimental research. *Evidence-Based Practice Briefs*, *4*, 1–11.

Smith, T., Scahill, L., Dawson, G., Guthrie, D., Lord, C., Odom, S., Rogers, S., & Wagner, A. (2007). Designing research studies on psychosocial interventions in autism. *Journal of Autism and Developmental Disorders*, *37*(2), 354–366. https://doi.org/10.1007/s10803-006-0173-3

Task Force on Evidence-Based Interventions in School Psychology. (2003). *Procedural and coding manual for review of evidence-based interventions*. http://www.indiana.edu/ ebi/documents/_workingfiles/EBImanual1.pdf

Tate, R., McDonald, S., Perdices, M., Togher, L., Schultz, R., & Savage, S. (2008). Rating the methodological quality of single-subject designs and n-of-1 trials: Introducing the single-case experimental design (SCED) scale. *Neuropsychological Rehabilitation*, *18*(4), 385–401. https://doi.org/10.1080/09602010802009201

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., McDonald, S., Togher, L., Shadish, W., Horner, R., Kratochwill, T., Barlow, D. H., Kazdin, A., Sampson, M., Shamseer, L., & Vohra, S. (2016). The single-case reporting guideline in behavioural interventions (SCRIBE) 2016: Explanation and elaboration. *Archives of Scientific Psychology*, *4*(1), 10. https://doi.org/10.1037/arc0000027

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. (2016b). The single-case reporting guideline in behavioural interventions (SCRIBE) 2016 statement. *Physical Therapy*, *96*(7), e1–10. https://doi.org/10.2522/ptj.2016.96.7.e1

Tate, R. L., Perdices, M., Rosenkoetter, U., Wakim, D., Godbee, K., Togher, L., & McDonald, S. (2013). Revision of a method quality rating scale for single-case experimental designs and n-of-1 trials: The 15-item risk of bias in N-of-1 trials (RoBint) scale. *Neuropsychological Rehabilitation*, *23*(5), 619–638. https://doi.org/10.1080/09602011.2013.824383

Tate, R. L., Rosenkoetter, U., Wakim, D., Sigmundsdottir, L., Doubleday, J., Togher, L., McDonald, S., & Perdices, M. (2015). *The risk of bias in N-of-1 trials (RoBint) scale: An expanded manual for the critical appraisal of single-case reports*. Author.

Test, D. W., Fowler, C. H., Richter, S. M., White, J., Mazzotti, V., Walker, A. R., Kortering, L. ... Kortering, L. (2009). Evidence based practices in secondary transition. *Career Development for Exceptional Individuals*, *32*(2), 115–128. https://doi.org/10.1177/0885728809336859

U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse. (2016, December). *Children and Students with an Autism Spectrum Disorder intervention report: Pivotal Response Training*. http://whatworks.ed.gov

Valentine, J. C., Tanner-smith, E. E., Pustejovsky, J. E., & Lau, T. S. (2016). Between-case standardized mean difference effect sizes for single-case designs: A primer and tutorial using the scdhlm web application. *Campbell Systematic Reviews*, *12*(1), 1–31. https://doi.org/10.4073/cmdp.2016.1

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

Wang, S. Y., & Parrila, R. (2008). Quality indicators for single-case research on social skill interventions for children with autistic spectrum disorder. *Developmental Disabilities Bulletin*, *36*(1 & 2), 81–105.

What Works Clearinghouse. (2011). *What Works Clearinghouse™ Procedures and Standards Handbook* (*version 2.1*). https://ies.ed.gov/ncee/wwc/Handbooks

What Works Clearinghouse. (2014). *What Works Clearinghouse™ Procedures and Standards Handbook* (*version 3.0*). https://ies.ed.gov/ncee/wwc/Handbooks

What Works Clearinghouse. (2017). *What Works Clearinghouse™ Procedures Handbook* (*version 4.0*). https://ies.ed.gov/ncee/wwc/Handbooks

What Works Clearinghouse. (2017b). *What Works Clearinghouse™ Standards Handbook* (*version 4.0*). https://ies.ed.gov/ncee/wwc/Handbooks

What Works Clearinghouse. (2020). *What Works Clearinghouse™ Procedures Handbook* (*version 4.1*). https://ies.ed.gov/ncee/wwc/Handbooks

What Works Clearinghouse. (2020b). *What Works Clearinghouse™ Standards Handbook (version 4.1)*. https://ies.ed.gov/ncee/wwc/Handbooks

What Works Clearinghouse. (2021b). *Search Publication*. RetrievedJanuary, 2021b, from https://ies.ed.gov/ncee/wwc/Publication#/ContentTypeId:1