


Teacher-Delivered Strategies to Increase Students' Opportunities to Respond: A Systematic Methodological Review

Behavioral Disorders
2020, Vol. 45(2) 67–84
© Hammill Institute on Disabilities 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0198742919828310
journals.sagepub.com/home/bhd


Eric Alan Common, PhD, BCaBA¹, Kathleen Lynne Lane, PhD, BCBA-D, CF-LI², Emily D. Cantwell, MEd², Nelson C. Brunsting, PhD³, Wendy Peia Oakes, PhD⁴, Kathryn Ann Germer, MEd, BCBA⁵, and Leslie Ann Bross, MSEd²

Abstract

We conducted this systematic review to map the literature and classify the evidence-based status of teacher-directed strategies to increase students' opportunities to respond (OTR) during whole-group instruction across the K-12 continuum. Specifically, we conducted this review to determine whether OTR could be classified as an evidence-based practice according to Council for Exceptional Children's *Standards for Evidence-Based Practices in Special Education*. We examined the extent to which 21 included studies addressed quality indicators and evidence-based practice standards using a modified, weighted criterion for methodologically sound studies. Three studies met all eight quality indicators and 11 studies met or exceeded 80% of quality indicators following a weighted criterion to define methodologically sound studies. Results indicated teacher-directed OTR strategy of response cards in K-12 school settings to be a potentially evidence-based practice. Educational implications, limitations, and future directions are discussed.

Keywords

opportunities to respond, active student responding, choral responding, response cards, evidence-based practice

Research has demonstrated academic engagement to be a critical predictor of students' school achievement (Brophy & Good, 1986). In large-group situations, teachers' implementation of instructional strategies is an important determinant of engagement for students who engage in challenging behavior (Downer, Rimm-Kaufman, & Pianta, 2007). Students who are not academically engaged may become passive learners, give up easily on tasks, and become anxious, withdrawn, or angry about school—leading to unsuccessful school experiences (Montague & Bergeron, 1997). As such, it is important for teachers to use high-leverage practices to promote active student engagement to facilitate success (McLeskey et al., 2017). Collectively, these practices promote safe, positive learning environments that foster academic engagement and decrease disruption. One practice shown to be effective for students who persistently engage in behavior challenges (e.g., students with emotional or behavior disorders [EBD]) is increasing students' opportunities to respond (OTR; Adamson & Lewis, 2017).

(c) promoting rapid student response through various modalities (e.g., verbal, gestural, textual), and (d) providing immediate feedback. OTR can be teacher-mediated (e.g., choral responding), technology-mediated (e.g., gaming), or peer-mediated (e.g., peer-tutoring). Ideally, teachers present students with multiple and varied OTR during a lesson at a brisk pace, but not so rapid that students are unable to participate (Sutherland & Wehby, 2001). By making simple shifts during instructional activities, teachers can promote and support the engagement of multiple students. In addition to being associated with higher rates of on-task behavior and lower rates of disruption for students with EBD (Sutherland & Wehby, 2001), OTR strategy can promote fluency and automaticity in basic skills of any content

Increasing Students' Opportunities to Respond

Use of OTR includes procedures for (a) presenting materials, (b) asking students questions at a high rate,

¹University of Michigan–Flint, USA

²The University of Kansas, Lawrence, USA

³Wake Forest University, Winston-Salem, NC, USA

⁴Arizona State University, Tempe, USA

⁵Walnut Creek, CA, USA

Corresponding Author:

Eric Alan Common, Assistant Professor, Department of Education, University of Michigan–Flint, 430-H French Hall, 303 East Kearsley, Flint, MI 48502-1950, USA.

Email: ecommon@umflint.edu

area and be used to formatively assess students' proficiency with material (Lane, Menzies, Ennis, & Oakes, 2015).

Teacher-delivered OTR strategy comprises three main elements: (a) identifying the content or skills to be targeted, (b) preparing an extensive set of questions or prompts that offer students practice with the material, and (c) leading the session with a high rate of questioning, rapid student responding, and immediate teacher feedback (Lane et al., 2015). A variety of student response formats can be utilized, including verbal (e.g., choral responding), physical (e.g., thumbs up or down, response cards), and electronic (e.g., clickers).

In 1987, the Council for Exceptional Children (CEC) recommended OTR to occur (a) four to six times per min for new material, with students responding with 80% accuracy, and (b) eight to 12 times per min for review material with students responding with 90% accuracy. Stichter and colleagues (2009) suggested an optimal rate of 3.5 OTR per min. This rate is supported by results suggesting slight differences in students' on-task behavior at three and five OTR per min (Sainato, Strain, & Lyon, 1987). Naturally occurring rates of OTR tend to fall below recommended levels, with a reported average of 2.61 per min ($SD = 0.66$; Stichter et al., 2009).

Establishing an Evidence Base

Evidence-based practices (EBP) can refer to a process or an instructional technique (Cook, Cook, & Collins, 2016). For instance, the process of EBP considers instructional decision-making based on the best available evidence, professional judgment, and preferences and needs of students (Spencer, Detrich, & Slocum, 2012), whereas EBPs are strategies, practices, or programs (a) supported by a body of high-quality, peer-reviewed, experimental research and (b) that have undergone a systematic evidence-based review and classified as evidence based (Cook et al., 2016). Given mandates, such as Every Student Succeeds Act (2015), charging schools to provide high-quality instruction for all students, critical instructional techniques are appraised both for methodological quality and for magnitude of effect to identify EBPs.

Quality Appraisals

In 2005, Horner et al. and Gersten et al. introduced standards for identifying EBPs in special education using single-case research design (SCRD) and group-comparison designs, respectively. Lane, Kalberg, and Shepcaro (2009) field-tested SCR D standards (Horner et al., 2005). Following this initial application, Lane et al. suggested initial standards to classify EBPs may be too conservative of a standard in determining "what works." Lane et al. raised concerns that

overly rigorous criteria may lead to the unintended consequence of having too few EBPs for use. As a result, they recommended using an 80% criterion for identifying studies as methodologically rigorous (and, therefore, eligible to be considered when classifying the evidence base of instructional techniques), rather than a 100% criterion across all quality indicators (QIs). In 2014, CEC proposed new *Standards for Evidence-Based Practices in Special Education* (hereafter referred to as *Standards for EBPs*), which also required studies meet 100% of QIs to be considered methodologically sound and included in EBP reviews. Recently, reviews have begun to apply Lane et al.'s weighted criterion to the *Standards for EBPs* (Common, Lane, Pustejovsky, Johnson, & Johl, 2017; Ennis, Royer, Lane, & Griffith, 2017).

Quantifying the Evidence Base

The emergence of EBP as a priority in education, both as a process and in the identification of instructional techniques, has placed increased emphasis on not only the methodological rigor (e.g., QIs) but also the magnitude of the effect across rigorous (e.g., methodologically sound; CEC, 2014) studies. SCR D has historically emphasized visual analysis to assess and report the effects of treatments, and many scholars have been skeptical of whether syntheses employing statistical analyses can capture nuances of SCR D (Shadish, Hedges, Horner, & Odom, 2015). The extent to which SCR D effect sizes (e.g., between-case standardized mean difference [BC-SMD]) and other quantitative indices (e.g., percentage of nonoverlapping data [PND]) adequately estimate the direction and magnitude of functional relations remains a subject of continued interest and debate (Ledford & Gast, 2018).

Effect sizes. Even in non-meta-analytic reviews (e.g., EBP reviews), effect sizes are useful in comparing results across rigorous studies that could otherwise not easily be compared (CEC, 2014; Shadish et al., 2015). Effect sizes can be calculated within and across studies and be standardized or nonstandardized. Standardized effect sizes put study results on a scale with the same meaning across studies (e.g., standardized mean difference, risk ratios, odd ratios; Shadish et al., 2015) and are particularly important when examining a body of evidence comprising a range of methodologies (e.g., SCR D, group-comparison designs).

Ideally, effect sizes are metrics that can be validly compared across studies using various designs (Pustejovsky, 2018). Hedges, Pustejovsky, and Shadish (2012, 2013) introduced BC-SMD, an effect size for SCR D directly comparable to standardized mean difference effect sizes used in group-comparison designs. BC-SMD is based on a hierarchical model for the within-case and between-case variation in the dependent variable (DV) captured in SCR D employing

withdrawal/reversal (AB_k) design, multiple-probe design, and multiple-baseline design (MBD) with three or more cases (Shadish et al., 2015). Although comparable, metrics from SCRD are relatively new and tend to be larger than those from group designs and should be interpreted with caution (Barton, Pustejovsky, Maggin, & Reichow, 2017).

More recently, Pustejovsky (2018) introduced the log response ratio (LRR), another effect size for SCRD, that is not constrained by number of cases and less constrained by design. LRR effect size is a within-case effect size (ES_{wc}) particularly well suited for single-case demonstration designs, with behavioral outcomes measured through systematic direct observation (Pustejovsky, 2015, 2018). Although there may not be a consensus on whether and which metrics should be used to discern magnitude effect in SCRD, there is growing consensus that, when used, scholars must demonstrate how their selection of effect sizes/quantitative indices should be made in the context of the set of studies to be synthesized (Maggin, Lane, & Pustejovsky, 2017).

Opportunities to Respond: Lessons Learned

Reviews examining specific OTR strategies include examination of response cards (Horn, 2010; Randolph, 2007; Schnorr, Freeman-Green, & Test, 2015) and choral responding (Haydon, Marsicano, & Scott, 2013). Randolph (2007) meta-analyzed studies examining response cards and found statistically significant effect sizes for achievement ($d = 1.08$), as well as substantial increases in student participation (47.70%) and decreases in off-task behavior (34.34%). Horn (2010) extended this review of response card for students with disabilities and offered initial evidence for considering response cards as an EBP using Horner et al.'s (2005) guidelines. Although Horn provided descriptive information and concluded guidelines for EBP were met, a methodological quality appraisal of included studies was not reported.

Haydon et al. (2013) conducted a review of the literature comparing choral and individual responding. Findings suggested choral responding resulted in higher levels of active student responding and on-task, appropriate behavior, as well as decreases in students' disruptive and inappropriate behaviors. More recently, Schnorr et al. (2015) offered the first methodological appraisal of an OTR strategy and examined response cards in elementary settings. Results indicated sufficient support for response cards as an EBP with a moderate level of evidence for increasing OTR for elementary students. Yet, like previous reviews, their review, focused on a specific OTR strategy and not the full range of methods possible for student responding during whole-group OTR (e.g., choral responding, clickers).

MacSuga-Gage and Simonsen (2015) examined varying modalities of teacher-delivered OTR, with results indicating choral responding resulted in positive academic and behavioral outcomes across students when compared with

individual responding. They found no studies conclusively examined differential effects of OTR rates nor identified the optimal rate of teacher-delivered OTR. All studies exploring the impact of increased rates of OTR demonstrated positive outcomes for students with and without disabilities, including increased correct responses, student participation, and on-task behavior and decreased off-task and disruptive behavior. Yet, MacSuga-Gage and Simonsen's study did not evaluate the methodological rigor necessary for classifying the evidence base of teacher-delivered OTR.

Purpose

We conducted the current EBP review to examine the effectiveness of teacher-delivered OTR strategies during whole-group instruction across the K-12 continuum. Specifically, we (a) mapped descriptive characteristics of included studies, (b) appraised the methodological rigor of included studies, (c) determined the evidence-based classification of OTR strategy, and (d) described the magnitude effects of OTR across methodologically sound studies.

Method

Article Selection Procedures

Article procurement was conducted independently by two or more authors at each step and included electronic, hand, and ancestral searches of the literature, initially conducted in Spring 2016 and again in Winter 2017, with searches concluding in December 2017. The electronic search included four databases: *ERIC*, *ProQuest Research Libraries*, *PsycArticles*, and *PsycINFO*. The following search string was used to identify potential records: all("Choral Respon*") OR all("signal* system*") OR all("individual white board") OR all("student response system*") OR all("clicker*") OR all("communication cups*") OR all("response card*") OR all("Opport* to respond*") OR all("active student respond*"), NOT all("higher education" OR "medical students" OR "college students" OR "adult education" OR "distance learning" OR "community college" OR "college" OR "undergraduate").

Ancestral searches occurred for all included articles, as well as for other literature reviews examining OTR (Haydon et al., 2013; Horn, 2010; MacSuga-Gage & Simonsen, 2015; Randolph, 2007; Schnorr et al., 2015). Hand searches were conducted for journals with two or more included studies (*Behavioral Disorders, Education and Treatment of Children, Journal of Applied Behavior Analysis, Journal of Positive Behavior Interventions, and Preventing School Failure*) from 1979 to 2017 (including online first), beginning the search from the first published study (McKenzie & Henry, 1979). Primary and secondary coders independently read titles and abstracts of each article to determine whether the full article should be read to further evaluate its

eligibility. When a disagreement occurred between coders, the article was read in full and a consensus model was used until agreement was achieved. See Figure 1 for the identification and inclusion process.

Inclusion Criteria

We used a binary coding scheme of met/not met to determine whether studies met the inclusion criteria. First, all studies had to be conducted using group comparison or SCRD (CEC, 2014). Second, studies needed to include a teacher-delivered method of increasing students' OTR (e.g., choral responding, signals such as thumbs up/down, communication or signaling cups, response cards, student response system, clickers) as the independent variable (IV; MacSuga-Gage & Simonsen, 2015). As such, interventions targeting peer-mediated strategies (e.g., classroom-wide peer-tutoring, Greenwood, Delquadri, & Hall, 1989; numbered heads together, Maheady, Mallette, Harper, & Sacca, 1991) were not included in this review. Third, the study's intervention needed to be teacher directed during whole-group instruction toward K-12 children and youth. Interventions could take place in general or special education classrooms. Fourth, studies included at least one student-level academic or behavior outcome DV. Finally, studies not written in English or included in peer-reviewed journals were excluded.

Coding Procedures

Descriptive coding. To provide descriptive context, we mapped the literature by coding description of practice, context and settings, participants, intervention agent, implementation fidelity, internal validity, outcome measures/DV, and data analysis. Inter-rater agreement (IRA) was 94.89%.

Quality indicator coding. To appraise the methodological quality of included studies, two authors independently coded every article using the eight categories of QIs in the *Standards for EBP* (full descriptions to follow). Across these QIs, coding components included either the 22 items for SCRD studies or the 24 items for group-comparison studies (CEC, 2014). We used a coding protocol developed by Lane, Common, Royer, and Muller (2014). The first and second authors were trained to reliability at 85% or higher across three or more consecutive articles not included in this review. Average IRA across four training articles was 90.90% ($SD = 6.43$).

Given the methodological quality of a study exists on a continuum—ranging from no methodological rigor to a strong methodological rigor—we followed recommendations by Lane et al. (2009) to report the degree to which each QI was met by using a weighted coding scheme.

Rather than using an absolute coding scheme (QI met/QI not met), we allowed each component constituting an indicator that was present to contribute partially. We used a binary scale coding scheme for each component (*met* [1], *not met* [0], or *not applicable* [NA]) within an indicator. For each QI, the number of components met within each indicator (range: 1-6) was summed and divided by the total number of components scored. Components coded as not applicable were dropped from denominator. Weighted scores ranged from 0 to 1 (rather than 0 or 1). Disagreements were resolved through a consensus process. IRA across studies was 92.01% ($SD = 0.09$) and 93.38% ($SD = 0.08$) across components.

Methodological Quality Indicators

1.0 Context and setting. This indicator included one component. To meet *1.1 context/setting* description, investigators needed to describe critical features of the context or setting relevant to the review (CEC, 2014). This component was considered met if at least one setting/context feature (e.g., region, type of school/classroom) was described (Lane et al., 2014).

2.0 Participants. This indicator included two components. To meet *2.1 participant description*, investigators needed to describe participant demographics relevant to the review (CEC, 2014). This component was met if at least one demographic element (e.g., age, gender) was reported (Lane et al., 2014). To meet *2.2 participant disability/at-risk status*, investigators needed to describe participants' disability or risk status and method of determination (CEC, 2014; Lane et al., 2014). We did not require risk status to be reported when the whole class was the unit of analysis (Ennis et al., 2017). We considered the following as insufficient: (a) global definitions, such as behavioral disabilities, and (b) vague descriptions that were not described with replicable precision, such as teacher nomination (Lane et al., 2014). This component was considered nonapplicable for studies not including participants with disability/at-risk status.

3.0 Intervention agent. This indicator included two components. To meet *3.1 role description*, investigators needed to describe intervention agent's role (e.g., researcher, teacher; CEC, 2014). To meet *3.2 training description*, investigators needed to report information on how intervention agent(s) received training and how investigators checked for understanding (e.g., trained to criterion, role-play). Furthermore, if the intervention agent was both a teacher and an author, author affiliation and/or authors' notes were used to reasonably determine the extent to which the author was competent in OTR strategy (e.g., designed intervention as part of guided study, theses, or dissertation process).

4.0 Description of practice. This indicator included two components: To meet 4.1 *intervention procedure description*, investigators needed to provide details with replicable precision (CEC, 2014). For 4.2 *materials description*, investigators needed to include a description of materials needed to implement intervention or offer accessible references providing this information (CEC, 2014). The second component was considered nonapplicable to studies not requiring materials (Cook et al., 2015).

5.0 Implementation fidelity. This indicator included three components. To meet 5.1 *implementation fidelity*, investigators needed to assess and report implementation fidelity using direct, reliable measures of adherence. To meet 5.2 *dosage or exposure assessed/reported*, investigators needed to assess and report implementation fidelity related to dosage or exposure to treatment conditions (CEC, 2014). This was considered met by reporting length of time of intervention or how long the intervention was in place (e.g., available from time-series line graph). Finally, to meet 5.3 *assessed across relevant elements and/or throughout study*, investigators needed to (a) assess and report implementation fidelity regularly and throughout the intervention (e.g., beginning, middle, and end), and (b) specify when, where, and for whom fidelity was assessed and report fidelity (Cook et al., 2015). This was considered present if any mention of assessing implementation fidelity occurred across different time points of the intervention. Studies did not have to report a measure of fidelity for each condition if an aggregated measure across conditions was reported. If neither adherence (5.1) nor dosage (5.2) was assessed, 5.3 was not applicable (CEC, 2014).

6.0 Internal validity. This indicator included six components, three shared by SCRD and group-comparison designs (6.1, 6.2, and 6.3), with three additional components specific to SCRD (6.5, 6.6, and 6.7) and three specific to group-comparison designs (6.4, 6.8, and 6.9). To meet 6.1 *IV systematically manipulated*, investigators were required to control and systematically manipulate the IV (CEC, 2014) and measure treatment fidelity of intervention (Lane et al., 2014). To meet 6.2 *baseline description*, investigators needed to describe baseline or control/comparison group conditions. To meet 6.3 *no or limited access to IV during baseline*, investigators needed to explicitly state or measure that nonintervention conditions did not have exposure to intervention (Lane et al., 2014). To meet 6.4 *group assignment*, investigators needed to describe assignment to group, which must have involved unit of analysis (e.g., participants, schools) being assigned randomly, nonrandomly and matched, or nonrandomly with meaningful differences identified and statistically controlled. To meet 6.5 *three demonstrations of experimental effect*, investigators must have employed a design that allowed for the possibility of

three demonstrations or replications of an experimental effect at three different time points (CEC, 2014). To meet 6.6 *baseline: minimum three data points and established pattern*, investigators needed to include at least three baseline data points unless justified by the study author (CEC, 2014). This component was not applicable to SCRD not requiring baseline (e.g., alternating treatment designs [ATDs]) although if baseline was included this component was assessed. To meet 6.7 *controls for threats to internal validity*, investigators must have employed an accepted SCRD (Ledford & Gast, 2018) with procedural integrity (Lane et al., 2014). To meet 6.8 *overall attrition*, overall attrition needed to be low across groups (e.g., <30% in a 1-year study; CEC, 2014). Finally, to meet 6.9 *group attrition*, differential attrition between groups needed to be low (e.g., ≤10%) or controlled for (CEC, 2014).

7.0 Outcome measures/dependent variables. This indicator included six components, of which the first five (7.1-7.5) applied to both SCRD and group-comparison designs, and one additional component specific to group-comparison design. To meet 7.1 *socially important*, investigators needed to discuss (e.g., introduction or discussion) the social significance of the goals, social appropriateness of the procedures, and/or social importance of the effects and/or explicitly measured and reported social validity (Lane et al., 2014). To meet 7.2 *description of DV measures*, investigators needed to define and describe each DV and use a valid measurement system (CEC, 2014). To meet 7.3 *reports effects on the intervention on all measures*, investigators needed to report the effects of the intervention across all outcome measures (CEC, 2014). To meet 7.4 *measured repeatedly (minimum three data points per phase)*, investigators needed to measure outcomes with appropriate frequency and timing (e.g., minimum of three data points per phase [e.g., AB_k, MBD, changing criterion design]; at least four repetitions of alternating sequence [e.g., ATD]; Ledford & Gast, 2018). For 7.5 *adequate interobserver agreement (IOA)*, investigators needed to provide evidence of adequate IOA by meeting minimal standards (i.e., IOA ≥80%, κ ≥60%; CEC, 2014) across participants and DVs. This component was considered met for aggregated data if the study stated IOA occurred across participants or conditions, and if averages met specified levels and any reported range did not fall below 60% IOA (Lane et al., 2014). Finally, for group-comparison designs only, 7.6 *validity* was considered met if investigators reported either (a) adequate validity coefficients or (b) outcomes adequately represented content measured (i.e., content validity; CEC, 2014).

8.0 Data analysis. This indicator included two components specific to group-comparison design (8.1, 8.3) and one component specific to SCRD (8.2). To meet *QI 8.1. data analytic techniques*, group designs studies needed to employ (a)

statistical analysis procedures generally recognized as appropriate for comparing change in the performance of two or more groups, or (b) atypical procedures were used, but justified and explained. To meet 8.2 *graph clearly represents outcome data*, SCRDS need to clearly represent outcome data for all student outcome measures by providing graphs that allowed for the possibility of visual analysis (e.g., examine level, trend, and stability within and across conditions). Finally, to meet 8.3 *effect sizes*, studies need to report effect sizes or provide data from which appropriate effect sizes can be calculated.

Evaluation Procedures for Determining Evidence-Based Practices

To answer questions related to classifying the evidence base for OTR strategies, we restricted included studies by including only demonstration (e.g., AB_k, MBD) and comparison designs evaluating and comparing OTR strategy as an IV with other IVs different from OTR strategy. As such, studies comparing multiple variations of OTR strategy were quality appraised but excluded in evaluating and classifying the evidence base.

Classifying methodologically sound studies. CEC (2014) defined methodologically sound studies as meeting all of the QIs across components. We utilized a modified criterion (Lane et al., 2009) and defined methodologically sound as studies meeting 80% or more of all eight QIs. A weighted criterion to define methodologically sound articles is based on the logic that rigor exists as a continuum (e.g., no rigor, some rigor, high rigor) rather than as a dichotomy (present, absent). We acknowledge this does not strictly adhere to *Standards for EBPs*' recommendation for classifying effects of studies.

Classifying study effect. To classify effects of group-comparison studies deemed methodologically sound, we followed the recommendations of *Standards for EBPs*: negative effect if $ES \leq -0.25$, mixed/neutral effect if $-0.24 > ES > 0.24$, and positive effect if $ES \geq 0.25$. We calculated Hedges's *g* for studies that did not report effect sizes but reported information from which effect sizes could be calculated (e.g., *M*, *SD*, *n*; *t*, *df*). We dropped studies with inconsistent reporting of information necessary to calculate effect sizes (e.g., different *t* values).

To classify effects of SCRDS deemed methodologically sound and with three or more cases, we employed visual analysis to discern positive, neutral or mixed, or negative effects, based on (a) number and proportion of participants for whom a functional relation was established and (b) direction of functional relation (CEC, 2014). The presence of a functional relation was evaluated independently by two

authors examining graphed data within and across phases for changes in level (e.g., low, moderate, or high), trend (e.g., increasing, decreasing, or flat), and stability (stable, variable; Ledford & Gast, 2018). Studies were determined to have *positive effects* if (a) 75% of cases demonstrated a functional relation between the IV and therapeutic changes in the DV, (b) there was no evidence of counter-therapeutic effects, and (c) remaining cases were neutral or mixed (i.e., no negative effects). Studies were determined as having *negative effects* if 75% of its cases demonstrated a functional relation between IV and unfavorable changes in DV (e.g., counter-therapeutic effects). Studies were determined as having *mixed or neutral effects* if it neither qualified as having positive or negative effects. IRA between visual analysis coders was 100%.

Classifying the evidence base. According to the *Standards for EBP*, for a strategy, practice, or program to be considered *evidence based*, it must be supported by (a) two methodologically sound group-comparison studies with random assignment to groups and unit of analysis aligned with unit of assignment, positive effects, and at 60 or more participants across studies; four methodologically sound group-comparison studies with random assignment but unit of analysis not aligned with unit of assignment or non-random assignment to groups; positive effects and at 120 or more participants across studies; or five methodologically sound SCRDS with positive effects and at 20 or more participants across studies; or (b) meet at least 50% of criteria for two or more of the study designs (CEC, 2014, p. 8). In addition, no methodologically sound studies can have negative effects, and the ratio of positive to neutral/mixed effects must be 3:1 or greater. See CEC's (2014) *Standards for EBPs* for classification requirements for *potentially evidence-based*, *mixed evidence*, *insufficient evidence*, and *negative effects*.

Determining Magnitude of Effect

To complement visual analysis of methodologically sound SCRDS studies with three or more cases, we additionally calculated ES_{BC} and ES_{WC} . We digitized published graphs of methodologically sound studies using data extraction software WebPlotDigitizer (Version 3.12; Rohatgi, 2015). Digital data were extracted independently by one of two authors, cleaned and formatted for statistical software, and made reliable against original graphs by a second author prior to analyses. Data without clearly marked legend keys or titles explaining the data (e.g., DV, unit of analysis) were dropped from these analyses. Effect sizes were screened using Grubbs's test for outliers in R (Komsta, 2011). Omnibus effect sizes were not calculated across articles and are reported by article (BC-SMD) or case (LRR).

Between-case effect sizes. We selected BC-SMD (Hedges et al., 2012, 2013) for SCRDs. BC-SMD technical requirements include SCRDs that (a) use MBD, multiple probe, or AB_k designs; (b) contain three or more cases; and (c) assume no trend (Shadish et al., 2015). We calculated BC-SMD using the online BC-SMD calculator developed by Pustejovsky (2016). We used the restricted maximum likelihood estimation method and specified (a) fixed effect and random effect for the baseline phase (i.e., permitted intercept [level] across all baseline phases to be different from zero and vary across cases, respectively) and (b) fixed effect for the treatment phase level (i.e., permitted the intercept [level] across all intervention phases to vary from the baseline phase level). Furthermore, we specified random effect for the treatment phase (i.e., permitted treatment effect to vary across cases). Finally, following recommendations from Valentine, Tanner-Smith, Pustejovsky, and Lau (2016), we assumed treatment effects to be constant across cases by omitting random effects for treatment phase level. Not enough information was reported in group-comparison studies to calculate Hedges's g . We employed Shadish, Zelinsky, Vevea, and Kratochwill's (2016) descriptive quartiles, which divided 74 previously published BC-SMD estimates into four groups for interpretation: 0 to 0.36 = nominal effect, 0.37 to 0.97 = small effect, 0.98 to 1.86 = medium effect, and ≥ 1.87 = large effect.

Within-case effect sizes. We selected LRR as our ES_{WC} single-case parametric over regression-based metrics, which account for trend, because regression-based approaches have additional technical constraints related to (a) insufficient number of data points in the initial condition to predict accurately, and (b) too much (i.e., instability) or not enough (i.e., zero baselines) variance in baseline to accurately predict performance in adjacent conditions. LRR conceptualizes the proportionate change for an individual case across two adjacent conditions (e.g., A-B). It was important to select an ES_{WC} flexible enough to be (a) calculated across a range of studies that do not meet the technical requirements of BC-SMD (Common et al., 2017), and (b) used to draw conclusions about each case separately, a hallmark of SCRD (Ledford & Gast, 2018; Shadish et al., 2015). ES_{WC} (comparisons are made within individual participants) differs conceptually from ES_{BC} (comparisons are made between average performance across participants). As such, it is impossible to compare ES_{WC} with ES_{BC} (Shadish et al., 2015). This limits the extent to which ES_{WC} can be utilized in quantitative reviews examining both group comparison and SCRDs.

The technical requirements of LRR assume the pattern of behavior within each phase lacks time trends (e.g., stable from session to session). When applied to DVs on a

scale of 0% to 100%, LRR requires all outcomes to be defined in the same direction of therapeutic change. All but one DV in this review consisted of student-level outcomes with a therapeutic direction being upward; thus, one DV (percentage off-task) was recoded to percentage on-task (i.e., $100 - \% \text{ off-task} = \% \text{ on-task}$). LRRs were calculated using the online single-case effect size calculator (Pustejovsky, 2017). We followed recommendations set forth by Pustejovsky (2017) for AB_k design studies with multiple A-B comparisons and estimated LRR for each pair of adjacent phases and combined those estimates to average a single summary effect size for each case (Pustejovsky, 2018). Furthermore, we excluded cases for LRR calculations under the following conditions: (a) either phase in an A-B contrast has fewer than three data points, (b) there was zero responding within a baseline phase, or (c) there was near-zero responding in a phase followed by a ceiling effect in the next phase. For ATD, LRR was adopted and treated as an AB_k design (e.g., A-B, A-C, A-D, B-C, B-D; Zelinsky & Shadish, 2018). For interpretation, LRR effect sizes employ directionality, with negative values of LRR corresponding to decreases, values of zero corresponding to no changes, and positive values corresponding to increases (Pustejovsky, 2015). To aid in the interpretation of LRR, percentage change was calculated from the LRR parametric using the following formula: $100 \times [\text{exp}(\text{LRR}) - 1]$.

Results

Descriptive Characteristics of Included Studies

Twenty-one studies were included and published across nine unique journals from 1979 to 2017 (see Figure 1). One study employed a group-comparison design and 20 studies employed an SCRD ($AB_k = 13$, ATD = 6, within-subject cross-over design = 1). Studies included participants from K to 11th grade. Ten studies took place in an elementary school, five in middle school, and four in high school. Two studies did not specify the school level, but specified classroom grade: third grade (McKenzie & Henry, 1979) and fifth grade (Munro & Stephenson, 2009). See Table 1 for additional information pertaining to context and setting. Although all studies were implemented during whole-class instruction, not all students were selected as participants for data recording. Across studies, 166 students participated in SCRD studies and 52 students were assigned to two theoretically comparable treatment groups in a group-comparison study. The predominant description of practices were response cards ($k = 13$; 61.90%), followed by verbal or nonverbal choral responding ($k = 5$; 23.80%), mixed-mode responding ($k = 3$; 14.29%), and student response systems/clickers ($k = 2$; 9.52%).

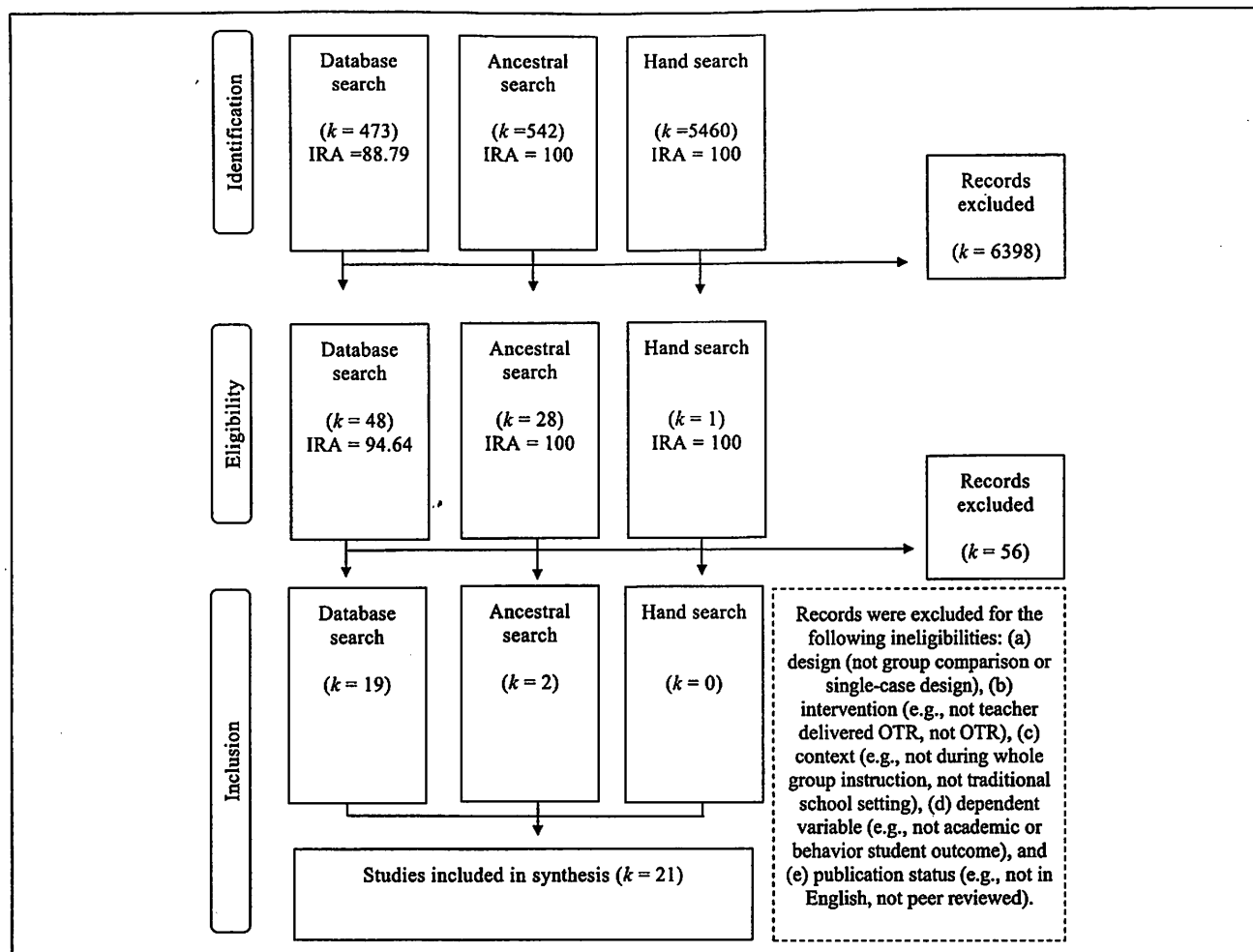


Figure 1. Article procurement flow diagram.

Note. Eligibility = full text articles assessed for eligibility; Identification = records identified; Inclusion = studies included in synthesis; IRA = inter-rater agreement; OTR = opportunities to respond.

Methodological Quality Indicators

Results of the methodological quality appraisal (CEC, 2014) are provided in Figure 2. Three studies (Adamson & Lewis, 2017; Haydon, Musti-Rao, & Alter, 2017; Messenger et al., 2017) met all eight QIs (mode = 4; range: 3-8). Eleven studies (see Figure 2) met or exceeded 80% or more of the QIs, which we defined as being methodologically sound ($M = 6.73$; $SD = 0.93$; range: 4.97-8.00).

Evaluation of the Practice

Four methodologically sound SCRD studies included three or more cases and examined the effectiveness of OTR strategy ($n = 17$; Adamson & Lewis, 2017; Clarke, Haydon, Bauer, & Epperly, 2016; Munro & Stephenson, 2009; Wood, Mabry, Kretlow, Lo, & Galloway, 2009). Two methodologically sound studies were excluded from

classifying the evidence base of the OTRs because they employed an ATD comparing more than one variation of an OTR strategy (Haydon et al., 2017; Messenger et al., 2017). Munro and Stephenson (2009) demonstrated positive effects of response cards on student-initiated responses. Wood et al. (2009) demonstrated positive effects of response cards on students' on-task behavior and participation. Clarke et al. (2016) demonstrated positive effects of response cards on student responding. Adamson and Lewis (2017) demonstrated positive effects of response cards when contrasted with class-wide peer-tutoring and guided notes on students' academic engaged time. Thus, teacher-delivered OTR strategy—specifically response cards—during whole-group instruction meets criteria for being a *potentially EBP* when a weighted criterion was used to define methodological rigor. See Table 2 for a summary of visual analysis of methodologically sound studies with three or more cases.

Table 1. Descriptive Results of Included Articles.

Study element	1979 McKenzie	1990 Narayan	1994 Gardner	1996 Cavanaugh
Description of practice	Nonverbal choral response (e.g., hand raise)	Response cards	Response cards	Response cards
Context and setting	3rd-grade classrooms (instructional sessions about mountains); suburban school	4th-grade classroom (social studies); urban public ES	5th grade (science); large Midwestern city ES	9th grade (earth science); suburban public HS
Participants				
Classroom	Two treatment groups: 52 students from two classes	Class: 8 boys, 12 girls; age 9-11	Class: 13 boys, 11 girls; age 10-12	Class: 23 students, 15 GE and 8 with disabilities (LD, EBD, ID) or at risk
Inclusion criteria	Whole class	Teacher nomination: range of overall skills levels	Teacher nomination: range of participation and academic performance	Whole class
Target students	Whole class	6 students	5 students	Whole class
Intervention agent	Classroom teacher	First author	First author as classroom teacher	Teacher
Implementation fidelity	—	DV of teacher presentation rate	Script and DV of teacher presentation rate	Procedural fidelity checklist
Internal validity				
Design	Group	A-B-A-B	A-B-A-B	ATD
Baseline	Control group	Hand-raising condition	Hand-raising condition	No baseline; response cards against passive review
Outcome measures/DVs	On-task behavior, text anxiety attitude items, and achievement	Teacher presentation rate; student responses, accuracy of student responses, and daily quiz scores	Teacher presentation rate; student responses, accuracy of student responses, next-day quiz scores, and biweekly review test scores	Next day and weekly tests
Social validity	Discussed	Interviews	Interviews	Discussed
Study element	1999 Armendariz	2002 Maheady	2003 Christle	2004 Davis
Description of practice	Response cards	Whole-group question and answer, response cards, and numbered heads together	Response cards	Response cards
Context and setting	Bilingual 3rd grade (math); urban ES	6th grade (general science); small urban MS	4th grade (math); urban ES (79.1% FRL)	7th and 8th grade (SCC English); MS
Participants				
Classroom	Class: 11 boys, 11 girls; age 8-9	Class: 7 boys, 14 girls; age 11-13; 3 students with LD, 1 student with SED, 1 student with ADHD, 4 receiving remedial reading instruction, and 2 students receiving ESL	Class: 9 boys, 15 girls; age 9-11; 8 students were Hispanic	Class: 11 students with LDs, including ESL learners
Inclusion criteria	Whole class	Whole class	Teacher nomination: range of academic skill, participation, and on-task behavior	Reported low levels of active responding and high rates of off-task behavior
Target students	Whole class	Whole class	5 students: 2 boys, 3 girls; age 9-11; 2 were Hispanic and 1 attended a special reading program for below-grade-level readers	4 students
Intervention agent	Classroom teacher	Classroom teacher	Classroom teacher	Classroom Teacher
Implementation fidelity	—	Procedural fidelity checklist	DO of planned teacher behavior	—

(continued)

Table 1. (continued)

Study element	1997 Armendariz	2002 Maheady	2003 Christle	2004 Davis
Internal validity				
Design	A-B-A	ATD	A-B-A	A-B-A-B
Baseline	Hand-raising condition	N/A	Hand-raising	Hand-raising condition
Outcome measures/DVs	Disruptive behavior	Primary: accuracy of responses. Secondary: instructional process variables; active pupil responses and on-task behavior	Number of student responses and initiated responses, weekly quiz score, and on-task behavior	Percentage of academic responses and off-task behavior
Social validity	Preference check	Consumer satisfaction survey	Preference check	Questionnaire
Study element	2006 Lambert	2009 Haydon	2009 Munro	2009 Wood
Description of practice	Response cards	Increased rate of questions and varied mode of questioning	Response cards	Response cards
Context and setting	4th grade (math); Midwestern urban ES (preschool-5th grade)	5th grade (science); north central Florida district ES	5th grade (English); urban public school in British Columbia, Canada	Kindergarten inclusion class (circle time: calendar); rural ES
Participants				
Classroom	2 Classrooms (15 and 16 students)	Class: 19 students	Class: 15 boys, 14 girls; age 10-11	Class: 12 boys, 11 girls; age 5-6; 21 White, 1 Hispanic, and 1 multiracial
Inclusion criteria	Teacher nomination: most disruptive, least attentive during math lessons, and lowest performance in math	(a) Demonstration of chronic disruptive behavior in the classroom, (b) significant rating indicating at-risk for EBD on the SSBD, and (c) teacher nomination	Teacher nomination: reluctant to respond during whole-class question-and-answer sessions	Teacher nomination: lack of participation and off-task behavior during group instruction
Target students	A total of 9 students—all eligible for FRL; -Classroom A: 2 boys, 2 girls; age 9; 3 Black, 1 White. Classroom B: 2 boys, 3 girls; age 9-10; all Black	1 student: girl. Age 11. Screened at-risk for EBD	5 students: age 10-11; 3 students immigrated 2-4 years prior to study	4 students: 2 boys, 2 girls; age 5-6; 3 White and 1 multiracial; 1 SL/LD and 1 DD
Intervention agent	Classroom teachers	Classroom teacher	Classroom teacher	Special education resource teacher
Implementation fidelity	Procedural integrity checklist	DO of teacher presentation rate and procedural checklist	DO of teacher presentation and feedback rates	Procedural reliability checklist
Internal validity				
Design	A-B-A-B	A-B-A	A-B-A-B	A-B-A-B
Baseline	Single-student responding (hand-raising)	Choral responding at naturally occurring rate	Hand-raising condition	Hand-raising condition
Outcome measures/DVs	Hand-raises and academic responses, correct responses, and disruptive behavior	Rate of question per minute; correct responses, on-task behavior, and disruptive behavior	Rate of teacher questions and feedback; student-initiated response opportunities and test scores	Off-task behavior and participation
Social validity	Interview questionnaire	Discussed	Discussed	Interview questionnaire
Study element	2010 Blood	2010 George	2010 Haydon	2011 Haydon
Description of practice	Student response system	Response cards	Individual responding, choral responding, and mixed-mode responding	Increased rate of questions and single student responding or unison hand-raising
Context and setting	9th-11th grade (self-contained; American history); suburban HS	6th-8th grade (emotional support classrooms); 4 suburban MS	2nd grade (sight words and syllable practice); 2 ES (1 urban, 1 suburban)	7th grade (health science class); large Midwestern urban MS (Grades 6-7)

(continued)

Table 1. (continued)

Study element	2010 Blood	2010 George	2010 Haydon	2011 Haydon
Participants				
Classroom	5 self-contained classrooms (class size range: 5-10)	5 classrooms	6 classrooms (class size ranged from 18 to 22 students). 50%-70% Black and 30%-50% White	20 students
Inclusion criteria	Off-task behavior, low response and participation, and good attendance	EBD	Whole class and consent to participants with chronic disruptive behavior, at-risk for EBD	Teacher nomination and demonstration of chronic off-task behavior
Target students	5 students 4 boys, 1 girl. Age 15-18; 4 White, 1 Native American/White. 2 EBD, 2 health impaired, and 1 autism	29 students: 23 boys and 6 girls. Age 11-15	6 students: 5 boys, 1 girl. Age 7-8; 5 Black, 1 White	2 students: male, Black: (a) Age 13—D and C student, frequently off-task; (b) Age 13—typically achieving, with few behavior problems
Intervention agent	Special education teacher	Special education teachers	Classroom teachers	Classroom teacher
Implementation fidelity	Frequency of questions asked	Frequency recording of teacher-posed questions	Procedural fidelity checklist and DO of teacher's implementation of the OTR procedure	Procedural fidelity checklist and DO of teacher following instructional sequence and DO of OTR rate
Internal validity				
Design	A-B-A-B-C	Within-subject cross-over design	ATD	A-B-C-B-C
Baseline	Business as usual	Business as usual	N/A	Typical instructional strategies (e.g., lecture, question and answer)
Outcome measures/DVs	Response rate, time on task, percentage correct on daily quizzes, percentage correct on end-of-phase quiz	Chapter posttest scores, academic responses, correct academic responses, on-task behavior, and student satisfaction surveys	Active student responding, off-task behavior, and disruptive behavior	Teacher-delivered praise statements and redirections; student correct responses, correct responses on test, and on-task behavior
Social validity	Discussed	Student satisfaction survey and open-ended questions for teachers	Teacher surveys	Teacher surveys
Study element	2015 Xin		2016 Clarke	2017 Adamson
Description of practice	Clickers		Response cards	Class-wide peer-tutoring, guided notes, and response cards
Context & setting	8th-grade SCC (language and math skills); urban MS		3rd grade (science and social studies); rural Midwestern ES	10th-11th grade (algebra); Midwestern HS
Participants				
Classroom	Number of students in class not specified		Class: 23 students	2 classrooms, 3 teacher-student dyads. Number of students per class not specified
Inclusion criteria	Not specified; risk status specified (2 with OHI [ADHD], 2 with EBD, 1 with LD with ADD)		High rates of on-task behavior but low rates of responding	Disability diagnosis, failing grade, and behavioral problems
Target students	5 students (4 boys, 1 girl. All Black. 14 years old)		5 students (3 boys, 2 girls); 8-9 years; all with ID and speech/language impairment)	3 students (all boys); age 15-16; 1 Black/White, 2 White, 2 OHI (ADHD)
Intervention agent	Special education teacher		Teacher	Teachers
Implementation fidelity	—		Procedural fidelity checklist	Procedural fidelity checklist

(continued)

Table 1. (continued)

Study element	2015 Xin	2016 Clarke	2017 Adamson
Internal validity			
Design	A-B-A-B	A-B-A-B	ATD
Baseline	Business as usual	Hand-raising condition	Business as usual
Outcome measures/DVs	On-task behavior, academic achievement, student satisfaction	Student responding and on-task behavior	Primary: AET Secondary: disruptive behavior
Social validity	Interviewed with 5 open-ended questions	Teacher Post-Intervention Acceptability and Importance of Effects Survey	Adapted Treatment Acceptability Rating Form from teacher and student perspectives
Study element	2017 Haydon		2017 Messenger
Description of practice	Choral responding + mnemonic device		Choral and mixed responding
Context and setting	7th-11th grade (social studies); urban Midwestern HS		4th grade (math); suburban Midwestern ES
Participants			Inclusive classroom: 21 students (11 boys)
Classroom	Class: 8 Students		At-risk internalizing behavior challenges and challenges working independently
Inclusion criteria	Class-wide; parent consent obtained to collect data;		2nd-4th grade girls (9-10 years; 1 White, 1 multiracial; 1 qualified LD during study)
Target students	4 students (13-15 years; all Black; all mild ID)		General educator
Intervention agent	Classroom teacher		Direct observation by outside observer and teacher self-report procedural fidelity checklist
Implementation fidelity	Direct measure of IVs using procedural fidelity checklist		
Internal validity			
Design	ATD		ATD
Baseline	NA		NA
Outcome measures/DVs	On-task behavior and correct responses		Active student responding and accuracy of responses
Social validity	Teacher and student surveys		Teacher: Intervention Rating Profile. Students: Children's Intervention Rating Profile

Note. A-B-A or A-B-A-B or A-B-A-B-C or A-B-C-B-C = withdrawal or reversal design; ADHD = attention deficit/hyperactivity disorder; AET = academic engaged time; ATD = alternating treatment design; CEI = critical events index (Walker & Severson, 1992); CFI = combined frequency index (Walker & Severson, 1992); DD = developmental delay; DO = direct observation; DV = dependent variable; EBD = emotional or behavioral disorder; ELL = English language learner; ES = elementary school; ESL = English as a second language; FRL = free and/or reduced lunch; GE = general education; HS = high school; ID = intellectual disability; LD = learning disability; MS = middle school; NA = nonapplicable; NHT = numbered heads together; OHI = other health impaired; OTR = opportunities to respond; RC = response cards; SCC = self-contained classroom; SED = serious emotional disturbance; SL = speech or language impairment; SSBD = Systematic Screening for Behavioral Disorders (Walker & Severson, 1992); — = not reported.

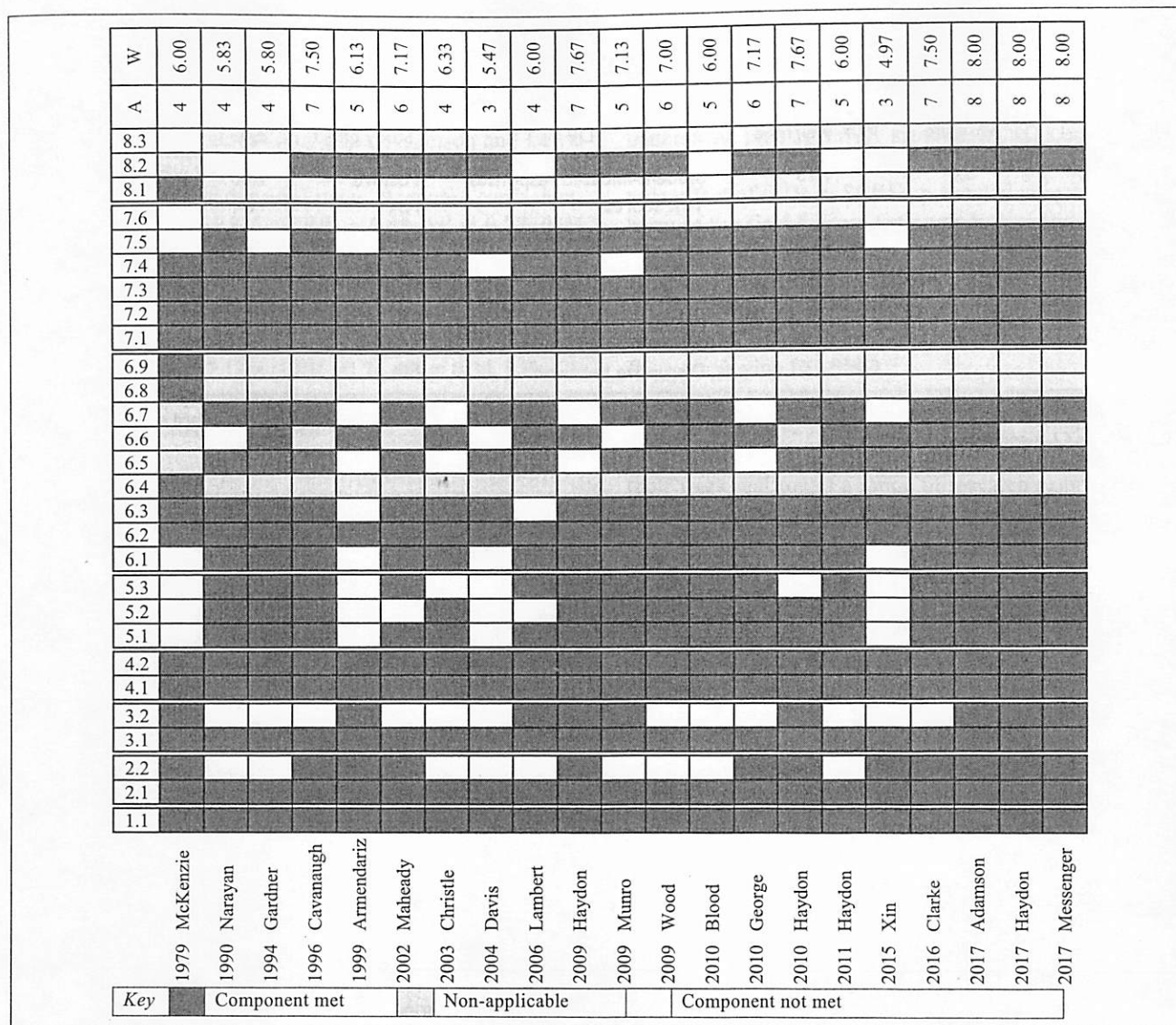


Figure 2. Scatter box plot of quality indicators (Council for Exceptional Children, 2014) of included studies. Note. A = absolute coding; W = weighted coding; 1.1 Context/Setting description; 2.1 Participant description; 2.2 Participant disability/at-risk status; 3.1 Role description; 3.2 Training description; 4.1 Intervention procedure description; 4.2 Materials description; 5.1 Implementation fidelity assessed/ reported; 5.2 Dosage or exposure assessed/ reported; 5.3 Assessed across relevant elements/ throughout study; 6.1 Independent variable (IV) systematically manipulated; 6.2 Baseline description; 6.3 No or limited access to IV during baseline; 6.4 Group assignment; 6.5 Three demonstrations of experimental effect; 6.6 Baseline: minimum three data points and established pattern; 6.7 Controls for threats to internal validity; 6.8 Overall attrition; 6.9 Group attrition 7.1 Socially important; 7.2 Description of dependent variable measures; 7.3 Reports effects on the intervention of all measures; 7.4 Measured repeatedly (minimum three data points per phase); 7.5 Adequate interobserver agreement; 7.6 Validity; 8.1 Data analytic techniques; 8.2 Graph clearly represents outcome; and 8.3 Effect sizes.

Between-case effect sizes. We calculated four BC-SMD estimates for three methodologically sound studies meeting the technological requirements (Clarke et al., 2016; Munro & Stephenson, 2009; Wood et al., 2009; see Table 2). A Grubbs’s test for outliers revealed that an effect size of 33.38 from Clarke et al. (2016) was an outlier ($G = 1.38, p = .16$). Upon visual inspection of normal probability plot, an effect size of 14.76 from Wood et al. (2009) was also identified as an outlier. We confirmed effect sizes were not an error and we dropped them from our analysis because the

large effect sizes were artifacts of near-zero responding to ceiling effect (Zelinsky & Shadish, 2018). Munro and Stephenson (2009) examined the effects of response cards, suggesting large effects, $BC-SMD = 2.60, SE = 1.45$; 95% confidence interval (CI): [1.10, 7.78], on student-initiated response opportunities. Wood et al. (2009) examined the effects of response cards, which also demonstrated large effects ($BC-SMD = 3.27, SE = 0.38$; 95% CI: [2.55, 4.05]) on student’s on-task behavior (originally coded off-task and reversed).

Table 2. Visual Analysis and Between-Case Standardized Mean Difference of Methodologically Sound Single-Case Design Studies With Three or More Cases Meeting Technical Requirements

Article	Quality indicators		Visual analysis		BC-SMD		
	Absolute	80% weighted	DV	Study effect	Est.	SE	95% CI
2009 Munro	5.0	7.13	Student-initiated responses	Positive	2.60	1.45	[1.10, 7.78]
			Test scores	NA	NA	NA	NA
2009 Wood	6.0	7.00	On-task (off-task reversed)	Positive	3.27	0.38	[2.55, 4.05]
			Participation	Positive			<i>Dropped: outlier</i>
2016 Clarke	7.0	7.50	Active student responding	Positive			<i>Dropped: outlier</i>
			On-task behavior	NA	NA	NA	NA
2017 Adamson	8.0	8.00	Academic engaged time	Positive	NA	NA	NA
			Disruptive behavior	NA	NA	NA	NA

Note. BC-SMD = between-case standardized mean difference effect size; DV = dependent variable; Est. = estimate; SE = standard error; NA = nonapplicable; CI = confidence interval.

Table 3. Within-Case Effect Sizes of Methodologically Sound Studies With Three or More Cases Meeting Technical Requirements of LRR.

Article	Dependent variable	Design: Contrast/case (phase)	LRR Est.	SE	95% CI	% change
2009 Munro	Student-initiated responses	AB _k : Alice, Leo	<i>Dropped: zero responding in baseline, <3 data</i>			
		AB _k : Brenda (average across)	<i>Dropped: outlier</i>			
		AB _k : Sam (average across)	1.32	0.08	[1.17, 1.47]	274.34
		AB _k : Nicky (average across)	1.26	0.06	[1.14, 1.38]	252.54
2009 Wood	Participation On-task (off-task reversed)	AB _k : Morgan, Thomas, Adam, & Valerie	<i>Dropped: near-zero responding/ceiling effect</i>			
		AB _k : Morgan (average across)	1.21	0.31	[0.60, 1.82]	235.35
		AB _k : Thomas (average across)	0.98	0.31	[0.38, 1.58]	166.45
		AB _k : Adam (average across)	0.79	0.28	[0.25, 1.32]	120.34
2016 Clarke	Active student responding	AB _k : Valerie (average across)	0.86	0.22	[0.43, 1.28]	136.32
		AB _k : Brandy, Ramona, Destiny, Danny	<i>Dropped: near-zero responding/ceiling effect</i>			
2017 Adamson	Academic engaged time	ATC: A ₁ to RC/S1	<i>Dropped: outlier</i>			
		ATC: RC to CWPT/ S1	0.16	0.20	[-0.22, 0.55]	17.35
		ATC: RC to GN/S1	0.45	0.27	[-0.09, 0.98]	56.83
		ATC: A ₁ to RC/S2	1.21	0.46	[0.32, 2.11]	235.35
		ATC: RC to CWPT/S2	0.24	0.33	[-0.41, 0.90]	27.12
		ATC: RC to GN/S2	0.53	0.43	[-0.32, 1.38]	69.89
		ATC: A ₁ to RC/S3	1.19	0.50	[0.22, 2.16]	228.71
		ATC: RC to CWPT/S3	0.14	0.40	[-0.65, 0.93]	15.03
		ATC: RC to GN/S3	0.64	0.43	[-0.20, 1.48]	89.65

Note. LRR = log response ratio effect size; Est. = estimate; SE = standard error; CI = confidence interval; AB_k = withdrawal/reversal design; ATC = active student responding; A₁ = baseline; RC = response card; CWPT = class-wide peer-tutoring; GN = guided notes; S_k = Student 1, Student 2, or Student 3.

Within-case effect sizes. We considered 16 LRR estimates from four methodologically sound studies meeting the technological requirements to calculate LRR. An additional nine cases were dropped for not meeting the technical requirements due to near-zero responding/ceiling effect, and one case was dropped for having fewer than three data points within an A-B contrast (see Table 3). A Grubbs's test for outliers revealed that an effect size of 1.75 from Munro and Stephenson (2009) was an outlier ($G = 1.65, p = .71$). Upon visual inspection of normal probability plot, an effect

size of 1.72 from Adamson and Lewis (2017) was also identified as an outlier. We confirmed effect sizes were not an error and dropped them from our analysis because the large effect sizes were artifacts of near-zero responding to ceiling effect. Munro and Stephenson (2009) examined effects of response cards, which demonstrated between 252.54% (LRR = 1.26, SE = 0.06, 95% CI: [1.14, 1.38]) and 274.34% (LRR = 1.32, SE = 0.08, 95% CI: [1.17, 1.47]) increase in student-initiated responding between hand-raising and response cards for two cases. Wood et al. (2009)

examined effects of response cards, which demonstrated between 120.34% (LRR = 0.79, $SE = 0.28$, 95% CI: [0.25, 1.32]) and 235.35% (LRR = 1.21, $SE = 0.31$, 95% CI: 0.60, 1.82) increase in on-task behavior between hand-raising and response cards across four cases. Adamson and Lewis (2017) compared effects of response cards against class-wide peer-tutoring and guided notes. Response cards demonstrated between 56.83% (LRR = 0.45, $SE = 0.27$, 95% CI: [-0.09, 0.98]) and 89.65% (LRR = 0.64, $SE = 0.43$, 95% CI: [-0.20, 1.48]) increase when compared against guided notes across three cases. Response cards demonstrated between 15.03% (LRR = 0.14, $SE = 0.40$, 95% CI: [-0.65, 0.93]) and 27.12% (LRR = 0.24, $SE = 0.33$, 95% CI: [-0.41, 0.90]) increase when compared against class-wide peer-tutoring across three cases.

Discussion

Increasing students' OTR is a high-leverage practice teachers can use to facilitate school success by increasing student engagement and decreasing challenging behavior (Lane et al., 2015; McLeskey et al., 2017). Across included studies, teachers implemented OTR as a part of general classroom management (e.g., Armendariz & Umbreit, 1999), as well as to offer additional support to students at-risk for EBD (e.g., Haydon et al., 2010; Messenger et al., 2017). Using *Standards for EBP*, we employed a modified definition for methodologically sound studies (i.e., 80% or more weighted criterion; Lane et al., 2009) to identify articles that were sufficiently methodologically rigorous (Common et al., 2017). Across 21 studies, 11 studies met our criterion for being methodologically sound, with three studies meeting 100% of CEC's QIs across components. Findings of this review indicated the majority (52.38%) of studies examining OTR strategies (e.g., choral responding, clickers, response cards) were methodologically rigorous.

Four methodologically sound studies with three or more cases ($n = 17$) examined the effectiveness of OTR strategy—specifically response cards—and demonstrated positive effects on student-initiated responses (Munro & Stephenson, 2009), on-task behavior (Wood et al., 2009), active student responding (Clarke et al., 2016), and academic engaged time (Adamson & Lewis, 2017). Thus, teacher-delivered OTR strategies, and specifically response cards, meet *Standards for EBPs* requirements as a potentially EBP following a modified definition of rigor. We note teacher-delivered OTR strategies would have been classified as having insufficient evidence had we followed CEC's (2014) definition of rigor.

We also calculated ES_{BC} and ES_{wc} for methodologically sound SCRD studies with three or more cases. Although three of the four studies met the methodological qualitative appraisals and technical standards for BC-SMD, two estimates were dropped from analysis for being outliers.

Similarly, three of the four studies met the methodological qualitative appraisals and technical standards for LRR, with two estimates also being dropped from analysis for being outliers. These results are similar to other studies showing patterns of inflated effect sizes in SCRD (Barton et al., 2017; Zelinsky & Shadish, 2018). Examining the magnitude effect of OTR in SCRDs presented unique challenges because the field has not yet come to consensus on which ES_{BC} and ES_{wc} should be employed. Specifically, few SCRD effect sizes are designed for use with ATD, and floor to ceiling effects failed to meet the technical requirements of many cases screened for LRR and produced inflated ES_{BC} employing BC-SMD.

Overall, results from this EBP review are similar to Schnorr et al.'s (2015) review in which the authors found sufficient support for response cards as an EBP. We extended their work and found a range of research examining OTR strategies (e.g., response cards, choral responding) to be methodologically rigorous and effective. We also examined the magnitude effect of teacher-delivered OTR across methodologically sound studies with three or more cases, of which those meeting the technical requirements of BC-SMD and LRR were large and in a therapeutic direction. These findings are consistent with our visual analysis.

Limitations and Future Directions

We encourage consideration of the following limitations and recommendations for future research when interpreting these findings. First, OTR strategies include a broad range of teacher-, peer-, and technologically mediated practices. We evaluated the evidence base of teacher-driven strategies to increase students' OTR. Future reviews are needed to examine the methodological quality of peer-mediated and technologically mediated OTR strategies. In addition, more research is needed to examine the empirical support for choral responding, clickers, and varied modes of responding. OTR strategy is particularly well suited for promoting fluency and automaticity in content knowledge, which are associated with increased gains in engagement, academic achievement, and desired student behaviors. Future research is needed to explore such student outcome effects across varying configurations (e.g., effects of student and classroom characteristics as well as various OTR strategies).

Second, in this review, we included all ATD to map the literature, including quality appraisal of the methodological rigor. Unlike demonstration designs (e.g., AB_K and MBD), which specifically examine the efficacy of the IV, ATDs are comparison designs, which address questions related to which IV is more effective (Ledford & Gast, 2018). As such, two ATD that compared different—albeit similar—IVs considered to be an OTR strategy were dropped from our evaluation of the evidence base (Haydon et al., 2017; Messenger et al., 2017). Furthermore, BC-SMD and LRR

were initially designed for demonstration designs, although Zelinsky and Shadish (2018) posited comparison designs might be adopted to allow A-B contrasts. Effect sizes for SCRD employing comparison designs should be interpreted with caution, as more research is needed to examine their theoretical and technical constraints.

Third, CEC's (2014) *Standards for EBPs* classifies SCRDs as having positive, neutral or mixed, or negative effects based in part by "the number and proportion of participants in a study for whom a functional relationship between IV and the DV was established" (p. 7). In this review, a number of methodologically sound studies reported aggregate class-wide data (e.g., Cavanaugh, Heward, & Donelson, 1996; George, 2010) and were excluded from visual analysis and further consideration in supporting the evidence base. Future research is necessary to explore the extent to which SCRDs reporting aggregate data should be included for visual inspection and considered when classifying the evidence base of instructional practices.

Fourth, whereas BC-SMD is theoretically on the same scale as Cohen's *d* and Hedges's *g*, early work has shown effect sizes for SCRD (i.e., ES_{BC} , ES_{wc}) to be much larger than *d* or *g* in group studies (Barton et al., 2017). This is consistent with theoretical expectations. For example, visual analysis of SCRD allows for the detection of large effects better than small or moderate effects. Thus, there is a strong publication preference for studies with larger effects (Barton et al., 2017; Shadish et al., 2016). More research is needed to examine the extent to which BC-SMD, *d*, and *g* are truly on the same scale, and, if so, whether they should be interpreted similarly or differently within systematic reviews that examine the magnitude effect across group comparison and SCRD research articles.

Finally, in this systematic review we included only studies published in peer-reviewed journals. As such, generalization may be limited due to the omission of theses, dissertations, and other studies that may have included null outcomes and/or included methodological decisions that may have prevented their publication. Also, although all studies reported effects across outcome measures, not all student outcome measures were graphed in original studies (e.g., reporting outcomes in tabular form). This led to the exclusion of some outcome data for further visual and statistical analysis. Furthermore, several graphs were dropped from analysis due to underreporting and inconsistent reporting of information necessary to interpret the data. Future researchers should ensure all necessary information is reported in studies to facilitate synthesis in systematic reviews.

Summary

Our goal was to classify the evidence base of teacher-delivered strategies to increase students' OTR across the K-12 continuum during whole-class instruction. We applied

Standards for EBPs (2014) utilizing a modified definition to identify methodologically sound studies. Eleven studies met or exceeded our modified criterion of 80% or more of the QIs. Five of these studies included three or more cases ($n = 21$) and demonstrated positive effects. Effect sizes demonstrated large magnitude effects in the therapeutic direction. We, therefore, classify teacher-directed OTR strategy in K-12 school settings to be a *potentially EBP*.

Declaration of Conflicting Interests

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

Funding

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

References

- Note.* References marked with an asterisk are included in the literature review.
- *Adamson, R. A., & Lewis, T. J. (2017). A comparison of three-opportunities-to-respond strategies on the academic engaged time among high school students who present challenging behavior. *Behavioral Disorders, 42*, 41–51. doi:10.1177/0198742916688644
 - *Armendariz, F., & Umbreit, J. (1999). Using active responding to reduce disruptive behavior in a general education classroom. *Journal of Positive Behavior Interventions, 1*, 152–158. doi:10.1177/109830079900100303
 - Barton, E. E., Pustejovsky, J. E., Maggin, D. M., & Reichow, B. (2017). Technology-aided instruction and intervention for students with ASD: A meta-analysis using novel methods of estimating effect sizes for single-case research. *Remedial and Special Education, 38*, 371–386. doi:10.1177/0741932517729508
 - *Blood, E. (2010). Effects of student response systems on participation and learning of students with emotional and behavioral disorders. *Behavioral Disorders, 35*, 214–228. doi:10.1177/019874291003500303
 - Brophy, J., & Good, T. L. (1986). Teacher behavior and student achievement. In Wittrock M. C. (Ed.), *Handbook of research on teaching* (3rd ed.; pp. 328–375). New York, NY: MacMillan.
 - *Cavanaugh, R., Heward, W., & Donelson, F. (1996). Effects on response cards during lesson closure on the academic performance of secondary students in an earth science course. *Journal of Applied Behavior Analysis, 29*, 403–406. doi:10.1901/jaba.1996.29-403
 - *Christle, C. A., & Schuster, J. W. (2003). The effects of using response cards on student participation, academic achievement, and on-task behavior during whole-class, math instruction. *Journal of Behavioral Education, 12*, 147–165.
 - *Clarke, L. S., Haydon, T., Bauer, A., & Epperly, A. C. (2016). Inclusion of students with an intellectual disability in the general education classroom with the use of response cards.

- Preventing School Failure*, 60, 35–42. doi:10.1080/1045988x.2014.96680
- Common, E. A., Lane, K. L., Pustejovsky, J. E., Johnson, A. H., & Johl, L. E. (2017). Functional assessment-based interventions for students with or at-risk for high-incidence disabilities: Field-testing single-case synthesis methods. *Remedial and Special Education*, 38, 331–352. doi:10.1177/0741932517693320
- 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, 220–234. doi:10.1177/0741932514557271
- Cook, B. G., Cook, S. C., & Collins, L. W. (2016). Terminology and evidence-based practice for students with emotional and behavioral disorders: Exploring some devilish details. *Beyond Behavior*, 25(2), 4–13. doi:10.1177/107429561602500202
- Council for Exceptional Children. (1987). *Academy for effective instruction: Working with mildly handicapped students*. Reston, VA: Author.
- Council for Exceptional Children. (2014). *CEC standards for evidence-based practices in special education*. Arlington, VA: Author.
- *Davis, L. L., & O'Neill, R. E. (2004). Use of response cards with a group of students with learning disabilities including those for whom English is a second language. *Journal of Applied Behavior Analysis*, 37, 219–222. doi:10.1901/jaba.2004.37-219
- Downer, J. T., Rimm-Kaufman, S. E., & Pianta, R. C. (2007). How do classroom conditions and children's risk for school problems contribute to children's behavioral engagement in learning? *School Psychology Review*, 36, 413–432.
- Ennis, R. P., Royer, D. J., Lane, K. L., & Griffith, C. E. (2017). A systematic review of precorrection in PK-12 settings. *Education and Treatment of Children*, 40, 465–495. doi:10.1353/etc.2017.0021
- Every Student Succeeds Act (2015). S.1177—114th Congress (2015-2016).
- *Gardner, R. G., Heward, W. L., & Grossi, T. A. (1994). Effects of response cards on student participation and academic achievement: A systematic replication with inner-city students during whole-class science instruction. *Journal of Applied Behavior Analysis*, 27, 63–71. doi:10.1901/jaba.1994.27-63
- *George, C. L. (2010). Effects of response cards on performance and participation in social studies for middle school students with emotional and behavioral disorders. *Behavioral Disorders*, 35, 200–213. doi:10.1177/019874291003500302
- Gersten, R., Fuchs, L. S., Compton, D., Coyne, M., Greenwood, C., & Innocenti, M. S. (2005). Quality indicators for group experimental and quasi-experimental research in special education. *Exceptional Children*, 71, 149–164. doi:10.1177/001440290507100202
- Greenwood, C. R., Delquadri, J. C., & Hall, R. V. (1989). Longitudinal effects of classwide peer tutoring. *Journal of Educational Psychology*, 81, 371–383.
- *Haydon, T., Conroy, M. A., Scott, T. M., Sindelar, P. T., Barber, B. R., & Orlando, A. M. (2010). A comparison of three types of opportunities to respond on student academic and social behaviors. *Journal of Emotional and Behavioral Disorders*, 18, 27–40. doi:10.1177/1063426609333448
- *Haydon, T., & Hunter, W. (2011). The effects of two types of teacher questioning on teacher behavior and student performance: A case study. *Education & Treatment of Children*, 34, 229–245. doi:10.1353/etc.2011.0010
- *Haydon, T., Mancil, G. R., & Van Loan, C. (2009). Using opportunities to respond in a general education classroom: A case study. *Education & Treatment of Children*, 32, 267–278. doi:10.1353/etc.0.0052
- Haydon, T., Marsicano, R., & Scott, T. M. (2013). A comparison of choral and individual responding: A review of the literature. *Preventing School Failure*, 57, 181–188. doi:10.1080/1045988x.2012.682184
- *Haydon, T., Musti-Rao, S., & Alter, P. (2017). Comparing choral responding and a choral responding plus mnemonic device during geography lessons for students with mild to moderate disabilities. *Education & Treatment of Children*, 40, 77–95. doi:10.1080/1045988x.2012.682184
- Hedges, L. G., Pustejovsky, J., & Shadish, W. R. (2012). A standardized mean difference effect size for single-case designs. *Research Synthesis Methods*, 3, 224–239. doi:10.1002/jrsm.1052
- Hedges, L. G., Pustejovsky, J., & Shadish, W. R. (2013). A standardized mean difference effect size for multiple baseline designs across individuals. *Research Synthesis Methods*, 4, 324–341. doi:10.1002/jrsm.1086
- Horn, C. (2010). Response cards: An effective intervention for students with disabilities. *Education and Training in Autism and Developmental Disabilities*, 45, 116–123.
- 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, 165–179. doi:10.1177/001440290507100203
- Komsta, L. (2011). Tests for outliers. R package (Version 0.14). Retrieved from <https://CRAN.R-project.org/package=outliers>
- *Lambert, M. C., Cartledge, G., Heward, W. L., & Lo, Y. Y. (2006). Effects of response cards on disruptive behavior and academic responding during math lessons by fourth-grade urban students. *Journal of Positive Behavior Interventions*, 8, 88–99. doi:10.1177/10983007060080020701
- Lane, K. L., Common, E. A., Royer, D. J., & Muller, K. (2014). *Group comparison and single-case research design quality indicator matrix using Council for Exceptional Children 2014 standards*. Unpublished tool. Retrieved from <http://www.ci3t.org/practice>
- Lane, K. L., Kalberg, J. R., & Shepcaro, J. C. (2009). An examination of the evidence base for function-based interventions for students with emotional and/or behavioral disorders attending middle and high schools. *Exceptional Children*, 75, 321–340. doi:10.1177/001440290907500304
- Lane, K. L., Menzies, H. M., Ennis, R. P., & Oakes, W. P. (2015). *Supporting behavior for school success: A step-by-step guide to key strategies*. New York, NY: Guilford Press.
- Ledford, J. R., & Gast, D. L. (2018). *Single case research methodology: Applications in special education and behavioral sciences* (3rd ed). New York, NY: Routledge.
- MacSuga-Gage, A. S., & Simonsen, B. (2015). Examining the effects of teacher-directed opportunities to respond on student outcomes: A systematic review of the literature. *Education*

- & *Treatment of Children*, 38, 211–239. doi:10.1353/etc.2015.0009
- Maggin, D. M., Lane, K. L., & Pustejovsky, J. E. (2017). Introduction to the special issue on single-case systematic reviews and meta-analyses. *Remedial and Special Education*, 38, 323–330. doi:10.1177/0741932517717043
- Maheady, L., Mallette, B., Harper, G. F., & Sacca, K. (1991). Heads together: A peer-mediated option for improving the academic achievement of heterogeneous learning groups. *Remedial and Special Education*, 12, 25–33. doi:10.1177/074193259101200206
- *Maheady, L., Michielli-Pendl, J., Mallette, B., & Harper, G. F. (2002). A collaborative research project to improve the academic performance of a diverse sixth grade science class. *Teacher Education and Special Education*, 25, 55–70. doi:10.1177/08884064020250010
- *McKenzie, G. R., & Henry, M. (1979). Effects of testlike events on on-task behavior, test anxiety, and achievement in a classroom rule-learning task. *Journal of Educational Psychology*, 71, 370–374. doi:10.1037//0022-0663.71.3.370
- McLeskey, J., Barringer, M.-D., Billingsley, B., Brownell, M., Jackson, D., Kennedy, M., & Ziegler, D. (2017, January). *High-leverage practices in special education*. Arlington, VA: Council for Exceptional Children & CEEDAR Center.
- *Messenger, M., Common, E. A., Lane, K. L., Oakes, W. P., Menzies, H. M., Cantwell, E. D., & Ennis, R. P. (2017). Increasing opportunities to respond for students with internalizing behaviors: The utility of choral and mixed responding. *Behavioral Disorders*, 42, 170–183. doi:10.1177/0198742917712968
- Montague, M., & Bergeron, J. (1997). Using prevention strategies in general education. *Focus on Exceptional Children*, 29, 1–12. doi:10.17161/fec.v29i8.6754
- *Munro, D. W., & Stephenson, J. (2009). The effects of response cards on student and teacher behavior during vocabulary instruction. *Journal of Applied Behavior Analysis*, 42, 795–800. doi:10.1901/jaba.2009.42-795
- *Narayan, J. S., Heward, W. L., Gardner, R., Courson, F. H., & Omness, C. K. (1990). Using response cards to increase student participation in an elementary classroom. *Journal of Applied Behavior Analysis*, 23, 483–490. doi:10.1901/jaba.1990.23-483
- Pustejovsky, J. E. (2015). Measurement-comparable effect sizes for single-case studies of free-operant behavior. *Psychological Methods*, 20, 342–359. doi:10.1037/met0000019
- Pustejovsky, J. E. (2016). *scdhlm: A web-based calculator for between-case standardized mean differences* (Version 0.3.1). Retrieved from <https://jepusto.shinyapps.io/scdhlm>
- Pustejovsky, J. E. (2017). *Single-case effect size calculator* (Version 0.2) Web application. Retrieved from <https://jepusto.shinyapps.io/SCD-effect-sizes/>
- Pustejovsky, J. E. (2018). Using response ratios for meta-analyzing single-case designs with behavioral outcomes. *Journal of School Psychology*, 68, 99–112. doi:10.1016/j.jsp.2018.02.003
- Randolph, J. J. (2007). Meta-analysis of the research on response cards: Effects on test achievement, quiz achievement, participation, and off-task behavior. *Journal of Positive Behavior Interventions*, 9, 113–128. doi:10.1177/1098300707009002020
- Rohatgi, A. (2015). *WebPlotDigitizer* (Version 3.12). Retrieved from <http://arohatgi.info/WebPlotDigitizer>
- Sainato, D. M., Strain, P. S., & Lyon, S. R. (1987). Increasing academic responding of handicapped preschool children during group instruction. *Journal of the Division for Early Childhood*, 12, 23–30. doi:10.1177/105381518701200104
- Schnorr, C. I., Freeman-Green, S., & Test, D. W. (2015). Response cards as a strategy for increasing opportunities to respond: An examination of the evidence. *Remedial and Special Education*, 37, 41–54. doi:10.1177/074193251557561
- Shadish, W. R., Hedges, L. V., Homer, R. H., & Odom, S. L. (2015). *The role of between-case effect size in conducting, interpreting, and summarizing single-case research (NCER 2015-002)*. Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education.
- Shadish, W. R., Zelinsky, N. A. M., Vevea, J. L., & Kratochwill, T. R. (2016). A survey of publication practices of single-case design researchers when treatments have small or large effects. *Journal of Applied Behavior Analysis*, 49, 656–673.
- Spencer, T. D., Detrich, R., & Slocum, T. A. (2012). Evidence-based practice: A framework for making effective decisions. *Education and Treatment of Children*, 35, 127–151.
- Stichter, J. P., Lewis, T. J., Whittaker, T. A., Richter, M., Johnson, N. W., & Trussell, R. P. (2009). Assessing teacher use of opportunities to respond and effective classroom management strategies: Comparisons among high- and low-risk elementary schools. *Journal of Positive Behavior Interventions*, 11, 68–81. doi:10.1177/109830070832659
- Sutherland, K. S., & Wehby, J. H. (2001). Exploring the relationship between increased opportunities to respond to academic requests and the academic and behavioral outcomes of students with EBD: A review. *Remedial and Special Education*, 22, 113–121. doi:10.1177/074193250102200205
- 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*. Oslo, Norway: The Campbell Collaboration. doi:10.4073/cmdp.2016.1
- Walker, H. M., & Severson, H. (1992). *Systematic screening for behavior disorders: Technical manual*. Longmont, CO: Sopris West.
- *Wood, C. L., Mabry, L. E., Kretlow, A. G., Lo, Y.-Y., & Galloway, T. W. (2009). Effects of preprinted response cards on students' participation and off-task behavior in a rural kindergarten classroom. *Rural Special Education Quarterly*, 28, 39–47.
- *Xin, J. F., & Johnson, M. L. (2015). Using clickers to increase on-task behaviors of middle school students with behavior problems. *Preventing School Failure*, 59, 49–57. doi:10.1080/1045988x.2013.823593
- Zelinsky, N. A., & Shadish, W. (2018). A demonstration of how to do a meta-analysis that combines single-case designs with between-groups experiments: The effects of choice making on challenging behaviors performed by people with disabilities. *Developmental Neurorehabilitation*, 21, 266–278. doi:10.3109/17518423.2015.1100690