Single-Case Synthesis Tools I: Comparing Tools to Evaluate SCD Quality and Rigor

Kathleen N. Zimmerman[1], Jennifer R. Ledford[1], Katherine E. Severini[1], James E. Pustejovsky[2], Erin E. Barton[1], & Blair P. Lloyd[1]

[1]Vanderbilt University [2]University of Texas at Austin

January 2, 2018. Forthcoming in *Research in Developmental Disabilities*. This paper is not the version of record and may not exactly replicate the final, published version of the article. The final article will be available, upon publication, at https://doi.org/10.1016/j.ridd.2018.02.003

Author acknowledgements: The research reported in this article was supported in part by Grant R305D160002 from the Institute of Educational Sciences, U.S. Department of Education. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

Abstract

Tools for evaluating the quality and rigor of single case research designs (SCD) are often used when conducting SCD syntheses. Preferred components include evaluations of design features related to the internal validity of SCD to obtain quality and/or rigor ratings. Three tools for evaluating the quality and rigor of SCD (Council for Exceptional Children, What Works Clearinghouse, and Single-Case Analysis and Design Framework) were compared to determine if conclusions regarding the effectiveness of antecedent sensory-based interventions for young children changed based on choice of quality evaluation tool. Evaluation of SCD quality differed across tools, suggesting selection of quality evaluation tools impacts evaluation findings. Suggestions for selecting an appropriate quality and rigor assessment tool are provided and across-tool conclusions are drawn regarding the quality and rigor of studies. Finally, authors provide guidance for using quality evaluations in conjunction with outcome analyses when conducting syntheses of interventions evaluated in the context of SCD.

Keywords: quality evaluation, rigor evaluation, single case research, sensory-based interventions

Single-Case Synthesis Tools I: Comparing Tools to Evaluate SCD Quality and Rigor

Guidelines for identifying evidence based practices (EBPs) have been established with the purpose of improving outcomes for children with disabilities in educational environments. Single case research designs (SCDs) are one experimental research method that can be used to evaluate if an intervention is an EBP (What Works Clearinghouse, 2014; Shadish, Hedges, Horner, & Odom, 2015). SCDs offer rigorous, experimental evaluations of intervention effectiveness, and unlike group design research, there is no agreed upon method to evaluate them in the context of EBP reviews (Ledford, Wolery, & Gast, 2014; Maggin, O'Keefe, & Johnson, 2011; Shadish et al., 2008 2015; Wolery, Busick, Reichow, & Barton, 2010). SCDs have often been excluded from EBP reviews due to the historical use of visual analysis as a method to determine intervention effectiveness rather than quantitative metrics found in group designs such as effect sizes (Shadish et al., 2015).

Hierarchical frameworks that separately evaluate internal validity then study outcomes have been suggested for use (Maggin, Briesch, Chafouleas, Ferguson, & Clark, 2014; O'Keefe, Slocum, Burlingame, Snyder, & Bundock, 2012; Wendt & Miller, 2012) to ensure outcomes are considered relative to experimental rigor. Evaluations of quality may be used to describe the weaknesses and strengths of the designs included in a synthesis, determine study inclusion or exclusion in quantitative evaluations of outcomes, and explain variability in outcomes across studies (Pustejovsky & Ferron, 2017).

Quality and rigor tools provide guidelines for systematically evaluating studies. In 2005, Horner and colleagues provided the first widely-used set of quality indicators for SCD, including guidelines related to reporting (e.g., participant and implementer demographics, setting, procedures), internal validity (e.g., potential demonstrations of effect, procedural fidelity, reliability), and outcomes. Since then, professional organizations, federal agencies, and other researchers (Cook et al., 2014; Kratochwill et al., 2013; What Works Clearinghouse, 2014; see Maggin et al., 2014 for review) have developed additional tools to evaluate design quality. Most frameworks require assessments based on dichotomous responses (CEC; Cook et al., 2014); others include dichotomous responses followed by categorical ratings with three potential responses (e.g., fewer than three data points per condition, at least three but fewer than five, five or more; Kratochwill et al., 2013). The *Single-Case Analysis and Review Framework* (SCARF) includes analysis of quality and rigor and provides graphic displays of results via a scatterplot of single study visual analysis outcomes (rated on a 0-4 scale) alongside quality and rigor scores (also on a 0-4 scale; Authors, 2016).

Given the breadth of tools available, it is not surprising that variable content exists across tools (Maggin et al., 2014; O'Keefe et al., 2012; Wendt & Miller, 2012). To address this variability, critical components of quality evaluation tools have been proposed (Maggin et al., 2014; O'Keefe et al., 2012; Wendt & Miller, 2012). First, researchers suggest use of hierarchical tools—including those that evaluate rigor prior to evaluating outcomes (Maggin et al., 2014; O'Keefe et al., 2012; Wendt & Miller, 2012). Researchers have also suggested unequal weighting of indicators, based on whether or not the design feature should prohibit interpretation of outcomes (Gersten et al., 2005; Maggin et al., 2014; Wendt & Miller, 2012). Finally, researchers propose authors should include clear directions for interpreting final evaluation scores when creating tools (Maggin et al., 2014; Wendt & Miller, 2012).

Despite some agreement regarding use of SCD synthesis tools, researchers disagree about which components are critical (Maggin et al., 2014; O'Keefe et al., 2012; Wolery, 2013). As of 2012, Wendt and Miller suggested that an "ideal quality tool" had not been developed (Wendt &

Miller, 2012, p. 259), although at least two frameworks have been proposed for use since that conclusion was drawn (Authors 2016; Goldstein, Lackey, & Schneider, 2014) and the commonly used What Works Clearinghouse guidelines have been updated. Given disagreements about critical features, differential results across reviews may be attributable to the tools used to evaluate intervention effects (Chen, Hyppa-Martin, Reichle, & Symons, 2016; Maggin et al., 2014; Maggin et al., 2011; O'Keefe et al., 2012; Wendt & Miller, 2012).

Synthesis of Antecedent Sensory-Based Interventions

Sensory-based interventions (SBI) are interventions designed to support individuals' abilities to process environmental sensory input, based on the theory of sensory integration (Blanche, Chang, Guiterrez, & Gunter, 2016). Proponents of SBI argue sensory dysfunction may increase maladaptive behaviors, decrease self-regulation, and negatively impact learning (Ayres, 1979), but there is limited empirical evidence to support these claims (American Academy of Pediatrics [AAP], 2012). Interventions based on Ayres' (1979) sensory integration theory vary extensively (Parham et al., 2011). For the purpose of this review, authors define *antecedent* SBI as any tools or materials (e.g., weighted vests, therapy cushions, therapy balls, and weighted blankets; Blanche et al., 2016) designed to be implemented *before an expected behavior is to occur* (antecedent) to provide an individual support in processing environmental sensory input in an instructional setting. These interventions may not adhere to current protocols defined by the Ayres Sensory Integration® program (Roley, Mailloux, Miller-Kuhaneck, & Glennon, 2007).

Seven systematic reviews, literature reviews, or meta-analyses (some including SCD and group design studies [e.g., RCTs]) have been published on the topic in peer-reviewed journals from 2008 through 2015 (Barton, Reichow, Schnitz, Smith, & Sherlock, 2015; Case-Smith, Weaver, & Fristad, 2015; Lang et al., 2012; Leong, Carter, & Stephenson, 2015; May-Benson &

Koomar, 2010; Watling & Hauer, 2015; Yunus, Liu, Bissett, & Penkala, 2015). All reviews evaluated the quality of articles, although conclusions regarding the quality and outcomes varied across reviews. One review concluded SBI were effective (May-Benson & Koomar, 2010), three reviews concluded SBI were ineffective (Barton et al., 2015; Lang et al., 2012; Leong et al., 2015), and three reviews yielded mixed results (Case-Smith, Weaver, & Fristad, 2015; Watling & Hauer, 2015; Yunus, et al., 2015).

Disparate results across syntheses might be confusing for families and educational personnel regarding the use of resources and expectations for treatment, which might increase stress. For example, families of children with autism spectrum disorders (ASD) often request the use of SBI (May-Benson & Koomar, 2010) and report that they prefer it as a method of treatment for occupational therapy (Goin-Kochel, Myers, & Mackintosh, 2007), despite a lack of consistently positive outcomes in research. Also, the AAP (2012) issued a policy statement highlighting the lack of evidence supporting sensory processing disorders and interventions. However, families spend an estimated $16,000 on SBI over the course of a child’s lifetime (Zane, Davis, & Rosswurm, 2008). An evaluation of synthesis tools might be useful to elucidate what contributes to the variability in synthesis conclusions.

Purpose and Research Questions

The purpose of this review is to evaluate differences among synthesis tools and to assess results derived from the tools related to SCD studies evaluating antecedent SBI. Two commonly used tools (Cook et al., 2014; WWC, 2014) as well as a new tool (SCARF; Authors 2016) were selected. The SCARF was selected because it is a novel, hierarchical tool that allows for weighting of indicators (Wendt & Miller, 2012). Furthermore, the tool provides a unique quantitative outcome with graphical display and the ability to visually analyze outcomes across studies. The following research questions guided the study: (1) Do conclusions regarding the quality and rigor of SCD differ across quality evaluation tools? (2) How does the composition of each tool contribute to variability in conclusions? and (3) To what extent have antecedent SBI been evaluated via rigorous SCD studies?

Method

Search Procedures

A systematic search was conducted to identify peer-reviewed articles, reviews, and gray literature. A structured electronic search using PsycINFO and ERIC was conducted with the following search terms for the intervention, participants, and outcomes. Each string of search terms was connected with the word ‘and’ and Boolean operators were used (Borenstein, Hedges, Higgins, & Rothstein, 2009). Intervention terms included: *environmental arrangement*, *materials manipulation*, *curricular revision*, *antecedent intervention*, *antecedent functional analysis*, *structural analysis*, *antecedent-based intervention*, or *antecedent function-based intervention*, *intervention*, or *sensory-based*. Participant terms included: *young child*, *children*, *preschool*, *elementary*, *Head Start*, *early intervention*, *day care*, *pre-kindergarten*, *kindergarten*, *first grade*, *second grade*, or *primary*. Outcome terms included *problem behavior*, *challenging behavior*, *prosocial behavior*, *interaction*, or *social*. Authors did not restrict the search by publication date. Additionally, a search of published online conference proceedings from the past two years from two early childhood conferences (*Division for Early Childhood of the Council for Exceptional Children* and *Conference on Research Innovations in Early Childhood Special Education*) and one behavior analysis conference (*Association of Behavior Analysis International*) was conducted. Reference lists of reviews and meta-analyses were forward- and backward-searched to identify additional articles.

Article Eligibility and Selection

Inclusion criteria were: (a) participants 8 years or younger *or* in preschool or kindergarten-2nd grade, (b) intervention included changing physical materials, (c) use of an antecedent sensory-based material (including weighted vests, stability balls, therapy cushions, or materials identified as "multi-sensory"), and (d) use of an SCD. Material manipulation could include addition or removal of a physical material or change in format of a material in the environment. Participant disability status, intervention setting, outcome measure, implementer, publication year, or publication language were not considered for inclusion. Articles were excluded if the intervention involved: (a) non-material manipulation intervention components or (b) consequence-based components that were in effect during only one condition (e.g., reinforcement for target behaviors in only the intervention condition). Designs were defined as stand-alone SCD (e.g., alternating treatments design [ATD], A-B-A-B, multiple baseline across participants); thus designs could include one (e.g., A-B-A-B) or more participants (multiple baseline across participants). If multiple SCD were reported in a source, only designs that met inclusion criteria were included (e.g., if an article included three designs, two could be excluded from the review if only one included an eligible participant). Any designs with an eligible participant were included in the review. Abstracts and then full texts were screened for eligibility, consistent with PRISMA guidelines (Moher, Liberati, Tetzlaff, & Altman, 2009), as outlined in Figure 1.

Coding Procedures

The first author— a doctoral student trained in SCD methodology, special education, and applied behavior analysis— served as the primary coder. Initially author, publication year, article title, and all experimental comparisons were coded. Then, descriptive information was coded, including: number of potential demonstrations of effect, participant demographics, setting, dependent variables, measurement procedures, intervention description, procedures, reliability, fidelity, social validity, maintenance, and generalization (coding manual available from first author). Demonstrations of effect were defined as the total number of potential condition changes across two adjacent conditions repeated over time. This criterion was used to evaluate if a SCD was an *experimental* SCD (Gast, Ledford, & Severini, 2018), defined as a design with at least three potential demonstrations of effect (e.g., A-B-A-B but not A-B-A). Finally, articles were coded using each evaluation tool in the order of specificity of tools as determined by the first and second authors (described below).

SCD Quality Evaluation Tools

CEC. The CEC created evaluation standards for EBPs in special education specific to SCD (Cook et al., 2014). The CEC tool was based on previous quality indicators (Gersten et al., 2005; Horner et al., 2005) and includes items related to both single case and group design studies. The CEC tool evaluates research at the article (manuscript) level—for example, an article including three A-B-A-B designs is evaluated as a single unit. The presence of at least three demonstrations of effect is assessed, but it is not used as a gating criterion (i.e., all studies are evaluated for all indicators). Eight general indicators are evaluated as present or absent (yes/no): context and setting, participants, intervention agent, practice description, implementation fidelity, internal validity, outcome measures/dependent variables, and data analysis (Cook et al., 2014). All indicators hold equal weight, and final decisions about the effectiveness of an intervention are classified into three categories: positive, neutral, or mixed effects (Cook et al., 2014). An intervention is classified as an EBP if there are (a) five articles that meet all quality indicators with positive effects and 20 participants and (b) there are no articles that meet all quality indicators with negative effects and (c) “at least a 3:1 ratio” of articles that met all quality indicators with positive effects to articles with neutral/mixed effects (Cook et al., 2014, p. 211).

WWC. What Works Clearinghouse (WWC, 2014) created a tool for evaluating SCD to assist in identifying EBPs (Kratochwill et al., 2013) as a complement to previously created group design standards (WWC, 2008). The tool is based on quality indicators originally identified by Horner and colleagues (2005). Designs are the units of analysis—for example, an article including three A-B-A-B designs is evaluated as three separate units. WWC uses a hierarchical framework across three items with yes/no responses for each (independent variable manipulation, reliability, and number of demonstrations) and two items with categorical responses (number of data points; outcomes). To proceed to the second level of the tool, the following three indicators *all* must be met: manipulation of the independent variable, collection of reliability data in each condition for 20% of measurement occasions with at least 80% agreement (0.6 kappa), and at least three potential demonstrations of effect. If a design meets the three aforementioned indicators, then the number of data points are evaluated as the second level of the hierarchical framework to determine whether a design *does not meet* WWC standards, *meets standards with reservations,* or *meets standards without reservations* (WWC, 2014). Finally, outcomes are evaluated for designs that meet standards with or without reservations. Outcomes are classified in three categories: no, moderate, or strong evidence of effectiveness. An intervention is classified as an EBP using SCD if positive outcomes are present for at least 5 studies (meeting WWC standards with or without reservations), with at least 20 participants, conducted by at least 3 different research teams (Kratochwill et al., 2013; WWC, 2014).

SCARF. The *Single-Case Analysis and Design Framework* (Authors, 2016) is a synthesis tool used to evaluate SCD on the design level using a hierarchical framework (cf. Authors, 2017). Designs are only evaluated if at least three potential demonstrations of effect are present. Scores of 0-4 are possible in 10 categories; they are generated via responses to several yes/no questions for 7 categories (data sufficiency, reliability, fidelity, social and ecological validity; condition, participant, and dependent variable descriptions) and are generated via several yes/no questions and 0-4 categorical ratings for 3 categories (maintenance, response generalization, and stimulus generalization measurement. For example, a score of 0 is provided for reliability if reliability data are not measured; a score of 1 is provided if it was not measured for at least 20% of sessions in both conditions; a score of 2 indicates it was measured in 20% or more of each condition; a score of 3 indicates at least 80% agreement in addition to requirements for a score of 2; and a score of 4 indicates requirements for 3 were met *and* data collectors were blind to study condition (see *website redacted for peer review* for questions across categories).

The ten categories are divided into *rigor* (dependent variable reliability, procedural fidelity, and sufficiency of data) and *quality/breadth of measurement* (social and ecological validity, participant and condition descriptions, dependent variables, response and stimulus generalization, and maintenance). The total rigor and quality score is calculated such that rigor scores are weighted twice compared to quality scores. Following rigor and quality assessments, outcomes are assessed separately for primary, maintained, and generalized effects, also on a 0-4 scale. Results are presented in a scatterplot with average design quality and rigor on the x-axis and the outcomes score on the y-axis. An intervention is classified as an EBP using SCD if the majority of studies are high quality with evidence of positive effects (top right quadrant of the scatterplot) and no high quality studies that show evidence of negative or null effects are present (lower right quadrant of the scatterplot; Authors 2016; Authors, 2017). Additional information on calculations of study rigor, quality, and outcomes can be found at (website redacted for review).

Reliability Calculations

Reliability data were collected for at least 33% of units of analysis for each tool by a trained second coder who was a graduate student in special education (third author). Interobserver agreement was calculated using point-by-point agreement (number of agreements/total number of data points) multiplied by 100 (Ayres & Ledford, 2014). Average overall agreement coding by study was 93% for general study characteristics (range 75-100%), 84% for CEC coding (range 75-92%), 97% for WWC coding (range 78-100%), and 95% for SCARF coding (range 78-100%). Coders resolved discrepancies via consensus; most resulted from coder errors (e.g., failing to follow tool directions). Seven disagreements not resolved via consensus (one each for CEC and WWC tools; five for SCARF) were resolved via a third independent coder who developed the SCARF tool and was a WWC certified reviewer in SCD (i.e., completed a rigorous training process with WWC, including an assessment demonstrating mastery of the standards and their applications to SCD; second author).

Results

General Study Characteristics

Eleven articles (published manuscripts) including 69 designs (SCD) and 51 participants across 10 research teams were included in the review; each article included at least 2 designs (e.g., two A-B-A-B designs in a single manuscript) and between 1 and 20 participants. Two articles were dissertations (Krombach, 2016; Olson, 2015); all other articles were published in peer-reviewed journals. Sixteen A-B-A-B designs were used in four articles (Hodgetts, Magill-Evans, & Misiaszek, 2011; Olson, 2015; Tunson & Candler, 2010; Umeda & Deitz, 2011), 33 alternating treatments designs (ATDs) were used in two articles (Cox, Gast, Luscre, & Ayres, 2009; Reichow, Barton, Sewell, Good, & Wolery, 2010), 4 multiple baseline across participants designs (MBPs) were used in two articles (Krombach, 2016; Leew, Stein, & Gibbard, 2010), 14 A-B-C designs were used in two articles (Bagatell, Miriglianai, Patterson, Reyes, & Test, 2010; Kane, Luiselli, Dearborn, & Young, 2004), and 2 A-B-A designs were used in one article (Reichow, Barton, Good, & Wolery, 2009). Twenty-seven participants were diagnosed with ASD, 19 were typically developing, 2 were diagnosed with developmental delays, 2 were diagnosed with intellectual disabilities, and 1 had an unreported disability. Ten participants were female and ages ranged from 2-12 years; one article did not report participant ages but reported participants were in kindergarten to first grade (Bagatell et al., 2010). Race and socio-economic status of participants were not reported in any article. Various interventions were assessed: weighted vests (Cox et al., 2009; Hodgetts et al., 2011; Kane et al., 2004; Leew et al., 2010; Reichow et al., 2009; Reichow et al., 2010), therapy balls (Bagatell et al., 2010; Krombach, 2016; Olson, 2015), a therapy cushion (Umeda & Deitz, 2011), and a multi-sensory environment (Tunson & Candler, 2010). Participant and intervention characteristics are displayed in Table 1.

SCD Quality Evaluation Tools

CEC. All 11 articles were included in the CEC analysis (see Table 2) and results are displayed by article in Table 3. Articles met an average of 53% of indicators (range 14-86%) and no articles met all criteria. All measured socially important outcomes and all but one reported graphs depicting results of all dependent variables (Bagatell et al., 2010). Neutral or mixed effects were present in all studies except one (Krombach, 2016), which reported positive effects. The most common deficiencies included intervention dosage; intervention agent training and role; evidence of systematic manipulation of the intervention and lack of access to treatment during baseline; and using direct, observational measures to assess fidelity. Because none of the articles met all quality indicators, insufficient evidence existed to classify the interventions.

WWC. Sixty-nine designs in 11 articles were included in the WWC analysis (see Table 2); results are displayed by design in Table 4. Eight designs did not meet standards due to inadequate reliability data; 4 did not include an adequate number of potential demonstrations of effect; 7 did not include at least three data points per condition; and 14 failed to meet multiple criteria. Ten designs met standards with reservations and 26 met standards without reservations. Moderate evidence of effects was present in five designs (Hodgetts et al., 2011; Krombach, 2016; Olson, 2015; Reichow et al., 2010); all other designs that met design standards demonstrated no evidence of effects (31 designs). Overall, antecedent SBI were not identified to be an EBP for improving outcomes for young children as assessed via this tool.

SCARF. Fifty-one designs in eight articles were included in the SCARF analysis (see Table 2); 18 designs were excluded for failing to include three potential demonstrations of effect. Results by item are displayed in Table 5; graphic displays of outcomes and quality/rigor scores are shown in Figure 2. All designs measured outcomes in typical environments, but few designs measured social validity (Cox et al., 2009; Krombach, 2016; Olson, 2015; Umeda & Deitz, 2011). For all designs, scores of 0 (possible range 0-4) were given for all indicators of stimulus and response generalization, maintenance, and fidelity measurement. The average primary effect across studies was 0.14 out of 4 (range 0–4) with only 3 of 51 designs scoring higher than 0 (Krombach, 2016; Reichow et al., 2010; Tunson & Candler, 2010). Most designs (n= 34) demonstrated low quality evidence and minimal or negative effects (bottom left quadrant in Figure 2); some (n = 15) demonstrated high quality evidence and minimal or negative effects (bottom right quadrant in Figure 2); 2 designs demonstrated low quality evidence and positive effects (top left quadrant in Figure 2; Krombach, 2016; Reichow et al., 2010); and no designs demonstrated high quality evidence and positive effects (top right quadrant in Figure 2). Overall, antecedent SBI were not identified as an EBP for improving outcomes for young children due to relatively few high quality studies, all of which showed negative or no effects.

Evaluation Across Tools

The three tools consistently concluded antecedent SBI were not an EBP for improving outcomes for young children (Table 2). The average primary effect was 0.1 (possible 0 – 4) using SCARF. One article (9%) included positive effects based on CEC indicators. No designs showed strong effects using WWC and 19% of designs (5 designs in 4 articles) showed moderate effects. The positive effect (CEC) and 3 of 5 moderate effects (WWC) were evaluated in non-peer-reviewed studies.

Associations between quality ratings across tools were generally in agreement (Figure 3), although some variability in scores was present (Table 2). There was general agreement on the worst and best quality studies across all tools (Bagatell et al., 2010; Cox et al., 2009). The study with the fewest CEC quality indicators received a score of 0 on WWC (did not meet standards) and SCARF (not evaluated; Bagatell et al., 2010). The study with the greatest number of CEC quality indicators received 1s and 2s from WWC (bottom panel, Figure 3) and the highest relative ratings on SCARF (Cox et al., 2009; top panel, Figure 3). Twenty-two designs in five articles did not meet the threshold for acceptable quality and rigor across all tools (Bagatell et al., 2010; Kane et al., 2004; Reichow et al., 2009; Tunson & Chandler, 2010; Umeda & Deitz, 2011). All designs identified by WWC as failing to meet standards also received scores less than 2 on the SCARF except one (Reichow et al., 2010; middle panel, Figure 3). Thirteen designs met standards using WWC, but were rated as low quality by CEC (failing to meet all quality indicators) and SCARF (scores less than 2; Hodgetts et al., 2011; Krombach, 2016; Leew et al., 2010; Olson, 2015; Reichow et al., 2010).

Differences in weighting schemes and the unit of analysis contributed to variable results across tools. Inadequate descriptions of intervention agents and dosage were captured by all tools, but variability in the weighting schemes for indices contributed to differential conclusions in one article (5 designs; Cox et al., 2009). Only WWC and SCARF captured within-article variability due to differences in the unit of analysis between tools. Measures of the sufficiency of data across designs in an article contributed to differential conclusions across tools (27 designs; Reichow et al., 2010). The SCARF determination was that more data were needed in some designs to draw confident conclusions; the WWC determination was that the minimum number of data points had not been reached for some designs but due to a numerical minimum rather than variability in data patterns. Additionally, within-study variations did not align between WWC and SCARF for one article (4 designs; Umeda & Deitz, 2011) due to inflexible gating procedures for reliability. Authors failed to report if reliability data collection occurred across conditions; all designs did not meet standards using WWC, but were given some credit for reporting overall reliability averages by SCARF. Continued analysis using SCARF found differential quality between the four designs in the article due to data instability in 3 of the 4 designs (overall quality/rigor scores 1.1-1.8).

Discussion

The results of the current review of quality and rigor evaluation tools provide important implications for SCD quality and rigor evaluation as well as antecedent SBI.

Differences in Quality Conclusions

Despite all three tools leading to the same general conclusion with respect to EBP classification, the current review confirms conclusions of previous reviews that different synthesis tools can produce dissimilar conclusions (Maggin et al., 2014; O'Keefe et al., 2012; Wendt & Miller, 2012). Among tools, ratings related to potential demonstrations of effect, number of data points, and procedural fidelity resulted in differential results. Moreover, the use of gating or weighting also influenced outcomes. Each of these is described below.

WWC and SCARF tools each evaluate potential demonstrations of effect and number of data points as a critical gating quality feature. Both require three potential demonstrations of effect, a common contemporary standard (Gast & Baekey, 2014); the WWC further requires five demonstrations in ATD designs. Both require a sufficient number of data points and potential demonstrations to be present before outcomes are evaluated, but WWC requires a fixed number of data points and SCARF evaluates data sufficiency using visual inspection of stability. Thus, the tools offer different interpretations for high quality designs with stable data paths but fewer than five data points. For example, a design with three stable data points at 0% correct responding would meet standards with reservations using the WWC tool (more than 3 but fewer than 5 data points), but could receive the highest possible rating on the SCARF (no additional data needed due to stability at floor levels).

The use of a fixed minimum standard versus use of visual analysis to evaluate data sufficiency has been debated (Ledford, Lane, & Tate, 2018; Wolery, 2013). Data from this analysis suggest using a fixed minimum standard versus visual analysis to determine data sufficiency may impact decisions regarding the inclusion of studies in SCD syntheses. The fixed minimum standard of three or five data points, absent visual analyses of data sufficiency, results in a lower rigor rating that might be arbitrary. For example, a design with three stable data points in baseline might receive a lower rigor rating than a design with five variable data points in baseline using the WWC tool. Whereas, the three stable data points in baseline design might receive a higher score than the five variable data points in baseline design due to meeting a fixed minimum of three data points *and* visual analysis of data sufficiency using the SCARF tool.

Inconsistencies and errors in quality/rigor evaluations or exclusions of studies with satisfactory rigor may lead consumers to conclude that effective interventions have insufficient evidence. The exclusion of quality studies may be particularly problematic when evaluating behaviors expected to be at zero levels before intervention (e.g., academic responding). Thus, quality/rigor tool selection may also be driven by the type of outcomes analyzed in an SCD synthesis. Outcomes expected to be at near-zero levels before intervention (e.g., skill acquisition) or variable levels across conditions (e.g., problem behavior) may be best evaluated using a tool with data sufficiency rules using visual analysis (SCARF) rather than fixed minimums (WWC) to ensure stable responding was achieved at adequate levels. More comparisons of tools using fixed minimum versus visual analysis standards to evaluate data sufficiency may assist researchers in understanding if tool selection could increase the likelihood of Type II error when identifying EBPs.

Procedural fidelity data, considered by some to be a critical feature of high-quality SCD research (Horner et al., 2005, 2012; Tate et al., 2016; Wolery & Lane, 2014), is not evaluated using the WWC tool. Four designs that met standards for WWC received low quality scores for SCARF due to an absence of fidelity measurement. Several researchers have considered the absence of fidelity measurement as a critical gating feature a considerable weakness of the WWC tool (Maggin et al., 2014; Wolery, 2013).

Although the absence of fidelity measurement resulted in low scores on SCARF for some studies that met criteria for WWC, both tools used gating or weighting procedures when calculating overall quality and rigor. The CEC tool assesses potential demonstrations of effect, data sufficiency, and fidelity, but weights these features equally to other study characteristics such as descriptions of intervention agents and settings. The absence of weighting the critical design components led to differential findings in this review for the CEC tool. Weighting and gating tools that require designs to include critical quality elements may assist researchers in identifying the variability in important quality elements between designs in an SCD review. Variability in meaningful quality elements that do not prohibit the interpretation of outcomes (e.g., participant and setting descriptions; materials descriptions) can still be detected, but will not prohibit the evaluation of potentially meaningful outcomes. Moreover, critical flaws (e.g., insufficient number of potential demonstrations) will be identified as such.

Antecedent Sensory-based Interventions

Antecedent SBI were not identified as an EBP by any tool. Over 50% of evaluations of antecedent SBI were conducted in SCD that failed to meet the acceptable quality and rigor thresholds of all three tools. Only one article received acceptable quality ratings from at least two tools; none received acceptable quality ratings from all three tools. Moreover, most outcomes were not positive regardless of quality rating. When outcomes were positive, they almost exclusively occurred in unpublished dissertations (Krombach, 2016; Olson, 2015). The results of this review are consistent with those of previous reviews published in special education journals (Barton et al., 2015; Lang et al., 2012; Leong et al., 2015), including reviews evaluating randomized control trials (Barton et al., 2015), indicating SBI are more likely to be ineffective than effective for improving outcomes for young children. Given recent findings indicating long-term academic and social outcomes for individuals with ASD (the disability group represented in the highest proportion) are poor (Steinhausen, Jensen, & Lauristen, 2016) and the results of this review, we do not recommend their use with young children.

Limitations and Future Research

Several limitations should be considered when interpreting the results of this review. First, the review included a relatively small body of literature (11 articles). However, the total number of possible units of analysis was 69 designs and two of three tools required assessment at the design level. Future comparisons of synthesis tools should be conducted with larger bodies of literature to evaluate if the results of this review are broadly applicable. Bodies of literature with mostly positive or heterogeneous outcomes should be evaluated. Furthermore, although half of the peer-reviewed studies included in the review were published in occupational therapy journals, the scope of the literature evaluated during the article selection process may have been limited by the fact that we did not search medically-based electronic databases or online conference proceedings from occupational therapy conferences. Although outside the scope of this review, future reviews evaluating SBI may also consider evaluating the effectiveness of interventions with and without occupational therapist implementers.

Additionally, the same reviewers coded across all tools; we attempted to control for carryover by coding with less-specific tools prior to more-specific tools, but it is possible coding across tools led to differential assessment. Two coders (first and second author) contributed to the development of the SCARF. The feasibility of application by unaffiliated research groups is currently being investigated; there is some evidence the tool can be applied by other researchers. Additional future applications of the tool by other researchers may provide insights into its usability. Final reports regarding the effectiveness of the intervention should include evaluations of both the quality of the evidence *and* the magnitude of the effects. Although each tool incorporates components of visual analysis to evaluate outcomes, quantitative measures are also frequently recommended in addition to visual analysis (e.g., overlap measures and SCD effect sizes; Shadish et al., 2015). A comparison of tools to quantitatively evaluate outcomes using the antecedent SBI literature are evaluated in a companion paper to this review, *Single-Case Synthesis Tools II: Comparing Quantitative Outcome Measures for Synthesizing Single Case Designs* (Authors, 2017).

Implications for Researchers

The results of this review suggest researcher selection of quality and rigor evaluation tools may contribute to variability in SCD synthesis conclusions. Although overall EBP determination was not different across tools in this review, the analysis of quality in individual designs was highly variable. It may be that EBP conclusions might also vary by tool in different groups of studies, given that many included studies in this review had null effects, which are relatively rare in published research (Shadish, Zelinsky, & Vevea, 2016; Sham & Smith, 2014). Additional research is needed to determine the extent to which EBP conclusions or conclusions regarding the likelihood of positive outcomes vary based on tool.

True variability in design quality is expected to explain differential results rather than tool selection. Subsequently, researchers should consider how the attributes of a tool may have contributed to conclusions (Maggin et al., 2014; Wendt & Miller, 2012). When conducting syntheses of interventions evaluated via SCD, we offer the following recommendations for using quality evaluations in conjunction with outcome analyses:

- When using the CEC quality indicators, we suggest consumers include assessment at the *design* (SCD study) level to identify (a) design features that contributed to an article meeting or not meeting criteria and (b) within-study variability that may not be captured in the review. In addition, we suggest that researchers explicitly denote items that are necessary for establishing a functional relation as “critical” and others as “important”.

- When using the WWC standards, we suggest consumers consider using additional analyses of sufficiency of data and procedural fidelity measurement to identify if stable data patterns were present, particularly in designs with fewer than five data points, *and* if adequate fidelity measurement and reporting was present across studies.
- When using the SCARF tool, we suggest interpreting results with caution because the tool is new and has only been applied in one published study (Authors, 2017). Although the tool is not missing critical evaluation components (e.g., fidelity measurement), it would be prudent to use the SCARF tool in conjunction with another tool.

References

References marked with an asterisk (*) indicate studies included in the review.

American Academy of Pediatrics (2012). Sensory integration therapies for children with developmental and behavioral disorders. *Pediatrics*, *129*, 1186–1189.

Authors (2016, April). Single case analysis and review framework (SCARF). Retrieved from (website redacted for peer review).

Authors (2017). Single-Case Synthesis Tools II: Comparing Overlap Measures and Parametric Effect Sizes for Synthesizing Sensory-Based Interventions. *Manuscript under review*.

Authors (2017). Evidence for the effectiveness of social narratives: Students without ASD. *Journal of Early Intervention*.

Ayres, A. J. (1979). *Sensory integration and the child*. Los Angeles, CA: Western Psychological Services.

Ayres, K., & Ledford, J. R. (2014). Dependent measures and measurement systems. In D. L. Gast & J. R. Ledford (Eds.), *Single case research methodology: Applications in special education and behavioral sciences* (pp. 124-153). New York, NY: Routledge.

*Bagatell, N., Mirigliani, G., Patterson, C., Reyes, Y., & Test, L. (2010). Effectiveness of therapy ball chairs on classroom participation in children with autism spectrum disorders. *The American Journal of Occupational Therapy, 64*, 895-903.

Barnek, G. T., Watson, L. R., Boyd, B. A., Poe, M. D., David, F. J., & McGuire, L. (2013). Hyporesponsiveness to social and nonsocial sensory stimuli in children with autism, children with developmental delays, and typically developing children. *Development and Psychopathology*, *25*, 307-320.

Barton, E. E., Reichow, B., Schnitz, A., Smith, I. C., & Sherlock, D. (2015). A systematic review of sensory-based treatments for children with disabilities. *Research in Developmental Disabilities, 37*, 64-80.

Ben-Sasson, A., Hen, L., Fluss, R., Cermak, S. A., Engel-Yeger, B., & Gal, E. (2009). A meta-analysis of sensory modulation symptoms in individuals with autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *39,* 1–11.

Blanche, E. I., Chang, M. C., Gutierrez, J., & Gunter, J. S. (2016). Effectiveness of a sensory-enriched early intervention group program for children with developmental disabilities. *American Journal of Occupational Therapy*, *70*, 7005220010.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. West Sussex, England: John Wiley & Sons.

Boyd, B. A., Baranek, G. T., Sideris, J., Poe, M. D., Watson, L. R., Patten, E., et al. (2010). Sensory features and repetitive behaviors in children with autism and developmental delays. *Autism Research, 3*, 78–87.

Case-Smith, J., Weaver, L. L., & Fristad, M. A. (2015). A systematic review of sensory processing interventions for children with autism spectrum disorders. *Autism*, *19*, 133-148.

Chen, M., Hyppa-Martin, J. K., Reichle, J. E., & Symons, F. J. (2016). Comparing single case design overlap-based effect metrics from studies examining speech generating device interventions. *American Journal on Intellectual and Developmental Disabilities*, *121,* 169-193.

Cook, B., Buysse, V., Klingner, J. Landrum, T., McWilliam, R., Tankersley, M., & Test, D. (2014). Council for exceptional children: Standards for evidence-based practices in special education. *Teaching Exceptional Children, 46*, 206-212.

*Cox, A. L., Gast, D. L., Luscre, D., & Ayres, K. M. (2009). The effects of weighted vests on appropriate in-seat behaviors of elementary-age students with autism and severe to profound intellectual disabilities. *Focus on Autism and Other Developmental Disabilities, 24*, 17-26.

Detrich, R., & Lewis, T. (2012). A decade of evidence-based education: Where are we and where do we need to go? *Journal of Positive Behavior Interventions, 15*, 214-220.

Gast, D. L., & Baekey, D. H. (2014). Withdrawal and reversal designs. In D. L. Gast & J. R. Ledford (Eds.), *Single Case Research Methodology,* (pp. 211-250). New York, New York: Routledge.

Gast, D. L., Ledford, J. R., & Severini, K. E. (2018). Withdrawal and reversal designs. In J. R. Ledford & D. L. Gast (Eds.), *Single Case Research Methodology,* (pp. 215-238). New York, New York: Routledge.

Gersten, R., Fuchs, L. S., Compton, D., Coyne, M., Greenwood, C. R., & Innocenti, M. S. (2005). Quality indicators for group experimental and quasi-experimental research in special education. *Exceptional Children, 71*, 149-164.

Goin-Kochel, R. P., Myers, B. J., & Mackintosh, V. H. (2007). Parental reports on the use of treatments and therapies for children with autism spectrum disorders. *Research in Autism Spectrum Disorders, 1*, 195–209.

Goldstein, H., Lackey, K. C., & Schneider, N. J. B. (2014). A new framework for systematic reviews: Application to social skills interventions for preschoolers with autism. *Exceptional Children, 80*, 262-286.

*Hodgetts, S., Magill-Evans, J., & Misiaszek, J. E. (2011). Weighted vests, stereotyped behaviors and arousal in children with autism. *Journal of Autism and Developmental Disorders, 41*, 805-14.

Horner, R., Carr, E., 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*, 165-180.

Horner, R. H., Swaminathan, H., Sugai, G., & Smolkowski, K. (2012). Considerations for the systematic analysis and use of single-case research. *Education and Treatment of Children*, *35*, 269–290.

*Kane, A., Luiselli, J. K., Dearborn, S., & Young, N. (2004). Wearing a weighted vest as intervention for children with Autism/Pervasive developmental disorder: Behavioral assessment of stereotypy and attention to task. *The Scientific Review of Mental Health Practice: Objective Investigations of Controversial and Unorthodox Claims in Clinical Psychology, Psychiatry, and Social Work, 3*, 19-24.

Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskoff, D. M., & Shadish, W. R. (2013). Single-case intervention research design standards. *Remedial and Special Education, 34,* 26-38.

*Krombach, P. A. (2016). *The effects of stability ball seating on children with autism spectrum disorder.* Available from ProQuest Dissertations & Theses Global. (1781592012).

Lang, R., O'Reilly, M., Healy, O., Rispoli, M., Lydon, H., Streusand, W., Davis, T., Kang, S., …Sigafoos, J. (2012). Sensory integration therapy for autism spectrum disorders: A systematic review. *Research in Autism Spectrum Disorders, 6*, 1004-1018.

Leaf, J. B., Oppenheim-Leaf, M. L., Leaf, R. B., Taubman, M., McEachin, J., Parker, T., ... &

Mountjoy, T. (2015). What is the proof? A methodological review of studies that have utilized social stories. *Education and Training in Autism and Developmental Disabilities, 50*, 127.

Ledford, J. R., Lane, J. D., & Tate, R. (2018). Evaluating quality and rigor in single case design. In J. R. Ledford & D. L. Gast (Eds.), *Single Case Research Methodology,* (pp. 365-394). New York, New York: Routledge.

Ledford, J. R., Wolery, M., & Gast, D. L. (2014). Controversial and critical issues in single case research. In D. L. Gast & J. R. Ledford (Eds.), *Single Case Research Methodology,* (pp. 377-396). New York, New York: Routledge.

*Leew, S. V., Stein, N. G., & Gibbard, W. B. (2010). Weighted vests' effect on social attention for toddlers with autism spectrum disorders. *The Canadian Journal of Occupational Therapy, 77*, 113-24.

Leong, H. M., Carter, M., & Stephenson, J. (2015). Systematic review of sensory integration therapy for individuals with disabilities: Single case design studies. *Research in Developmental Disabilities, 47*, 334-351.

Maggin, D. M., Briesch, A. M., Chafouleas, S. M., Ferguson, T. D., & Clark, C. (2014). A comparison of rubrics for identifying empirically supported practices with single-case research. *Journal of Behavioral Education, 23*, 287-311.

Maggin, D. M., O'Keefe, B. V., & Johnson, A. H. (2011). A quantitative synthesis of methodology in meta-analysis of single-subject research for students with disabilities: 1985-2009. *Exceptionality, 19*, 109-135.

May-Benson, T. A., & Koomar, J. A. (2010). Systematic review of the research evidence examining the effectiveness of interventions using a sensory integrative approach for children. *American Journal of Occupational Therapy*, *64*, 403-414.

Miller, L. J., Anzalone, M. E., Lane, S. J., Cermak, S. A., & Osten, E. T. (2007). Concept evolution in sensory integration: A proposed nosology for diagnosis. *American Journal of Occupational Therapy, 61*, 135–140.

Moher D., Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). *P*referred *R*eporting *I*tems for *S*ystematic Reviews and *M*eta-*A*nalyses: The PRISMA Statement. Retrieved from http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1000097.

O'Keefe, B. V., Slocum, T. A., Burlingame, C., Snyder, K., & Bundock, K. (2012). Comparing results of systematic reviews: Parallel reviews of research on repeated reading. *Education and Treatment of Children, 35*, 333-366.

*Olson, N. A. (2015). *Investigating stability balls in the classroom: Effects on student behavior and academic productivity.* Available from ProQuest Dissertations & Theses Global. (1686537353).

Parham, D. L., Roley, S. S., May-Benson, T. A., Koomar, J., Brett-Green, B., Burke, J. P., Cohen, E. S., Mailloux, Z., Miller, L. J., & Scaff, R. C. (2011). Development of a fidelity measure for research on the effectiveness of the Ayres Sensory Integration intervention. *Department of Occupational Therapy Faculty Papers, Paper 24*. Retrieved from http://jdc.jefferson.edu/cgi/viewcontent.cgi?article=1020&context=otfp.

Pustejovsky, J. E., & Ferron, J. M. (2017). Research synthesis and meta-analysis of single-case designs. In J. M. Kauffman, D. P. Hallahan, & P. C. Pullen (Eds.), *Handbook of Special Education*, 2nd Edition. New York, NY: Routledge.

*Reichow, B., Barton, E. E., Good, L., & Wolery, M. (2009). Brief report: Effects of pressure vest usage on engagement and problem behaviors of a young child with developmental delays. *Journal of Autism and Developmental Disorders, 39*, 1218-21.

*Reichow, B., Barton, E. E., Sewell, J. N., Good, L., & Wolery, M. (2010). Effects of weighted vests on the engagement of children with developmental delays and autism. *Focus on Autism and Other Developmental Disabilities, 25*, 3-11.

Roley, S. S., Mailloux, Z., Miller-Kuhaneck, H., & Glennon, T. (2007). Understanding Ayres sensory integration®. *OT Practice, 12,* CE-1-CE-7.

Shadish, W. R., Hedges, L. V., Horner, R. H., & Odom, S. L. (2015). The role of between-case effect sizes 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. (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.

Sham, E., & Smith, T. (2014). Publication bias in studies of an applied behavior-analytic intervention: An initial analysis. *Journal of Applied Behavior Analysis, 47*, 663-678.

Steinhousen, J., Jensen, C., & Lauristen, M. (2016). A systematic review and meta-analysis of the long-term overall outcome of autism spectrum disorders in adolescence and adulthood. *Acta Psychiatrica Scandinavica, 133*, 445-452.

Tate, R. L., Perdices, M., Rosenkoetter, U., Shadish, W., Vohra, S., Barlow, D. H., …Wilson, B. (2016). The single-case reporting guidelines in behavioural interventions (SCRIBE) 2016 statement. *Journal of Clinical Epidemiology, 73*, 142-152.

*Tunson, J., & Candler, C. (2010). Behavioral states of children with severe disabilities in the multisensory environment. *Physical & Occupational Therapy in Pediatrics, 30*, 101-110.

*Umeda, C., & Deitz, J. (2011). Effects of therapy cushions on classroom behaviors of children with autism spectrum disorder. *The American Journal of Occupational Therapy*, *65*, 152-159.

Watling, R., & Hauer, S. (2015). Effectiveness of Ayres Sensory Integration® and sensory-based interventions for people with autism spectrum disorder: A systematic review. *American Journal of Occupational Therapy*, *69*, 1-8.

Wendt, O., & Miller, B. (2012). Quality appraisal of single-subject experimental designs: An overview and comparison of different appraisal tools. *Education and Treatment of Children*, *35*, 235-268.

What Works Clearinghouse. (2008). Procedures and standards handbook. (Version 2.0). Retrieved from http://ies.ed.gov/ncee/wwc/documentsum.aspx?sid=19

What Works Clearinghouse. (2014). Procedures and standards handbook. (Version 3.0). Retrieved from http://ies.ed.gov/ncee/wwc/pdf/reference_resources /wwc_procedures_v3_0_standards_handbook.pdf.

Wolery, M. (2013). A commentary: Single-case design technical document of the what works clearinghouse. *Remedial and Special Education*, *43*, 39-43.

Wolery, M., & Lane, K. L. (2014). Writing tasks: Literature reviews, research proposals, and final reports. In D. L. Gast & J. R. Ledford (Eds.), *Single Case Research Methodology*, (pp. 50-104). New York, New York: Routledge.

Wolery, M., Busick, M., Reichow, B., & Barton, E. E. (2010). Comparison of overlap methods for quantitatively synthesizing single-subject data. *The Journal of Special Education, 44*, 18–28.

Wong, C., Odom, S. L., Hume, K., Cox, A. W., Fettig, A., Kucharczyk, S., …Schultz, T. R.

(2014). Evidenced-based practices for children, youth, and young adults with Autism Spectrum Disorder. Retrieved from http://autismpdc.fpg.unc.edu/content/ebp-update.

Yunus, F. W., Liu, K. P. Y., Bissett, M., & Penkala, S. (2015). Sensory-based intervention for children with behavioral problems: A systematic review. *Journal of Autism and Developmental Disorders, 45*, 3565-3579.

Zane, T., Davis, C., & Rosswurm, M. (2008). The cost of fad treatments in autism. *Journal of Early and Intensive Behavior Intervention, 5*, 44-51.

Table 1

Participant and Intervention Characteristics

Author	Age	Gender	Disability	Materials	Setting	Implementer	Dependent Variable(s)
Bagatell 2010	NR	M	ASD	Therapy Ball	School	NR	Engagement
	NR	M	ASD				
	NR	M	ASD				
	NR	M	ASD				
	NR	M	ASD				
	NR	M	ASD				
Cox 2009	5	F	ASD, SLI	Weighted vest	School	Researcher	Engagement
	6	M	ASD, SLI				
Hodgetts 2011	8	M	ASD	Weighted vest	School	Classroom aide	Problem behavior
	6	M	ASD				
	3	F	ASD				
	5	M	ASD				
	4	M	ASD				
Kane 2004	8	M	ASD	Weighted vest	School	Therapist	Engagement Problem behavior
	8	M	PDD-NOS				
Krombach 2016	12	M	ASD	Therapy Ball	Home	Therapist	Engagement
	6	M	ASD				
	7	M	ASD				
	4	M	ASD				
Leew 2010	2	M	ASD	Weighted Vest	Home	Parent	Problem behavior Social skills
	2	M	ASD				
	2	M	ASD				
	2	M	ASD				
Olson 2015	7-8	M (11) F (9)	TD	Therapy Ball	School	Teacher Researcher	Engagement Unengagement Academic Task
Reichow 2009	4	M	DD	Weighted Vest	School	Researcher	Engagement Problem behavior
Reichow 2010	4	M	DD	Weighted Vest	Daycare	Researcher	Engagement Problem behavior
	5	M	ASD				
	5	M	ASD				
Tunson 2010	7	M	ID	Multisensory Environment	Hospital	Researcher	Environmental states
	3	M	ID				
Umeda 2011	5	M	ASD	Therapy Cushion	School	NR	Engagement
	6	M	ASD				

Note. Studies identified by first author. Age in years. M=male. F=female. ASD=autism spectrum disorder. SLI=speech language impairment. PDD-NOS: pervasive developmental disorder-not otherwise specified. TD=typically developing. DD=developmental delay. ID=intellectual disability. The Olson study included 20 participants; numbers indicated in parentheses in gender column.

Table 2

Study Design and Analysis

		Evaluation Conducted			Meets Quality/Rigor Standard			EBP Classification		
	Design	CEC	WWC	SCARF	CEC	WWC	SCARF	CEC	WWC	SCARF
Bagatell	ABC	•	•	-	-	-	-	Not EBP	Not EBP	Not EBP
Cox	ATD	•	•	•	-	•	•			
Hodgetts	ABAB	•	•	•	-	⊕	-			
Kane	ABC	•	•	-	-	-	-			
Krombach	MBP	•	•	•	-	•	-			
Leew	MBP	•	•	•	-	•	-			
Olson	ABAB	•	•	•	-	⊕	-			
Reichow[a]	ABA	•	•	-	-	-	-			
Reichow[b]	ATD	•	•	•	-	⊕	⊕			
Tunson	ABAB	•	•	•	-	-	-			
Umeda	ABAB	•	•	•	-	-	-			

Note. Articles are indicated by first author. a=Reichow et al., 2009. b=Reichow et al., 2010. • = present. - = absent. ⊕ = some designs met standard in article, but not all. EBP classification based on SCD classification guidelines only; group design studies were not included in the review.

Table 3

Council of Exceptional Children Results

Components	Bagatell	Cox	Hodgetts	Kane	Krombach	Leew	Olson	Reichow[a]	Reichow[b]	Tunson	Umeda	% of Articles
Context and setting		●			●			●	●			36
Participant demographics	●	●	●	●	●	●		●	●	●	●	91
Participant disability status		●	●	●		●				●	●	55
Intervention agent role			●			●					●	27
Intervention agent training							●					9
Description of procedures		●	●	●	●	●			●	●		64
Description of materials		●	●	●	●	●		●	●	●	●	82
Adherence to procedures		●	●								●	27
Intervention dosage			●									9
Fidelity: measurement		●										9
Systematic IV manipulation		●	●									18
Description of baseline		●	●			●			●	●		5
Access to treatment		●									●	18
Demonstrations of effect		●			●	●	●		●	●	●	64
Internal validity: Baseline	●	●		●	●			●	●		●	64
Threats to internal validity		●		●	●		●		●	●	●	64
Socially important outcomes	●	●	●	●	●	●	●	●	●	●	●	100
Outcomes measurement		●			●	●	●	●	●	●	●	73
Outcome data reporting		●	●	●	●	●	●	●	●	●	●	91
Dependent variable		●	●	●	●		●			●	●	64
Interobserver reliability	●	●	●		●	●	●	●	●		●	82
Data analysis		●	●	●	●	●	●	●	●	●	●	91
% of indicators	14	86	64	45	59	55	41	41	59	55	68	
Positive effects					●							9
Negative effects												0
Neutral/mixed effects	●	●	●	●		●	●	●	●	●	●	91

Note. Percentage of indicators calculated as a total number of indicators present divided by the total number of indicators possible. ● = indicator present. a=Reichow et al., 2009. b.=Reichow et al., 2010.

Table 4

What Works Clearinghouse Results

Author	Number of Designs	Sufficient Reliability Data	Sufficient Demonstrations of Effect	Design Standards Conclusion			Outcomes			
				DNM	Meets with Reservations	Meets	Not Evaluated	None	Mod	Strong
Bagatell	10	0	0	10	0	0	10	0	0	0
Cox	6	6	5	0	2	4	0	6	0	0
Hodgetts	5	5	3	2	1	2	2	2	1	0
Kane	4	0	0	4	0	0	4	0	0	0
Krombach	2	2	2	0	0	2	0	0	2	0
Leew	2	2	2	0	0	2	0	2	0	0
Olson	5	3	3	2	3	0	2	2	1	0
Reichow[a]	2	0	0	2	0	0	2	0	0	0
Reichow[b]	27	27	27	7	4	16	7	19	1	0
Tunson	2	0	0	2	0	0	2	0	0	0
Umeda	4	0	0	4	0	0	4	0	0	0
TOTAL	69	45	42	33	10	26	33	31	5	0

Note. Totals represent the number of designs in each article that met criteria for each index. All designs met criteria that researchers systematically manipulated independent variable; thus, these data are not presented above. IOA=interobserver agreement. DNM=does not meet. Mod=moderate. a=Reichow et al., 2009. b.=Reichow et al., 2010. Criteria for SCD Design Standards and Outcomes can be found at https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_procedures_v3_0_standards_handbook.pdf.

Table 5

Single Case Analysis and Review Framework Results

		Rigor				Quality and Breadth of Measurement									Outcomes		
Author	Number of Designs	Reliability	Fidelity	Sufficient Data	**Total Rigor Score**	Ecological/Social Validity	Participant Descriptions	Condition Descriptions	DV Descriptions	SG Measurement	Response Generalization	Maintenance Measurement	**Total Quality Score**	**Combined Rigor and Quality Score**	Primary Effect	Maintained Effect	Generalized Effect
Cox	6	3	3	1-4	**2.3-3.3**	2	4	1	3	0	0	0	**1.4**	**2.0-2.7**	0	0	0
Hodgetts	3	1	0	4	**1.7**	1	4	0	2	0	0	0	**1**	**1.4**	0	0	0
Krombach	2	2	0	0	**0.7**	2	4	0	2	0	0	0	**1.1**	**0.8**	0-1	0	0
Leew	2	1	0	0	**0.3**	1	3	1	4	0	0	0	**1-1.7**	**0.8**	0	0	0
Olson	5	0-1	0-1	1-2	**0.3-1.3**	3	0	0-1	4	0	0	0	**1-1.1**	**0.6-1.3**	0	0	0
Reichow[b]	27	3	0	1-4	**1.3-2.3**	1	3	4	4	0	0	0	**1.7**	**1.5-2.1**	0-4	0	0
Tunson	2	0	0	2	**0.3**	1	2	3	3	0	0	0	**1.3**	**0.7**	0-1	0	0
Umeda	4	1	1	1-4	**1-2**	2	4	0	3	0	0	0	**1.3**	**1.1-1.8**	0	0	0
Average	4.6	2.3	0.5	2.6	**1.8**	1.4	3.0	2.2	3.3	0	0	0	**1.5**	**1.7**	0.1	0	0

Note: Number of designs includes those with at least three demonstrations of effect. Designs with fewer than three (e.g., A-B-A designs) are not evaluated. Each design received a score; ranges refer to the minimum and maximum scores of designs within a single article. a=Reichow et al., 2009. b.=Reichow et al., 2010. DV=dependent variable. SG=stimulus generalization. Total rigor, quality, and combined score equations can be obtained from website redacted for review.

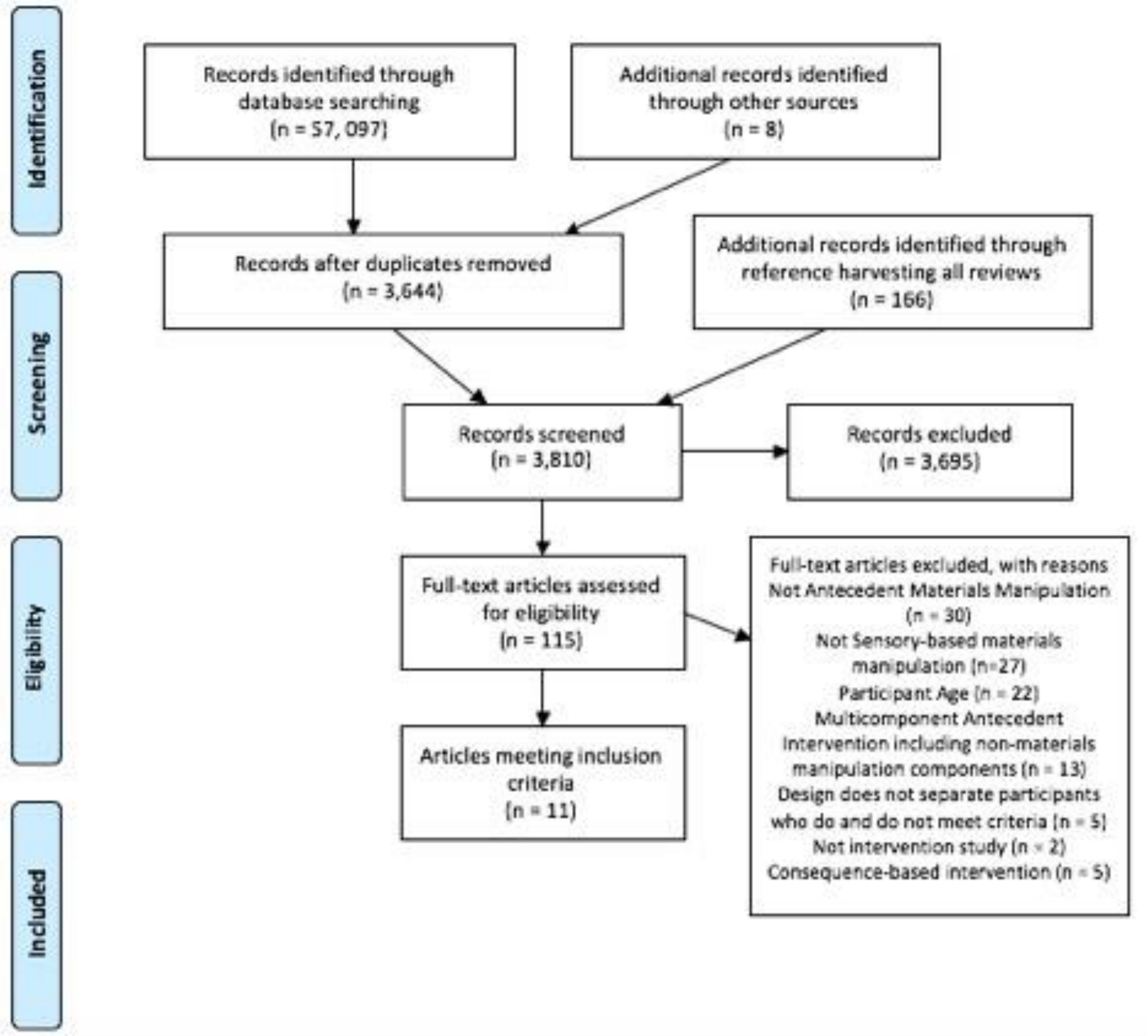
Identification
Screening
Eligibility
Included
Records identified through database searching (n = 57, 097)
Additional records identified through other sources (n = 8)
Records after duplicates removed (n = 3,644)
Additional records identified through reference harvesting all reviews (n = 166)
Records screened (n = 3,810)
Records excluded (n = 3,695)
Full-text articles assessed for eligibility (n = 115)
Full-text articles excluded, with reasons
Not Antecedent Materials Manipulation (n = 30)
Not Sensory-based materials manipulation (n=27)
Participant Age (n = 22)
Multicomponent Antecedent Intervention including non-materials manipulation components (n = 13)
Design does not separate participants who do and do not meet criteria (n = 5)
Not intervention study (n = 2)
Consequence-based intervention (n = 5)
Articles meeting inclusion criteria (n = 11)

Figure 1. PRISMA Search Procedure and Study Eligibility Reporting.

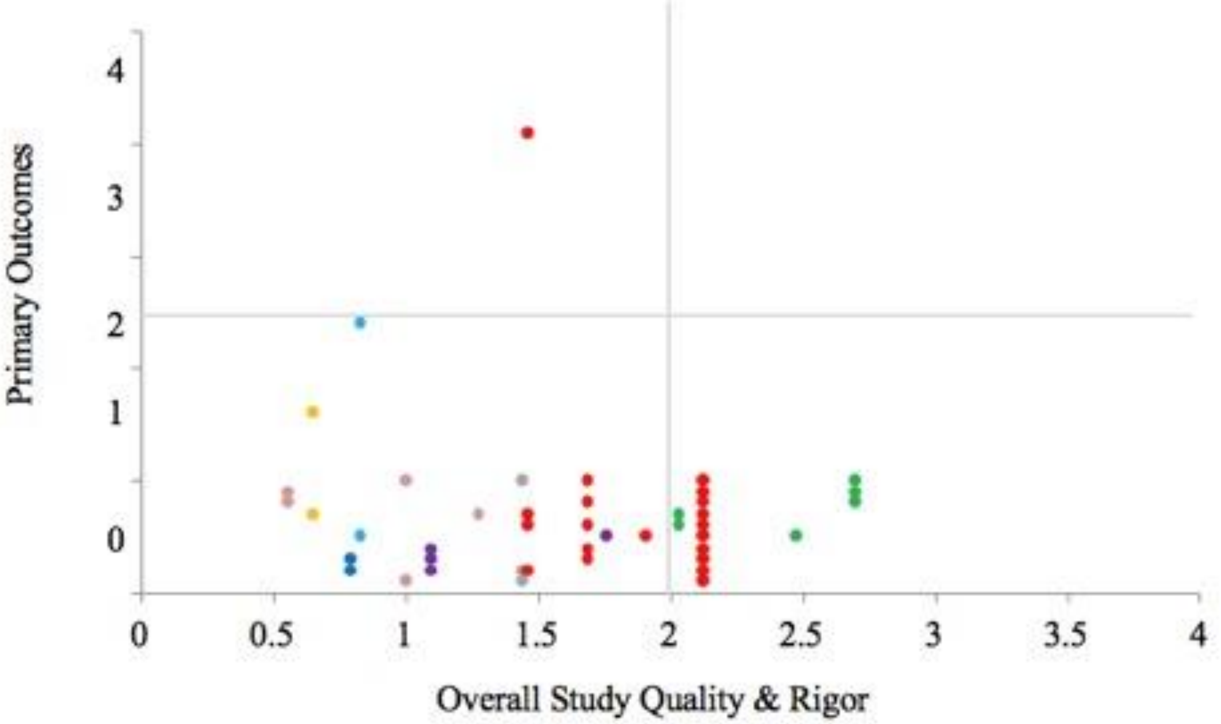

Figure 2. SCARF results showing most studies had low quality and rigor without positive outcomes (bottom left quadrant), some had higher quality and rigor without positive outcomes (bottom right) and one had low quality and rigor with a positive outcome (top left). Articles identified by first author. Green=Cox. Gray=Hodgetts. Light blue=Krombach. Dark blue=Leew. Pink=Olson. Red=Reichow 2010. Orange=Tunson. Purple=Umeda. Coordinate values by design available from first author.

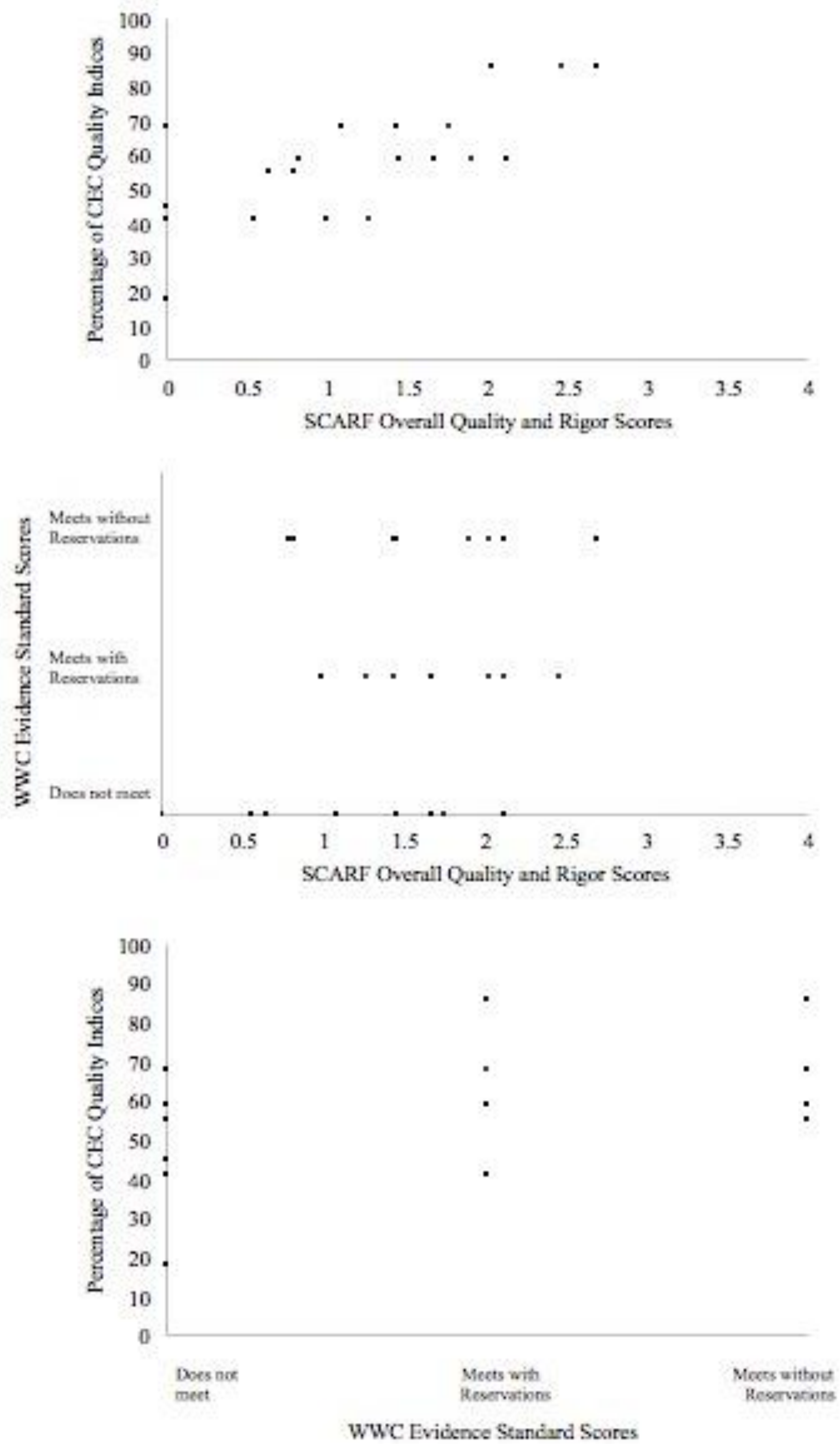

Figure 3. Associations between Quality Comparison Ratings across Tools.