Methodology: Systematic Review and Meta-analysis of the AAC Literature for People with Autism Spectrum Disorder or Intellectual Disabilities who have Complex Communication Needs through 2020

Jay B. Ganz

Texas A&M University

James E. Pustejovsky

University of Wisconsin-Madison

Joe Reichle

University of Minnesota

Kimberly J. Vannest

University of Vermont

Lauren M. Pierson

Sanikan Wattanawongwan

Texas A&M University

Man Chen

University of Wisconsin-Madison

Margaret Foster

Texas A&M University

Marcus C. Fuller

University of Vermont

April N. Haas

Life Skills Autism Academy

Bethany Hamilton

University of Texas at Austin

Mary R. Sallese

Texas A&M University

S. D. Smith

Southeast Missouri State University

Valeria Yllades

Texas A&M University

Corresponding author: Jay Ganz; 4225 TAMU, College Station, TX 77843, USA; Email: jayganz@tamu.edu

Author Contribution Statement using Contribution Roles Taxonomy (CRediT)

Authors are listed alphabetical order by tier with an explanation of the contributions that are indicated for each tier by using CRediT (Allen, O'Connell, & Kiermer, 2019).

Tier 1: Ganz, Pustejovsky, Reichle, Vannest (Principal Investigators)

Tier 2: Pierson, Wattanawongwan (Project staff)

Tier 3: Chen, Foster, Fuller, Haas, Hamilton, Sallese, Smith, Yllades (Additional staff or investigators who contributed substantively)

Jay B. Ganz: Conceptualization (lead); formal analysis (supporting); funding acquisition (lead); investigation (supporting); methodology (lead); project administration and supervision (lead); writing - original draft preparation (lead); writing - review and editing (lead). **James E. Pustejovsky**: Conceptualization (supporting); data curation (equal); formal analysis (lead); funding acquisition (equal); investigation (supporting); methodology (equal); project administration and supervision (equal); resources (equal); software (lead); visualization (lead); writing - original draft preparation (supporting); writing - review and editing (equal). **Joe Reichle**: Conceptualization (equal); data curation (supporting); formal analysis (supporting); funding acquisition (equal); investigation (supporting); methodology (equal); project administration and supervision (equal); resources (supporting); software (supporting); writing - original draft preparation (equal); writing - review and editing (equal). **Kimberly J. Vannest**: Conceptualization (equal); data curation (supporting); formal analysis (supporting); funding acquisition (supporting); investigation (equal); methodology (equal); project administration and supervision (equal); resources (equal); software (supporting); writing - original draft preparation (equal); writing - review and editing (equal). **Lauren Pierson**: Conceptualization (supporting); data curation (supporting); investigation (equal); methodology (supporting); resources (supporting); visualization (supporting); writing -

original draft preparation (supporting); writing - review and editing (supporting). **Sanikan Wattanawongwan**: Conceptualization (supporting); data curation (equal); formal analysis (supporting); investigation (equal); methodology (supporting); project administration and supervision (equal); resources (supporting); software (supporting); visualization (equal); writing - original draft preparation (equal); writing - review and editing (equal). **Man Chen**: Data curation (supporting); formal analysis (supporting); investigation (supporting); visualization (supporting). **Margaret Foster**: Conceptualization (supporting); data curation (supporting); investigation (supporting); methodology (supporting); project administration (supporting). **Marcus Fuller**: investigation (supporting). **April N. Haas**: investigation (supporting). **Bethany Hamilton**: data curation (supporting); investigation (supporting). **Mary R. Sallese**: investigation (supporting). **S. D. Smith**: investigation (supporting). **Valeria Yllades**: data curation (supporting); investigation (supporting).

Funding Acknowledgement

The research described here is supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R324A180110 to Texas A&M University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education

Abstract

A comprehensive meta-analysis examining the impacts of augmentative and alternative communication for individuals with complex communication needs was conducted, examining the relationship between social-communicative and educational outcomes and use of augmentative and alternative communication devices and across potential moderating variables. This document reports the methodology of the project as a whole, describing overarching procedures. Given the comprehensiveness of the review and meta-analysis, results of this review are reported in digestible groupings of types of research designs, types of research questions, and moderator groupings. Each of the resulting papers cite this primary document, as do additional reviews derived from the assembled data set.

The documents herewith report the overarching methodology of the project, including the following. The document searches occurred in 2018 and 2020, resulting in 7,327 documents reviewed for title/abstract indication of meeting inclusion criteria. Full text document review was conducted for 1,758 documents for the next inclusion/exclusion gate. Documents were divided into group design (n = 132) and single-case experimental design (SCED) documents (n = 547) and reviewed for basic design criteria, resulting in 59 group design documents and 257 SCED documents. Lead project authors conducted screening remaining documents for false positives, resulting in 20 group and 176 SCED documents remaining for further analysis. Data extraction and potential moderator variable coding procedures are described in detail, with relevant coding manuals and other materials attached. Effect size metrics used for meta-analytic procedures are detailed.

Methodology

A comprehensive literature review followed procedures outlined in the Cochrane guidelines (Higgins et al., 2019). The search and coding process began in 2018, concluded in 2020 reflecting literature between 1970 and 2020.

Search Procedures

Studies were identified through an electronic search utilizing the following databases: *Academic Search Complete, ERIC, PsycINFO, Conference Proceedings Citation Index – Social Science & Humanities (Web of Science),* and *Proquest Dissertations & Theses Global.* These databases were selected in consultation with a professional reference librarian (Meert et al., 2016) to be the most comprehensive, inclusive of conference proceedings, and best suited to include grey literature. The search included keywords based on: [((augmentative or alternative) within one word (w1) communicat*) or "sign language" or manual sign* or speech-generating device* or SGD or "voice output communication aid" or VOCA* or PECS or "picture exchange communication system" or AAC or "visual scene display" or "functional communication training"] AND [(down* w1 syndrome) or ((develop* or intellectual) w1 (delay* or disabil* or impair*)) or autis* or retard*]. Thesaurus terms matching these concepts in each database added additional synonymous terms to the search to widen the net of retrieved articles.

The initial database search occurred between October and December of 2018. A second search occurred in April 2020 to identify any additional articles. After electronic database searches, additional search methods included reference searches, first author searches, and forward searches. We used the *Web of Science* database to search each included document to review (a) references that have been cited in the included documents (reference searches), (b) first authors' other published studies (first author), and (c) any published studies that cited those included documents (forward search). The reference librarian trained

one phd level graduate research assistant to review for this stage. The reference librarian has conducted and consulted on systematic reviews for the past 10 years in a variety of disciplines including medicine, education, and public health. She is currently an Associate Professor, serving as the Systematic Reviews and Research Coordinator at the University library.

These search procedures identified 7,327 unique documents (duplicate title/abstracts removed). The publication types of documents before duplicates were removed include journal articles (6,573), dissertations (717), reports (252), conference (189), books (147), book chapters (140). See Appendix A for a PRISMA Flow Chart depicting the search and inclusion/exclusion steps.

Inclusion/Exclusion Criteria

Title and abstract (n = 7,327). We screened 7,327 documents against title/abstract criteria after duplicates were removed using the software Rayyan (Ouzzani et al., 2016). The documents were excluded when title or abstract indicated the research (a) did not involve an AAC intervention (including approaches to AAC that have been thoroughly discredited in the literature e.g., facilitated communication and rapid prompting method, supported typing), (b) did not include at least one participant with ASD, IDD with none having a complex communication need; or reported data on includes participant(s) that could not be disaggregated from the excluded participants, (c) did not involve social-communicative or challenging behavior outcomes, (d) did not utilize a single-case experimental design(s) (SCED) or between-groups design (GD), (e) was not available in English language. Articles were included in the next review step, the full-text review, unless specifically excluded (i.e. uncertainty in meeting or not meeting criteria). See Appendix B for a title and abstract inclusion/exclusion coding sheet.

Inter-rater reliability (IRR). A total of four raters evaluated 7,327 articles in the title/abstract stage. The four raters were doctoral students in special education with experience and expertise in conducting meta-analyses, systematic review, AAC, and SCED. All raters were trained to code for at least 80% accuracy from a small number of included documents. Then, all raters discussed any disagreement articles as a group before continuing to independently code for IRR purposes. All (100%) of articles were reviewed by two rates at this stage. Item by item percentage agreement was used to calculate IRR and the resulted in 88.50% for title/abstract stage.

Full text review (n = 1,758). Documents proceeding to full-text review numbered 1,758. The full-text inclusion criteria required that (a) the study was in English; (b) one or more participants with an intellectual delay, developmental disability(ies) (e.g. Angelman syndrome, cerebral palsy, autism spectrum disorder, Down syndrome etc. (IDD), such as Autism Spectrum Disorder (ASD), Intellectual disability (ID), other developmental disabilities (DD), with co-occurring complex communication needs (e.g., minimally- or non-verbal), mental retardation, cognitive disability, severe and profound, Down syndrome, Microcephaly, Apraxia, dyspraxia who received instruction (c) reported the results of an augmentative and alternative communication (AAC) intervention (AAC includes both unaided [e.g., sign language, sign system, gesture, manual sign] and aided systems [e.g., from low- mid and high-tech applications] to supplement or replace conventional speech for people with complex communication needs [CCN]); (d) was a SCED or GD; (e) measured social-communicative or social communicative and challenging behavior outcomes. Participants were excluded if they had primary diagnoses of physical impairments that impeded AAC use, had developmental disabilities other than ASD and did not report an IQ or cognitive assessment score demonstrating an intellectual impairment, or had multiple

disabilities (e.g., ASD and sensory impairments). See Appendix C for a full-text inclusion/exclusion coding sheet.

Inter-rater reliability (IRR). A total of four raters evaluated 1,758 articles in the Full-text stage. The same four doctoral students from the title/abstract stage were raters in this stage. All raters were trained from a sub number of articles to code for at least 80% accuracy at the beginning. Then, any discrepancies were discussed by all raters before starting to independently code. IRR calculations resulted in 88.39% accuracy across 39% of included articles. All raters met as a team to discuss any discrepancies between each coding batch.

Basic design quality standards review. A total of 679 documents met the inclusion criteria for full-text review and were continued for evaluating for basic design quality standards. Documents were divided into a set of SCED and a set of GD. Basic design quality standards criteria for both SCED and GD review were based on WWC basic standards (U.S. Department of Education [USDE], 2019).

Basic design quality standards screening: SCED Documents (n = 547). We reviewed remaining SCED documents for basic design quality standards, based on WWC standards (USDE, 2019). The six basic design standards include: a systematically manipulated independent variable; measured and reported inter-observer agreement (IOA); a minimum of 20% IOA collected across across data in baseline and intervention separately; at least 80% or .60 kappa IOA scores; at least three attempts data points by phases changes measured; at least three data points per baseline and intervention phases and at least four data per intervention phase for alternating treatment design.

For SCED documents, a total of 547 proceeded to a basic design quality standards review we termed an “efficient screening”. The screening method was used to reduce false positives as quickly as possible based on high-frequency fail criteria of IAO data adequacy

and 3 possible demonstrations of effect. Raters used an online form to identify prima facie violations of the inclusion rules for methodological quality.

The SCED documents resulted in 257 documents to be reviewed for against remaining design standards (systematically manipulated IV and data adequacy for the design) using WWC (Kratochwill et al., 2010, 2014; USDE, 2019). A total of 14 additional articles were excluded for this process. The SCED documents resulted in 243 documents to be screened during the next stage by the PIs.

Inter-rater reliability (IRR). A total of eleven raters evaluated 547 documents in the basic design standards screening stage for SCED articles. Two raters were professors in special education and served as a PI for the project. Both of them had an intensive experience conducting meta-analysis. Another nine doctoral students in special education were raters for this stage. All raters were trained to code and discuss criteria of the screening google form. IRR calculations resulted in 89.44% accuracy for the efficient screening across 20% of 537 documents. The IRR results in 94.33% accuracy for the remainder of the six basic design standards using WWC across 20% of included 257 documents. The PI and Co-PI met to discuss any discrepancies.

Basic design quality standards screening: GD Documents (n = 132). For GD documents, 132 documents proceeded to an "efficient screening" for basic design quality standards to identify clear violations of the inclusion rules for methodological quality. The standards for GD included (a) random selection and assignment to intervention and control group; (b) the sum of the overall attrition rate and five times the differential attrition rate did not exceed 60%; (c) equivalence was shown at baseline for the groups in the analytic sample. The GD screening identified 59 documents for the next step of review.

Inter-rater reliability (IRR). A total of three raters evaluated 132 articles in the basic design standards screening stage for GD articles. One rater was a professor in research

methods and statistics who served as a PI for the project, and who has extensive experience and expertise in meta-analysis. Another two raters were doctoral students in quantitative methods. All raters were trained to code and discussed criteria of the screening google form. IRR calculations resulted in 89.1% accuracy across 101 of the included articles. Discrepancies were reviewed by the PI rater and resolved through discussion among all three raters.

Screening to adjust for false positives of screening. Following screening for inclusion and exclusion criteria, the coding process was scheduled to occur. During the coding of dependent variables and participants, it became evident that the inclusion/exclusion criteria had a large number of false positives that were not evident based on the operational definitions despite initial testing of descriptions and highly reliable screening. The definitions were therefore reliable but not valid. Two PIs with decades of combined experience in the AAC professional world and literature engaged in an independent review of the 243 SCED documents and the 59 GD documents to re-assess inclusion/exclusion. Following this additional step, 176 SCED and 20 GD documents met criteria based on a re-reading of dependent variable procedures.

Inter-rater reliability (IRR). A minimum of two raters assessed 20% of SCED documents and 100% of GD documents against the revised full description of the inclusion/exclusion criteria. PhD raters engaged in the reliability assessment and produced 89% accuracy for SCED articles. Reliability for GD included two reviews and a discussion-based consensus method.

Full methodological quality standards review. After all screening gates were completed the total number of included articles was 195 (176 SCED, 20 GD) we assessed the documents for methodological quality using an aggregation from published standards by relevant national organizations, expert panels and federal guidelines (e.g., Council for

Exceptional Children [CEC], 2014; Horner et al., 2005; Reichow et al., 2008; USDE, 2019). Ratings for each standard reflect quality level (a) meets design standards, (b) meets design standards with reservations, or (c) does not meet standards. Ratings for GD studies were based on the relevant WWC design standards. The WWC standards for SCEDs were proposed more recently, continue to evolve, and have not achieved the same extent of field-wide consensus as the standards for GD studies. We therefore augmented the WWC standards with more detailed and extensive criteria. These extended methodological standards included criteria related to participants, settings, materials, implementers, procedures in baseline and intervention, the dependent variable, maintenance, generalizations, procedural fidelity, and social validity description. See Appendix D for the full methodological quality standards for SCED.

Inter-rater reliability (IRR). Two raters who are doctoral students in special education reviewed all articles for full methodological quality standards review. At the beginning of the stage, two raters were trained on each criteria of screening in the qualtric form by a PI and discussed any discrepancies. Raters were trained and practiced coding until they met 80% accuracy per each category of criteria. Any disagreements were reviewed by two raters until a consensus was reached, or a third rater who served as PI reviewed the discrepancy and made a final decision. The IRR resulted in a mean agreement of 89.87% (range 82%-96%) across 20% of included articles, by using percentage agreement.

Intervention Characteristics, Dependent Variables Characteristics, Participant Characteristics, and Extraction Procedures

The 196 documents (SCED: 176 documents; GD: 20 documents) were coded for variables related to: intervention characteristics, dependent variables characteristics, and participant characteristics. A total of eight graduate student coders were trained by a PI to code these variables. We coded intervention characteristics with regard to: (a) characteristics

of the intervention as described in the manuscript (i.e., social behavioral, functional behavioral), (b) instructional features (i.e., environmental arrangement, preference/reinforcement assessment, reinforcement, modeling, verbal prompting, physical prompting, prompt fading, graphic prompt), (c) named/manualized intervention, (d) functional communication training intervention, (e) setting of intervention implemented. Results were grouped and analyzed by intervention types/categories. We also coded dependent variables related to: (a) communicative function (i.e., behavior regulation, social interaction, joint attention), (b) expressive and receptive communication (i.e., communication production, communication comprehension), (c) communication mode (i.e., natural gesture, manual sign, low tech aided system, mid-to-high tech speech generating device, vocal, verbal), (d) function of the challenging behavior (if challenging behavior). Included articles were also coded for the following participants characteristics: (a) diagnosis (i.e, ASD, IDD), (b) age (i.e., pre-k, elementary, secondary), (c) number of words used prior to intervention, (d) combine symbols to phrases or sentences prior to intervention, (e) communication assessment, (f) cognitive/IQ assessment, (g) ASD diagnostic assessment, (h) communication mode used prior to study, (i) imitation, (j) joint attention. See Appendix E for details of coding for intervention characteristics, Appendix F for details of coding for dependent variables characteristics, and Appendix G details of coding for participant characteristics. After the variable coding stage, we combined some variables codes based on how the categories were developed.

Inter-rater reliability (IRR). Coding was conducted using online forms to reduce human error; variables were grouped into three categories to reduce workload. That is, coders worked in groups to code and review for reliability within three groups: intervention characteristics, dependent variables characteristics, and participant characteristics. For intervention characteristics, a total of three raters who are doctoral students in special

education coded articles for IRR purposes. At the beginning of the stage, all raters were trained on each criteria of each coding in the online form by PIs and discussed any discrepancies. They were trained and practiced coding until they all received 80% accuracy per each category of coding. Any disagreements were reviewed by all raters until a consensus was reached, or a third rater who served as PI was reviewed and gave a final decision. IRR results for intervention characteristic, dependent variable characteristic, and participant characteristic coding resulted in a mean agreement score of 93.11% (range = 85%-97.22%), 91.55% (range = 87.28%-96%), and 94.32% (range =84.62%-99.36%), respectively, using percentage agreement.

Race, Ethnicity, Home-language coding in Methodological Quality Review Project

We recorded the race, ethnicity, and home language of the 522 participants included in our review of the quality of the single-case experiments in our pool. Race, ethnicity, and home language were coded for each participant, divided by participant roles (i.e., person with ASD/ID [n=458], educator [n=28], parent of person with ASD/ID [n=17], peer or sibling [n=19]). The results found that information was reported for few of the participants on race (44%), ethnicity (7%), and home-language (13%). Details are provided in Appendix H.

Inter-rater reliability (IRR). Coding was conducted using an online form. Two raters discussed and practiced coding on each criteria and discussed any discrepancies. After all raters achieved more than 80% accuracy for each criteria of coding, they independently coded each document. IRR was conducted across 20% of documents with the accuracy results in a mean agreement score of 98.33% (range 96%-100%).

Outcome Data Extraction

Raw data were extracted from each A-B contrasts and outcome in each study from the graphs in the documents, with information about the A-B contrasts, outcome, phase, and session to which the data correspond. Included studies may have some, but not all

participants that meet the inclusion criteria; for example, a single study may include both typically developing participants and participants with disabilities, in which case typically-developing participants would be excluded.

Data were extracted by using Engauge Digitizer (Mitchell et al., 2017; markummitchell.github.io/engauge-digitizer), which is a freely available, open-source computer program for converting electronic images into numerical data. To improve accuracy, this tool allows for adjustment of pixel size, redefinition of the axis points, the addition of grid lines, overlays to compare extracted data to original graphs, and includes a wizard that can be used to complete each step of the process. Similar procedures have shown very high reliability (Shadish et al., 2009) and have been used in several previous systematic reviews of SCED research (Gage et al., 2012; Lequia et al., 2015; Losinski et al., 2014).

For any group-design studies that are identified, summary statistics (means, standard deviations, sample sizes) were extracted for each treatment group, outcome, and follow-up time where possible. If summary statistics were not available, other statistics (e.g., p-values, t- or F-statistics) were extracted, from which effect size estimates were determined using standard formulas (Borenstein et al., 2009).

Inter-rater reliability (IRR). For data extraction coding, a total of four raters who are doctoral students in special education double-coded a subset of articles for IRR purposes. At the beginning of the stage, all raters were trained how to extract data by PIs. IRR was measured using intra-class correlations (ICC) across raters based on calculated effect size estimates. We therefore report separate IRR results for each of three effect size indices: Tau, LRR, and the within-case SMD. IRR assessment was conducted on case-specific effect size estimates--rather than on the raw data--because subsequent analysis was all based on meta-analysis of calculated effect sizes. Coding discrepancies in the raw data that do not influence calculated effect size estimates are therefore inconsequential. For Tau, IRR results were

calculated across 190 cases from 40 studies, with an overall ICC of 0.99. For LRR and within-case SMD, IRR results were calculated after excluding cases where all baseline data points were zero; across 116 cases from 34 studies, overall ICC was 0.98 for LRR and 0.99 for within-case SMD.

Effect Size Measures

Effect size metrics quantify a change between treatment and non-treatment conditions in experimental designs. Confidence intervals describe a range of possible values within a degree of certainty and together provide an understanding of how much change might occur again under similar conditions with like participants. Quality of studies was assessed, and studies scored below criterion were not included in the analysis. Effect sizes calculated and reported here include parametric (response ratio & BC-SMD) and non-parametric analysis (Tau-U).

Response Ratio. The response ratio (Pustejovsky, 2014) is an effect size that quantifies treatment effects in terms of proportionate change from baseline, which may be particularly appropriate for SCED studies that use behavioral outcomes measured through direct observation. This effect size has the advantage of being intuitively interpretable (as percentage change) and is closely related to other effect sizes that have been used in synthesis of SCED research, such as the Mean Baseline Reduction (Campbell, 2004) and Suppression Index (Marquis et al., 2000). Response ratios are also used in meta-analysis of GDs (e.g., Hedges et al., 1999). The main conceptual limitation of the LRR is that it is not meaningful for measuring change in behavior from near-zero baseline levels. Further drawbacks of this effect size are that, as currently developed, it does not account for time trends and its standard error is sensitive to auto-correlation. However, the latter drawback can be addressed through the use of robust variance estimation in the meta-analysis.

Tau-U. Tau-U (Parker et al., 2011) is a non-overlap measure that combines Mann

Whitney U and Kendall's Tau to identify the magnitude of change across time and accounts for undesirable trend in baseline. Tau-U does not rely on distributional assumptions about the outcome measurements (e.g., normality) that may be inappropriate for data from single-case designs and Tau-U. Interpretive guidelines for Tau-U values for mid-to-high-tech AAC research were established in a prior work (Ganz et al., 2017). Tau-U values could roughly be interpreted as follows which may be connected to this study: study to very strong effects: 0.93-1.00, moderate effects: 0.80-0.92, low effects: 0.65-0.79, no to very effects: ≦ 0.64. New guidelines will be published as a result of the current, comprehensive project.

Tau-U procedures. Aggregation of data is achieved through compilation of AB phase contrasts within designs and also across studies as demonstrations of effects. In designs with replications across participants/settings/behaviors (e.g., multiple baseline design) each AB is included; in a reversal or withdrawal design (ABAB) the first AB is used. In alternating treatment designs, concurrent phases only are included. For each AB comparison the inverse of the variance for the data in that AB phase contrast is used for weighting. Sample size weighting was used, and results compared between methods. An additional aggregation included use of phase-length only and a comparison of phase-length vs variance weighting were conducted to answer questions about the effects of autocorrelation. These may answer additional questions about aggregating single case design results, provide more information about sampling distribution, provide more information about relative changes in relationship to number of data points in a series, and provide consistent or alternative explanations to differences in magnitude dependent on length of observation sessions used to challenging behavior; see Pustejovsky, 2015). Unknown distributions can be addressed through robust variance estimation; and the impact of sampling were mitigated by including relevant procedural details as moderators in the meta-analysis.

Between-case standardized mean difference (BC-SMD). BC-SMD is an ES that is

comparable, in principle, to the d-index from a between-groups randomized experiment conducted on the same population and with the same DV (Hedges et al., 2012, 2013). The estimation methods control for auto-correlation, time-trends, and between-case variability (Pustejovsky et al., 2014). BC-SMD provides a means for comparing the magnitude effects from SCEDs to those from between-groups designs. Technical drawbacks include: aggregation across individual-level results, potentially concealing variation that may be of substantive interest (Kratochwill & Levin, 2014). Three cases are needed for calculations, and so some studies must be excluded from analysis based on BC-SMD. The distribution theory supporting this method is based on parametric assumptions that may not be present in SCED studies, particularly those that examine behavioral outcomes near floor or ceiling levels.

Calculating effect sizes. Effect sizes were calculated from available data for each study. In order to ensure accuracy and full reproducibility, the calculations were performed in the R statistical computing environment. SCED effect size calculations were carried out using functionality developed by Co-PI Pustejovsky (Pustejovsky et al., 2014). GD effect size calculations were carried out using the metafor package for R (Viechtbauer, 2010), a well-developed tool that is used across many areas of meta-analysis.

Moderator Analysis

Separate analyses were conducted for each effect size index. A mixed-effects meta-regression models with robust variance estimation assumes true effect sizes may vary across cases (Borenstein et al., 2009). This analysis meets the characteristics of our data from heterogeneous populations and additionally provides information relevant for generalizing. Hierarchical linear modeling (HLM) was used for two case-level effect sizes (response ratio and Tau-U). Effect sizes for each case were nested within each study (Van den Noortgate & Onghena, 2008). A typical mixed-effects meta-regression model was also used when the

hierarchical model was not necessary because the effect size was already at the study level (i.e., represents an aggregation of effects across cases); instead, we used a typical mixed-effects meta-regression model. If sufficient effect sizes from GD studies are identified, they were analyzed using a typical mixed-effects meta-regression model.

Robust variance estimation. The proposed mixed-effects meta-analytic approach is complicated by the fact that the estimated sampling variances of the effect size measures may not be accurate (e.g., due to auto-correlation in the outcome measures, or for the between-groups d-index, due to the small number of cases per study). Conventional meta-analytic techniques can break down when effect size variance estimates are inaccurate. However, an innovation in meta-analysis called robust variance estimation (Hedges et al., 2010) provides a means to conduct valid meta-analysis and meta-regression when sampling variances are inaccurate or unknown, as well as when each study contributes multiple effect size estimates that may be statistically dependent. Robust variance estimation involves using a "working model" to estimate weighted least-squares estimates of meta-regression models, but uses variance estimation techniques that are valid even when the working model is incorrect. For the case-level effect sizes, we used robust variance estimation with the "hierarchical" working model, as described in Hedges et al. (2010); for the between-groups d-index, we used a random-effects working model. We followed recent recommendations by using finite-sample corrections for hypothesis tests based on robust variance estimation (Tipton & Pustejovsky, 2015; Tipton, 2014).

Publication/outcome reporting bias. An important threat to the validity of any meta-analysis is the possibility that the identified studies may not be representative of the full range of potential results, as might occur if studies that demonstrate clear and large effects have a higher chance of being published than studies demonstrating weak or inconsistent findings (Rothstein, Sutton, & Borenstein, 2005). There is a growing concern regarding publication

bias in syntheses of SCED studies (e.g., Ganz et al., 2012; Sham & Smith, 2014; Shadish et al., 2016). Our primary strategy for addressing the concern of publication bias was searched for and included both published and unpublished studies, including gray literature, and to test for differences in the magnitude of effects between published and unpublished studies.

In meta-analyses of GD studies, it is common to also use graphical diagnostics such as funnel plots and statistical tests (e.g., Egger, Smith, Schneider, & Minder, 1997; Stanley & Doucouliagos, 2014) to assess, and potentially to adjust for, the threat of publication bias. However, these diagnostics are designed for the GD literature and are may not be entirely suitable for application to SCED research, where the process that leads to publication bias in the SCED literature is likely to be driven by factors such as visual determinations of experimental control and functional relationships rather than by the statistical significance of results (Shadish et al., 2015). Lacking methods that are better-suited for single-case research, we followed the precedent of other recent meta-analyses of SCEDs (e.g., Heyvaert et al,, 2012; Shadish et al., 2014) and use funnel plot diagnostics and associated tests to investigate the possibility that studies are selected based on the significance of findings, but interpret the results with a measure of caution.

Software and reproducibility. In order to ensure the accuracy and reproducibility of the investigation, all analysis was carried out in the R statistical computing environment, using the metafor (Viechtbauer, 2010) and robumeta packages (Fisher & Tipton, 2014); all of this software is open-source. Datasets and computer code for replicating all reported analyses was made available in the supplementary materials for each of the planned publications.

References

Allen, L., O'Connell, A., & Kiermer, V. (2019). How can we ensure visibility and diversity in research contributions? How the Contributor Role Taxonomy (CRediT) is helping the shift from authorship to contributorship. *Learned Publishing, 32*(1), 71-74.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). Fixed-effect versus random-effects models. *Introduction to Meta-analysis, 77*, 85.

Campbell, J. M. (2004). Statistical comparison of four effect sizes for single-subject designs. *Behavior modification, 28*(2), 234-246.

Council for Exceptional Children. (2014). Council for Exceptional Children standards for evidence-based practices in special education. Retrieved from http://www.cec.sped.org/Standards/Evidence-Based-Practice-Resources-Original

Fisher, Z., & Tipton, E. (2014). robumeta: An R-package for robust variance estimation in meta-analysis. Working paper.

Gage, N. A., Lewis, T. J., & Stichter, J. P. (2012). Functional behavioral assessment-based interventions for students with or at risk for emotional and/or behavioral disorders in school: A hierarchical linear modeling meta-analysis. *Behavioral Disorders*, *37*(2), 55-77.

Ganz, J.B., Earles-Vollrath, T.L., Heath, A.K., Parker, R.I., Rispoli, M.J., & Duran, J.B. (2012). A meta-analysis of single case research studies on aided augmentative and alternative communication systems with individuals with autism spectrum disorders. Journal of Autism and Developmental Disorders, 42, 60–74. Retrieved from http://dx.doi.org/ 10.1007/s10803-011-1212-2

Ganz, J. B., Morin, K. L., Foster, M. J., Vannest, K. J., Genç Tosun, D., Gregori, E. V., & Gerow, S. L. (2017). High-technology augmentative and alternative communication for individuals with intellectual and developmental disabilities and complex

communication needs: A meta-analysis. *Augmentative and Alternative Communication, 33,* 224-238.

Hedges, L. V., Gurevitch, J., & Curtis, P. S. (1999). The meta-analysis of response ratios in experimental ecology. *Ecology*, *80*(4), 1150-1156.

Hedges, L. V., Pustejovsky, J. E., & Shadish, W. R. (2012). A standardized mean difference effect size for single case designs. *Research Synthesis Methods*, *3*(3), 224-239.

Hedges, L. V., Pustejovsky, J. E., & Shadish, W. R. (2013). A standardized mean difference effect size for multiple baseline designs across individuals. *Research Synthesis Methods*, *4*(4), 324-341.

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research synthesis methods*, *1*(1), 39-65.

Heyvaert, M., Maes, B., Van Den Noortgate, W., Kuppens, S., & Onghena, P. (2012). A multilevel meta-analysis of single-case and small-n research on interventions for reducing challenging behavior in persons with intellectual disabilities. *Research in developmental disabilities*, *33*(2), 766-780.

Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M.J., Welch, V.A. (editors). *Cochrane Handbook for Systematic Reviews of Interventions* version 6.0 (updated July 2019). Cochrane, 2019. Available from www.training.cochrane.org/handbook.

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children*, 71, 165-179.

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M & Shadish, W. R. (2010). Single-case designs technical documentation. Retrieved

from What Works Clearinghouse website: http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf.

Kratochwill, T. R., & Levin, J. R. (2014). Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue. *Psychological Methods*.

Lequia, J., Wilkerson, K. L., Kim, S., & Lyons, G. L. (2015). Improving transition behaviors in students with autism spectrum disorders: A comprehensive evaluation of interventions in educational settings. *Journal of Positive Behavior Interventions*, *17*(3), 146-158.

Losinski, M., Cuenca-Carlino, Y., Zablocki, M., & Teagarden, J. (2014). Examining the efficacy of self-regulated strategy development for students with emotional or behavioral disorders: A meta-analysis. *Behavioral Disorders*, *40*(1), 52-67.

Marquis, J. G., Horner, R. H., Carr, E. G., Turnbull, A. P., Thompson, M., Behrens, G. A., ... & Doolabh, A. (2000). A meta-analysis of positive behavior support. *Contemporary special education research: Syntheses of the knowledge base on critical instructional issues*, *11*, 137-178.

Meert, D., Torabi, N., & Costella, J. (2016). Impact of librarians on reporting of the literature searching component of pediatric systematic reviews. *Journal of the Medical Library Association: JMLA*, *104*(4), 267.

Mitchell, M., Muftakhidinov, B., & Winchen, T. (2017). Engauge digitizer software. *Webpage: http://markummitchell. github. io/engauge-digitizer. Accessed*, *11*.

Ouzzani, M., Hammady, H., Fedorowicz, Z. (2016). Rayyan—a web and mobile app for systematic reviews. *Syst Rev* 5, 210. https://doi.org/10.1186/s13643-016-0384-4

Parker, R. I., Vannest, K. J., Davis, J. L., & Sauber, S. B. (2011). Combining nonoverlap and trend for single-case research: Tau-U. *Behavior Therapy*, *42*(2), 284-299.

Pustejovsky, J. E. (2015). Measurement-comparable effect sizes for single-case studies of free-operant behavior. *Psychological Methods*, *20*(3), 342.

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

Reichow, B., Volkmar, F. R., & Cicchetti, D. V. (2008). Development of the evaluative method for evaluating and determining evidence-based practices in autism. *Journal of Autism and Developmental Disorders*, *38*(7), 1311-1319.

Shadish, W. R., Brasil, I. C., Illingworth, D. A., White, K. D., Galindo, R., Nagler, E. D., & Rindskopf, D. M. (2009). Using UnGraph to extract data from image files: Verification of reliability and validity. *Behavior Research Methods*, *41*(1), 177-183.

Shadish, W. R., Hedges, L. V., Horner, R. H., & Odom, S. L. (2015). The Role of Between-Case Effect Size in Conducting, Interpreting, and Summarizing Single-Case Research. NCER 2015-002. *National Center for Education Research*.

Shadish, W. R., Hedges, L. V., & Pustejovsky, J. E. (2014). Analysis and meta-analysis of single-case designs with a standardized mean difference statistic: A primer and applications. *Journal of School Psychology*, *52*(2), 123-147.

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

Tipton, E. (2014). How generalizable is your experiment? An index for comparing experimental samples and populations. *Journal of Educational and Behavioral Statistics*, *39*(6), 478-501.

Tipton, E., & Pustejovsky, J. E. (2015). Small-sample adjustments for tests of moderators and model fit using robust variance estimation in meta-regression. *Journal of Educational*

and Behavioral Statistics, *40*(6), 604-634.

Van den Noortgate, W., & Onghena, P. (2008). A multilevel meta-analysis of single-subject experimental design studies. *Evidence-Based Communication Assessment and Intervention*, *2*(3), 142-151.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of statistical software*, *36*(3), 1-48.

U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, What Works Clearinghouse. (2019, September). *What Works Clearinghouse: Procedures and Standards Handbook* (Version 4.1). http://whatworks.ed.gov